Testing the Boundaries of LLMs: Dialectal and Language-Variety Tasks

Fahim Faisal[1], **Antonios Anastasopoulos**[1,2]
[1]Department of Computer Science, George Mason University
[2]Archimedes/Athena RC, Greece
`{ffaisal,antonis}@gmu.edu`

Abstract

This study evaluates the performance of large language models (LLMs) on benchmark datasets designed for dialect-specific NLP tasks. Dialectal NLP is a low-resource field, yet it is crucial for evaluating the robustness of language models against linguistic diversity. This work is the first to systematically compare state-of-the-art instruction-tuned LLMs—both open-weight multilingual and closed-weight generative models—with encoder-based models that rely on supervised task-specific fine-tuning for dialectal tasks. We conduct extensive empirical analyses to provide insights into the current LLM landscape for dialect-focused tasks. Our findings indicate that certain tasks, such as dialect identification, are challenging for LLMs to replicate effectively due to the complexity of multi-class setups and the suitability of these tasks for supervised fine-tuning. Additionally, the structure of task labels—whether categorical or continuous scoring—significantly affects model performance. While LLMs excel in tasks like machine reading comprehension, their instruction-following ability declines in simpler tasks like POS tagging when task instructions are inherently complex. Overall, subtle variations in prompt design can greatly impact performance, underscoring the need for careful prompt engineering in dialectal evaluations.[1]

1 Introduction

Natural Language Processing (NLP) systems have traditionally focused on high-resource languages, leaving dialectal variations underexplored (Kantharuban et al., 2023). In this work, we address this gap by evaluating large language models (LLMs) on task-specific benchmark datasets curated for various dialects. Dialectal tasks often lack the resources available for standard languages, but they provide critical insights into a model's robustness across linguistic diversity (Joshi et al., 2024). To our knowledge, no prior studies have systematically assessed LLM performance on dialect-focused NLP tasks. We compare LLMs such as GPT-4 (OpenAI, 2023) and Aya-101 (Üstün et al., 2024) with state-of-the-art multilingual encoder models like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) to establish new baselines and identify areas where LLMs either excel or fall short.

Our Contributions: We make several key contributions to the understanding of LLM performance in dialect-specific tasks:

- We conduct the first systematic evaluation of LLMs on dialectal NLP tasks across seven NLP tasks, comparing instruction-tuned models (GPT-4, Aya-101) with fine-tuned encoder models (mBERT, XLM-R) to establish new baselines.
- Our findings reveal significant limitations of LLMs in complex multi-class dialect identification tasks, where in-context learning with large LLMs falls short compared to fine-tuned encoders. Adding more prompt examples yields only slight gains, while Aya-101 shows a strong bias, frequently misclassifying Arabic varieties as Sudanese Arabic.
- We show that LLM performance is influenced by task label structure (e.g., categorical vs. continuous), with challenges arising in score-based sentiment classification for specific dialects.
- LLMs excel in Machine Reading Comprehension but struggle with simpler tasks like POS tagging when instructions are complex, underscoring the need for clear task framing.

Overall, this study contributes to a deeper understanding of LLM behavior in low-resource, dialect-rich environments and emphasizes the need for

[1]Code repository: `https://github.com/ffaisal93/DialectBench`

Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 68–92
January 19, 2025.

tailored approaches when working with dialectal NLP tasks.

2 Dialectal Datasets and Benchmarking

DIALECTBENCH: To evaluate LLMs on dialect-specific tasks, we utilize the design framework and task dataset collections from DIALECTBENCH (Faisal et al., 2024), a benchmark that focuses on language varieties organized into structured language clusters. In this benchmark, a *language cluster* is a group of related language varieties that share a common linguistic origin and exhibit similarities in grammar and vocabulary. Each cluster includes several language varieties with shared ancestry, based on the Glottocode classification (Hammarström and Forkel, 2022). Within each cluster, a *cluster representative* is selected to serve as a standardized reference point for evaluating the entire group. This makes it easier to compare model performance across different dialects within the same cluster. For example, in the Arabic language cluster, Modern Standard Arabic (MSA) often acts as the representative variety when it is available for a task. This method allows for consistent and efficient evaluation of models across various dialectal forms.

Task Selection: We experiment with seven tasks from the DIALECTBENCH task collections. These tasks are:

1. Parts-of-Speech (POS) Tagging
2. Dialect Identification (DId)
3. Sentiment Analysis (SA)
4. Topic Classification (TC)
5. Natural Language Inference (NLI)
6. Multiple-Choice Machine Reading Comprehension (MRC)
7. Extractive Question Answering (EQA)

Table 1 provides an overview of the datasets used for each task, including the number of language clusters and varieties covered. These tasks were selected based on their data availability across diverse dialectal varieties. For instance, POS tagging, as a structured prediction task, utilizes the Universal Dependency dataset, which includes 11 clusters and 25 varieties. Classification tasks, such as Dialect Identification (DID), Sentiment Analysis (SA), Topic Classification (TC), and Natural Language Inference (NLI), draw from datasets like MADAR, DSL-TL, and TSAC, among others. Similarly, for question answering tasks, including Machine Reading Comprehension (MRC) and Extractive Question Answering (EQA), we utilize datasets like Belebele and SDQA, with these tasks covering between 4 to 5 clusters and multiple varieties. In Appendix Table 6, we report all the language clusters and their varieties explored in this study.

3 Experimental Setup

This section outlines the selected language models for evaluation, along with the training and evaluation configurations.

3.1 Models

We utilize four models with varying sizes and capabilities: mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), GPT-4 (OpenAI, 2023), and Aya-101 (Üstün et al., 2024). The first two, mBERT and XLM-R, are multilingual encoder-based models trained using masked language modeling and next-token prediction tasks across hundreds of languages. We finetune these pretrained models on task-specific datasets using supervised setups.

In contrast, GPT-4 and Aya-101 are large-scale generative models designed for instruction following. Aya-101 is an open-weight multilingual instruction-tuned model built on the T5 (Raffel et al., 2020) encoder-decoder architecture, and it has been trained on data covering 101 languages. On the other hand, GPT-4 is a closed-weight model. Due to GPT-4's large scale and diverse data exposure, we hypothesize that it may exhibit strong robustness across multilingual settings.

3.2 Training Configuration

DIALECTBENCH datasets have an uneven distribution of training data availability across tasks and varieties. As a result, we opted for a diverse set of task-specific finetuning configurations best suited for the available resource utilization. A summary of these configurations is reported in Table 2. The following subsections further clarify the different experimental setups:

1. **Cross-Lingual Transfer from English:** For several tasks, we faced low-resource training data for certain varieties. As a result, it wouldn't be a fair comparison to fine-tune some varieties on high-resource data while others are fine-tuned on low-resource data. To

Category	Task	Metric	#Clusters	#Varieties	Source Dataset
Structured Prediction	POS tagging	F1	11	25	Universal Dependency (Zeman et al., 2021), Singlish (Wang et al., 2017)
Classification	DId	F1	6	45	MADAR (Bouamor et al., 2018), DMT (Jauhiainen et al., 2019), Greek (Sababa and Stassopoulou, 2018), DSL-TL (Zampieri et al., 2023), Swiss Germans (Scherrer et al., 2019)
	SA	F1	1	9	TSAC (Medhaffar et al., 2017), TUNIZI (Fourati et al., 2021), DzSentiA (Abdelli et al., 2019), SaudiBank (Alqahtani et al., 2022), MAC (Garouani and Kharroubi, 2022), ASTD (Nabil et al., 2015), AJGT (Alomari et al., 2017), OCLAR (Al Omari et al., 2019)
	TC	F1	15	38	SIB-200 (Adelani et al., 2023)
	NLI	F1	15	38	XNLI (Conneau et al., 2018) translate-test
Question Answering	MRC	F1	4	11	Belebele (Bandarkar et al., 2024)
	EQA	Span F1	5	24	SDQA (Faisal et al., 2021)

Table 1: DIALECTBENCH tasks used to evaluate generative models against multilingual encoders. This table presents selected dialectal and variety-specific datasets, highlighting task metrics, number of language clusters, and varieties. The study extends the original benchmark to compare instruction-tuned LLM performance with traditional multilingual models.

Task	Encoder (finetune)				LLM (k-shot ICL)			
	English	Cluster-rep.	Variety	Combined	English	Cluster-rep.	Variety	Combined
SA	-	-	-	✓	-	-	-	✓
TC	✓	✓	-	-	✓	-	✓	-
NLI	✓	-	-	-	✓	-	-	-
MRC	-	-	-	✓	-	-	-	✓
EQA	✓	-	-	✓	✓	-	-	✓
POS tagging	✓	-	-	-	✓	-	-	-
DId	-	-	-	✓	-	-	-	✓

Table 2: Task-specific experimental configurations: Encoder models are fine-tuned on English data, representative languages of each cluster, or a mixture of language varieties. In contrast, LLMs employ k-shot In-Context Learning (ICL) using prompts in English, the representative language of the cluster, the target language variety, or a combination of these language varieties.

address this, we adopted a more practical approach: fine-tuning on standard English task data, which is almost always available, and performing zero-shot evaluations on all target varieties. We applied this method for POS tagging, Topic Classification, Extractive QA, and NLI.

2. **Finetuning on Cluster-representative:** In addition to cross-lingual transfer from standard English, we conducted an experiment where encoder models were fine-tuned on cluster representatives within the Topic Classification dataset. This approach was feasible because all cluster-representative training data for this task was equal in size. The result is a set of cluster-specific, fine-tuned Topic Classification models, which we then used to evaluate performance on their respective cluster varieties.
3. **Combined Fine-tuning:** Instead of fine-tuning on a single variety, for tasks such as Sentiment Classification and Dialect Identification, we fine-tune using a combined dataset from all varieties to create a supervised classification model. For tasks like Extractive QA and Machine Reading Comprehension, the training data is limited to multiple standard varieties. Consequently, for these tasks, we also fine-tune on the available combined training data and then evaluate performance on the other available dialects.
4. **In-Context Learning:** For LLMs, we skip fine-tuning and rely on in-context learning (ICL) with randomly chosen k-shot examples (k=3) in either English, the target cluster-representative, or the target variety itself. For classification tasks with a large number of categories (e.g., Dialect Identification), we provide one example per class to keep the prompt sequence manageable. Additionally, for tasks involving combined training data (e.g., Extractive QA and Machine Reading Comprehension), we sample out our k-shot examples from this aggregated set.

For all instruction prompts used in task-specific in-context learning, we keep the instructions as straightforward as possible, opting for the simplest form of task description. This approach ensures that the model's performance is primarily a reflection of its inherent capabilities rather than prompt engineering. All task-specific instruction prompts can be found in Appendix A.

3.3 Evaluation Criteria

Our study is structured to empirically identify failure cases in LLM performance using encoder models as baselines. In-context learning via prompting is exclusively employed for LLMs (Aya-101 and GPT-4). On the other hand, encoder models are evaluated using supervised fine-tuning setups, which are deterministic, unlike LLMs which can exhibit variability in responses depending on prompt phrasing and context. When we observe inconsistencies or failures, we analyze these cases further in the task analysis section to hypothesize potential root causes and conduct targeted ablation studies to investigate specific issues.

Metrics: For task-specific comparative evaluation, we calculate metrics such as F1 score and Accuracy for different tasks, as presented in Table 1. Guided by the task configurations outlined in Table 2, we identify the highest achievable performance for each language variety and task combination, comparing smaller, encoder-based models with larger LLMs. Using these performance scores, we establish two comparative metrics based on performance deltas, denoted as $\Delta_{\text{LLM-enc}}$ and $\Delta_{\text{closed-open}}$:

- $\mathbf{\Delta}_{\text{LLM-enc}}$: This metric represents a global comparison across all model types, measuring the performance difference between the best small-sized, non-instruction-tuned encoder models and instruction-tuned large language models (LLMs).
- $\mathbf{\Delta}_{\text{closed-open}}$: This metric is a local comparison within the LLM category, representing the performance gap specifically between a closed-weight instruction-tuned LLM (GPT-4) and an open-weight multilingual instruction-tuned LLM (Aya-101).

These two metrics are used to pinpoint anomaly cases and to identify general trends and differences when transitioning from non-instruction-tuned small-sized encoder models to instruction-tuned LLMs, as well as when comparing closed-weight and open-weight instruction-tuned LLMs.

Task	Metric	mBERT	XLM-R	GPT4	AYA
SA	Acc	78.8	**80.1**	69.1	<u>65.8</u>
TC	F1	75.3	<u>73.1</u>	**84.9**	79.2
NLI	F1	<u>58.4</u>	63.3	**68.9**	63.6
MRC	F1	<u>39.4</u>	40.3	**80.8**	71.7
EQA	F1	69.2	<u>67.2</u>	53.8	**73.1**
POS tagging	F1	52.5	51.2	**59.8**	<u>15.9</u>
DId	F1	**65.7**	59.3	27.9	<u>16.4</u>

Table 3: Average maximum task performance for each model under various configurations (e.g., transfer from English, in-cluster tuning, ICL). The bold values indicate the highest performance achieved for each task, while underlined values mark the lowest performance. GPT-4 generally outperforms other models across most tasks, while AYA struggles significantly with POS tagging and LLM generally fails on the multi-label Dialect Identification task.

4 Takeaway from Task-Specific Results

Table 3 presents a summary of average maximum task performance across various models. We observe that GPT-4 generally performs well in Machine Reading Comprehension (MRC) and Natural Language Inference (NLI) tasks, outperforming smaller encoder-based models in these areas. However, GPT-4 lags in tasks such as Parts of Speech (POS) tagging and Extractive Question Answering (EQA), where encoder-based models like mBERT and XLM-R outperform it. Aya-101, despite being multilingual, consistently struggles, especially in complex tasks like POS tagging and Dialect Identification (DID).

Table 4 highlights the variability in model performance based on different language varieties. For certain tasks like MRC and NLI, the performance gap between LLMs and encoder models is positive, indicating superior results for LLMs. However, for tasks like DID and POS tagging, LLMs underperform significantly compared to encoder-based models, especially when tasked with handling diverse or low-resource language varieties.

We provide detailed task-specific results in Appendix D Tables 8 to 14. Based on these results, our key takeaways are as follows:

Classification Gap Due to Label Differences The sentiment analysis task aggregates data at the level of different Arabic varieties from various sources, which contain a diverse set of task labels per dialect, significantly contributing to the differences in performance across dialects. The results

$\Delta_{\text{LLM-enc}}$					
Task	Avg	Min_Variety	Min	Max_Variety	Max
SA	-8.90	arabic, egyptian arabic	-41.79	arabic, arabic (a:jordan)	3.34
TC	7.70	sinitic, cmm sinitic (o:traditional)	-4.41	kurdish, central kurdish	58.85
NLI	6.59	sinitic, cantonese	-3.33	sotho-tswana (s.30), southern sotho	26.69
MRC	42.31	sotho-tswana (s.30), northern sotho	31.00	arabic, egyptian arabic	50.61
EQA	2.27	anglic, indian english (a:south)	-6.88	korean, korean (a:south-eastern, m:spoken)	47.45
POS tagging	3.61	anglic, english	-9.40	saami, north saami	20.76
DId	-38.15	(sinitic, m. chinese (a:taiwan, o:simp.))	-87.58	(anglic, north american)	-4.20

$\Delta_{\text{closed-open}}$					
Task	Avg	Min_Variety	Min	Max_Variety	Max
SA	3.29	arabic, moroccan arabic	-9.45	arabic, south levantine arabic	36.59
TA	5.08	sotho-tswana (s.30), northern sotho	-6.81	arabic, standard arabic	9.55
NLI	5.39	latvian, east latvian	-16.74	sw. shif. romance, portuguese (a:european)	20.42
MRC	9.14	sotho-tswana (s.30), northern sotho	-14.85	arabic, egyptian arabic	18.21
EQA	-17.46	bengali, vanga (a:west bengal)	-32.75	anglic, philippine english	-8.62
POS tagging	43.86	tupi-guarani subgroup i.a, old guarani	-0.55	high german, german	76.53
DId	11.47	(southwestern shifted romance, spanish)	-32.74	(arabic, rabat-casablanca arabic)	41.65

Table 4: Task-specific performance summary across $\Delta_{\text{LLM-enc}}$ and $\Delta_{\text{closed-open}}$ metrics. A positive $\Delta_{\text{LLM-enc}}$ indicates that LLMs with in-context learning (ICL) outperform supervised fine-tuning of smaller encoders, while a negative value suggests the opposite. A positive $\Delta_{\text{closed-open}}$ indicates GPT-4's closed-weight superiority over the open-weight Aya-101, whereas a negative value favors Aya-101. For each task, the table shows the average delta, along with minimum and maximum values across language varieties, identifying the language cluster and delta.

in Table 9 show that, in two cases—Tunisian Arabic and Egyptian Arabic—we observe a more pronounced performance gap ($\Delta_{\text{LLM-enc}}$) between the LLMs and encoder models. We find that the classification labels are `['positive', 'neutral', 'objective', 'negative']` and `['neutral', 'positive', 'negative']` for these two dialects, respectively. The results suggest that LLMs, especially when using in-context learning, struggle with the increased number of classification labels, which is further compounded by their limited grasp of these specific Arabic dialects.

Moreover, considering $\Delta_{\text{closed-open}}$ for South Levantine Arabic, we observe a notable gap between the two LLMs, GPT-4 and Aya-101. The classification labels for this dialect are `[1, 2, 3, 4, 5]`. Despite being a multilingual instruction-tuned model, it becomes evident that Aya-101 struggles with score-based sentiment classification. In contrast, GPT-4 does not face the same difficulty level, indicating a more robust ability to manage such tasks effectively.

Performance Disparity in Complex vs. Simplistic Classification Tasks In our experiment with sentiment classification and dialect identification, we observe that LLMs struggle with extreme multi-label classification using only in-context learning (ICL). This is largely due to label variation and the challenges of intensity-score-based evaluation. These factors result in performance gaps between different LLMs.

In contrast, we see superior performance from LLMs in natural language inference (NLI) and topic classification tasks. These tasks are also classification-based, but they are simpler. NLI has three classes, and topic classification involves seven topic classes. As a result, LLMs perform well and significantly surpass supervised encoder fine-tuning for low-resource languages such as Central Kurdish and Sotho dialects. The variety understanding gap becomes less apparent due to the LLMs' robust ability to handle simpler classification tasks effectively.

Machine Reading Comprehension: A Challenge for Fine-Tuned Encoder Models This task consists of a question, a context passage, and four answer options. For supervised fine-tuning with encoder models, each option was appended to the question and context, treating the task as a four-class classification problem. This setup led to suboptimal performance for fine-tuned encoder models. In contrast, both Aya-101 and GPT-4 performed moderately well with just in-context learning, similar to their success in topic classification and natural language inference (NLI). This improved performance can be attributed to the fact that LLMs can leverage their superior text-understanding capabilities to read the context, interpret the question, and select the correct answer, making the MRC task relatively easier for them.

LLMs Often Struggle With Complex Instruction Following and Output Formatting The task of Parts of Speech (POS) tagging uses a simple token classification setup for fine-tuning encoder-based models. However, transforming this task into an in-context learning scenario requires moderately complex instructions, including detailed descriptions of token tags, input formats, and output formats. When evaluating zero-shot performance, where encoder-based models are fine-tuned on English and LLMs are prompted with three-shot examples, GPT-4 outperforms the other models. In contrast, Aya-101, despite being a multilingual model, falls significantly behind. A deeper investigation reveals that Aya-101 struggles to consistently follow complex instructions and often fails to properly format the output, which contributes to its poor performance.

Interestingly, Aya-101 performs the best in the extractive question answering (QA) task, surpassing GPT-4. Surprisingly, GPT-4 also scores lower compared to smaller encoder-based models. Upon investigation, we find that, as with the POS tagging task, output formatting issues contribute to this discrepancy. Extractive QA with encoder-based models involves retrieving an answer span from the given context. To emulate this scenario for generative models, we instructed both Aya-101 and GPT-4 to provide only the specific answer from the given context. While Aya-101 adhered strictly to the instructions, GPT-4 often included additional tokens or information, resulting in subpar performance when evaluated under the same criteria as the other models.

LLMs Struggle With Dialect Identification In encoder-based models, dialect identification is generally approached as a supervised classification task, where the model is fine-tuned on labeled dialectal sentences and tasked with predicting the correct dialect class for each input sentence during evaluation. To adapt this setup for generative LLMs, we provided each model with at least one example sentence paired with its dialectal label, then asked the model to classify additional sentences. However, this method did not yield results comparable to those achieved by fine-tuned encoder models. On average, GPT-4 performed better than Aya-101, though this may be influenced by data contamination, as GPT-4 could have had prior exposure to some of the labeled datasets. Despite these advantages, generative models still struggled significantly with city-level Arabic dialect classification, failing to accurately identify the dialects in most cases.

The primary reason for this failure lies in the limitations of extreme multi-label classification when relying solely on in-context learning (ICL). Unlike tasks such as common-sense reasoning or sentiment analysis—where ICL has shown success in identifying familiar, intuitive categories—dialect classification requires distinguishing between subtle, complex labels that demand a deeper understanding of linguistic differences. As a result, using only ICL for this task proves suboptimal, as it lacks the structure and specificity necessary to accurately classify fine-grained dialectal variations. Prior research has demonstrated that a combination of candidate shortlisting with re-ranking (Zhu and Zamani, 2024) or the use of retriever-based models (D'Oosterlinck et al., 2024) is more effective. Given the task's complexity—26 distinct Arabic dialect classes—simply providing class labels and a single example per class proved insufficient for accurate identification.

5 Investigating Dialect Identification Failure

Including Explanation-Prompt Yields No Improvement To investigate further the challenges faced by LLMs in dialect identification task, we conducted an ablation study on prompt-engineering to improve dialect identification performance. The experiment involved presenting varying numbers of example sentences n=(1, 3, 10, 30, and 50 examples) per city-level dialect to GPT-4 and subsequently prompting it to generate refined instructions for the classification task (presented in Fig. 2). We then used these refined prompts to evaluate the performance of Aya-101. Table 5 presents the results of this prompt refinement study. Despite the iterative refinement process, the overall results did not show significant improvements. The highest

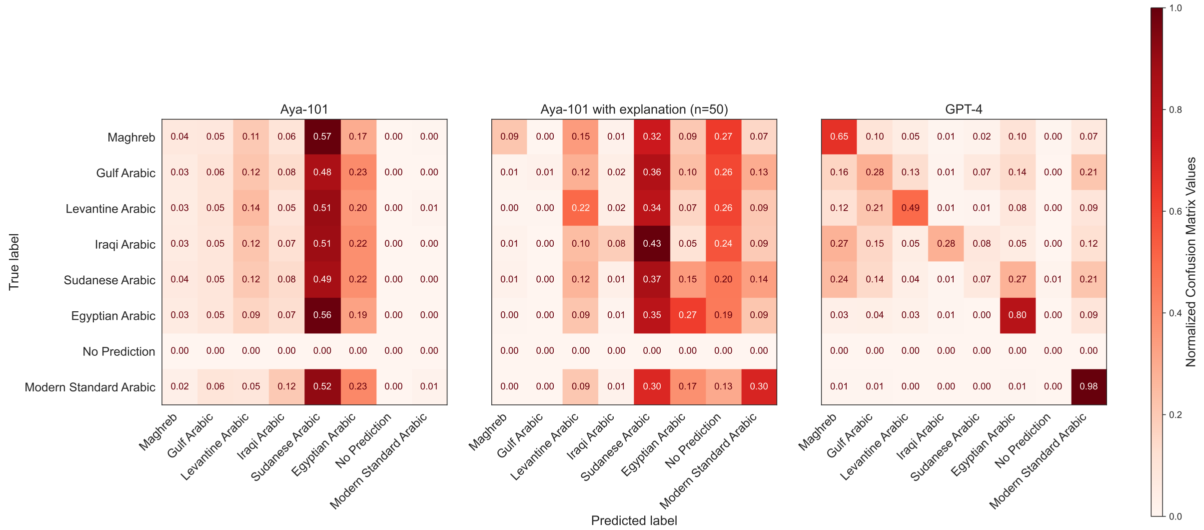

Figure 1: Confusion matrices for Arabic dialect classification across two LLMs: Aya-101 (prompting with one example per class as well as with additional explanation) and GPT-4. Here 26 city-level dialects are grouped into seven regional categories, providing a high-level view of model misclassifications and within-group confusions. Notably, Aya-101 shows a strong bias toward predicting Sudanese Arabic regardless of the true label, while the addition of explanation in the prompt reduces misclassification but introduces some "No Prediction" responses. GPT-4 demonstrates more balanced performance, with fewer confusions across dialect groups.

score was achieved with the "n=30" setup, which showed only a marginal improvement in F1 score. While most dialects exhibited limited gains, there were some exceptions, such as Rabat-Casablanca and Modern Standard Arabic (MSA) showed a slight increase in accuracy when more examples were provided. For instance, the score for MSA reached up to 17.0 with n=30, highlighting that some dialects might benefit from increased exposure during prompt refinement. This also suggests that the relatively better performance for these varieties might be attributed to Aya-101's prior exposure or broader representation of these dialects in the training data.

Nevertheless, the performance of LLMs for dialect identification remains inadequate, especially when relying solely on ICL for a large number of dialect classes.

Aya-101's Strong Bias Toward Sudanese Arabic In our initial setup, we began with a detailed set of 26 city-level Arabic dialects. To simplify analysis and improve model interpretability, we grouped these dialects into broader regional categories, such as Maghreb, Gulf Arabic, Levantine Arabic, and Egyptian Arabic, as reported in Table 7. This grouping provides a clearer perspective on how models handle regional dialect distinctions rather than granular city-level variations, allowing us to assess the models' generalization capabilities across similar dialects. Upon grouping the dialect classes, we visualized the confusion matrices for Aya-101, Aya-101 with explanation (n=50), and GPT-4 in Fig. 1.

We observe, Aya-101, without additional explanations, exhibits a strong tendency to misclassify a wide range of dialects as Sudanese Arabic, despite Sudanese Arabic representing only a small fraction (200 instances) of the dataset. This misclassification does not align with the true label distribution, where Maghreb (1400 instances), Gulf Arabic (1200), and Levantine Arabic (1000) are among the most represented dialects. Aya-101's errors are predominantly concentrated within Maghreb and Gulf Arabic groups, leading to a significant over-prediction of Sudanese Arabic.

When provided with a longer prompt including additional explanations, Aya-101 demonstrates improved differentiation, particularly in distinguishing Levantine and Egyptian Arabic from other groups. However, this extended prompting introduces a new issue: a portion of predictions are left blank, marked as "No Prediction", indicating instances where Aya-101 fails to respond with a specific classification. This is a significant limitation, as such non-responses reduce the model's effective prediction rate. Furthermore, Aya-101 continues to show substantial within-group confusion, especially among dialects within the Gulf and

Variety	(n-shot) n=1	With Explanation (n-shot) n=1	n=2	n=10	n=30	n=50
aleppo	2.9	3.0	5.0	7.0	6.0	6.0
algerian	0.0	0.0	1.0	11.0	4.0	2.0
ara. peninsula (a:yemen)	0.0	0.0	4.0	1.0	3.0	0.0
egyptian (a:alx)	0.0	2.0	0.0	0.0	0.0	0.0
egyptian (a:asw)	0.9	1.0	3.0	0.0	0.0	0.0
egyptian (a:cai)	6.4	7.0	0.0	11.0	13.0	12.0
egyptian (a:kha)	6.8	7.0	7.0	8.0	7.0	8.0
fez. meknes	0.7	4.0	1.0	8.0	4.0	0.0
gilit mesop.	4.8	4.0	9.0	5.0	6.0	3.0
gulf (a:doh)	4.0	4.0	0.0	4.0	4.0	0.0
gulf (a:jed)	1.5	8.0	12.0	8.0	0.0	3.0
gulf (a:mus)	0.0	0.0	3.0	0.0	0.0	0.0
gulf (a:riy)	2.5	0.0	0.0	0.0	0.0	0.0
levan. (a:north-dam)	2.7	6.0	10.0	7.0	7.0	10.0
libyan (a:ben)	1.6	0.0	0.0	0.0	2.0	3.0
north mesop. (a:bas)	1.0	0.0	0.0	0.0	0.0	0.0
north mesop. (a:mos)	0.0	2.0	0.0	8.0	20.0	0.0
rabat-casablanca	0.9	1.0	2.0	13.0	24.0	23.0
sfax	6.8	3.0	8.0	8.0	3.0	9.0
s. levan. (a:south-amm)	1.7	0.0	1.0	2.0	0.0	0.0
s. levan. (a:south-jer)	5.4	1.0	1.0	2.0	3.0	1.0
s. levan. (a:south-sal)	0.0	1.0	4.0	0.0	1.0	0.0
standard	1.9	11.0	16.0	11.0	17.0	14.0
sunni beiruti	5.0	1.0	1.0	14.0	14.0	14.0
tripolitanian	0.0	0.0	0.0	2.0	3.0	9.0
tunisian (a:tun)	1.0	3.0	9.0	16.0	6.0	0.0
Avg.	2.2	2.7	3.7	5.6	5.7	4.5

Table 5: Dialect Identification Results for Aya-101 with GPT-4 Explanation-Prompting. This table presents the F1 scores for dialect identification using Aya-101, where the model was prompted with explanations generated by GPT-4. The explanations were provided with varying numbers of examples (n-shots), from 1 to 50, for each dialect. The average F1 score across dialects is shown at the bottom, indicating limited improvements with increased examples.

Maghreb regions, even with additional explanation.

In comparison, GPT-4 demonstrates the most robust performance across dialects. It closely aligns with the true label distribution and shows higher accuracy in identifying key groups such as Maghreb, Levantine Arabic, and Modern Standard Arabic. Although GPT-4 still exhibits within-group misclassification—such as confusing Gulf Arabic with Iraqi Arabic—it effectively differentiates between dialects overall. This indicates that, while longer prompts with explanations enhance Aya-101's performance to some extent, GPT-4's inherent understanding of dialectal distinctions remains significantly stronger.

6 Related Work

The evaluation of language models has been a critical component in advancing natural language processing (NLP). Evaluation benchmarks are necessary to provide standardized, reproducible comparisons across models, ensuring that improvements in architecture or training result in tangible performance gains on a variety of tasks (Wang et al., 2018). Popular benchmarks such as XTREME (Hu et al., 2020) and GLUE (Wang et al., 2018) are designed to assess models, primarily focusing on standard language varieties and tasks like text classification and structural prediction.

With the development of large language models (LLMs), recent benchmarks have expanded to include reasoning capabilities and expert domain knowledge. Examples include benchmarks like SuperGLUE (Wang et al., 2019), BigBench (Srivastava et al., 2023), and MMLU (Hendrycks et al., 2021), which evaluate models on complex reasoning, knowledge-intensive tasks, and multi-domain expertise. These benchmarks are increasingly multilingual, but they still largely overlook dialectal and non-standard language varieties across diverse tasks.

Efforts in dialectal NLP have emerged, such as the Arabic dialect corpus MADAR (Bouamor et al., 2018) and resources developed through the VARDIAL workshop (Scherrer et al., 2024), such as DSL-TL (Zampieri et al., 2023) and Dialect-COPA (Ljubešić et al., 2024). However, these datasets remain largely scattered, and no unified benchmark exists to comprehensively evaluate language models on dialectal and non-standard varieties across multiple languages and tasks. DIALECTBENCH (Faisal et al., 2024) attempts to address this by aggregating dialectal datasets using a standardized approach with Glottocode mapping for language clusters and varieties. However, it primarily evaluates smaller encoder models and does not comprehensively explore dialectal tasks using recent advancements in large language models. Structured studies that leverage LLMs to evaluate a broad range of dialectal tasks remain largely unexplored.

7 Conclusion

In this study, we evaluated the performance of encoder-based models and LLMs on various dialect-specific NLP tasks. Our results indicate that while LLMs such as GPT-4 and Aya-101 excel in tasks like topic classification and natural language inference, they struggle with complex instructions and formatting, particularly in Parts of Speech (POS) tagging and dialect identification. In contrast, fine-tuned encoder models outperform LLMs in highly structured tasks such as POS tagging and extractive question answering. These findings suggest that while LLMs have potential, task-specific fine-tuning or hybrid approaches are still necessary for effectively handling nuanced, low-resource dialects.

Limitations

This study examines a limited selection of LLMs (one closed-weight and one open-weight) and solely relies on datasets provided by DIALECTBENCH.

Acknowledgements

The authors are thankful to the anonymous reviewers for their feedback. This work has been supported by the National Science Foundation under awards IIS-2125466 and IIS-2327143, and by the Microsoft Accelerate Foundation Models Research (AFMR) grant program. Computational resources for experiments were provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631). AA is also partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

Adel Abdelli, Fayçal Guerrouf, Okba Tibermacine, and Belkacem Abdelli. 2019. Sentiment analysis of Arabic Algerian dialect using a supervised method. In *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*, pages 1–6.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *Preprint*, arXiv:2309.07445.

Marwan Al Omari, Moustafa Al-Hajj, Nacereddine Hammami, and Amani Sabra. 2019. Sentiment classifier: Logistic regression for Arabic services' reviews in Lebanon. In *2019 International Conference on Computer and Information Sciences (ICCIS)*, pages 1–5.

Khaled Mohammad Alomari, Hatem M. ElSherif, and Khaled Shaalan. 2017. Arabic tweets sentimental analysis using machine learning. In *Advances in Artificial Intelligence: From Theory to Practice*, pages 602–610, Cham. Springer International Publishing.

Dhuha Alqahtani, Lama Alzahrani, Maram Bahareth, Nora Alshameri, Hend Al-Khalifa, and Luluh Aldhubayi. 2022. Customer sentiments toward Saudi banks during the Covid-19 pandemic. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 251–257, Trento, Italy. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. 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, Bangkok, Thailand. 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 Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Karel D'Oosterlinck, Omar Khattab, François Remy, Thomas Demeester, Chris Develder, and Christopher Potts. 2024. In-context learning for extreme multi-label classification. *Preprint*, arXiv:2401.12178.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chayma Fourati, Hatem Haddad, Abir Messaoudi, Moez BenHajhmida, Aymen Ben Elhaj Mabrouk, and Malek Naski. 2021. Introducing a large Tunisian Arabizi dialectal dataset for sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 226–230, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Moncef Garouani and Jamal Kharroubi. 2022. MAC: An open and free Moroccan Arabic corpus for sentiment analysis. In *Innovations in Smart Cities Applications Volume 5*, pages 849–858, Cham. Springer International Publishing.

Harald Hammarström and Robert Forkel. 2022. Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web Journal*, 13(6):917–924.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019. Discriminating between Mandarin Chinese and Swiss-German varieties using adaptive language models. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–187, Ann Arbor, Michigan. Association for Computational Linguistics.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *ArXiv*, abs/2401.05632.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024. DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to South Slavic dialects. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 89–98, Mexico City, Mexico. Association for Computational Linguistics.

Salima Medhaffar, Fethi Bougares, Yannick Estève, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 55–61, Valencia, Spain. Association for Computational Linguistics.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. ASTD: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. OpenAI technical report. Available at `https://openai.com/research/gpt-4`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hanna Sababa and Athena Stassopoulou. 2018. A classifier to distinguish between Cypriot Greek and Standard Modern Greek. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 251–255.

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Marcos Zampieri, Preslav Nakov, and Jörg Tiedemann, editors. 2024. *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*. Association for Computational Linguistics, Mexico City, Mexico.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy

Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xi-

aoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal Dependencies parsing for colloquial Singaporean English. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1732–1744, Vancouver, Canada. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language variety identification with true labels. *Preprint*, arXiv:2303.01490.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Ajede, and et al. 2021. Universal Dependencies 2.9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Yaxin Zhu and Hamed Zamani. 2024. ICXML: An in-context learning framework for zero-shot extreme multi-label classification. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2086–2098, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

A Task-Specific In-Context Learning Prompts

A.1 Parts of Speech Tagging (POS)

```
Instruction:
Given a sentence as space-separated tokens, predict the Part of Speech
    ↪ (PoS) tags for each token. You will need to use the tags defined
    ↪ below:
TAGS: ['NOUN', 'PUNCT', 'ADP', 'NUM', 'SYM', 'SCONJ', 'ADJ', 'PART',
    ↪ 'DET', 'CCONJ', 'PROPN', 'PRON', 'X', '_', 'ADV', 'INTJ', 'VERB',
    ↪ 'AUX', 'CONJ', 'root']

Input format:
Sentence: [space-separated tokens]
Output format:
1    [token1]    [predicted_tag1]
2    [token2]    [predicted_tag2]
...
n    [tokenn]    [predicted_tagn]

Input:
Sentence: {sentence}
Output: <entities to predict>
```

A.2 Natural Language Inference (NLI)

```
Instruction:
Given a premise and a hypothesis, determine the relationship between them.
The possible relationships are:
- Entailment: The hypothesis follows logically from the premise.
- Neutral: The hypothesis may or may not be true given the premise.
- Contradiction: The hypothesis contradicts or is inconsistent with the
    ↪ premise.

Premise: {premise}
Hypothesis: {hypothesis}
Relationship: <relation to predict>
```

A.3 Sentiment Analysis (SA)

```
Instruction:
Given a sentence, predict its sentiment as either {sentiment labels}

Sentence: {input_sentence}
Sentiment: <sentiment to predict>
```

A.4 Topic Classification (TC)

```
Instruction:
Given a sentence, predict its topic from one of the following categories:
    ↪ <topic classes>

Sentence: {sentence}
Topic: <topic to predict>
```

A.5 Extractive QA (EQA)

```
Instruction:
Given a context and a question, provide an answer to the question based
    ↪ on the information in the context.
The answer should be a span of text extracted directly from the context.
If the context does not contain enough information to answer the
    ↪ question, respond with "No answer".
Answer as concisely as possible in the same format as the examples below:

Context: {context}
Question: {question}
Answer: <answer to predict>
```

A.6 Dialect Identification (DID)

A.6.1 Standard

```
Instruction:
Given a sentence, predict in which dialect it is written. The options
    ↪ are: {dialect classes}

Sentence: {input_sentence}
Dialect: <dialect to predict>
```

A.6.2 GPT4-Refined Prompt from 50 Examples

In Fig. 2, we present the dialect markers obtained through prompting GPT-4 with 50 instances per Arabic dialect class. We utilize these dialect markers to design our prompt for dialect identification using Aya-101.

Dialect-Specific Markers:

- KHA (Khartoum): Sudanese Arabic featuring "داير" (want), local terms like "متين" (when), and polite formal requests.
- RAB (Rabat): Moroccan Arabic using "عافاك" (please), "بغيت" (want), and intricate negotiation-related terms.
- ALG (Algiers): Algerian Arabic marked by "واش" (what), French terms like "شحال" (how much), and mixed linguistic patterns.
- JED (Jeddah): Hejazi Arabic with "أبغا" (want), "فين" (where), and hospitality-driven expressions.
- CAI (Cairo): Egyptian Arabic with "عايز" (want), "فين" (where), and humor-tinged colloquialisms.
- MOS (Mosul): Iraqi Arabic with "چ" (ch sound), "گ" (g sound), and local vocabulary.
- ALE (Aleppo): Northern Syrian Arabic with "بدي" (I want), "قديش" (how much), and Turkish loanwords.
- SFX (Sfax): Tunisian Arabic featuring "باش" (will), "نحب" (want), and French-infused expressions.
- BEN (Benghazi): Libyan Arabic with "شن" (what), "توا" (now), and "نبي" (want).
- BAG (Baghdad): Central Iraqi Arabic marked by "شلون" (how), "ماكو" (none), and pronounced local pronunciation.
- RIY (Riyadh): Najdi dialect using "وش" (what), "تبغى" (want), and direct, formal phrasing.
- BEI (Beirut): Lebanese Arabic with "عم" (progressive), "إزا" (if), and blended French and English terms.
- MSA (Modern Standard Arabic): Formal Arabic used in media, academic, and professional settings.
- ASW (Aswan): Upper Egyptian Arabic with distinct local expressions and tonal shifts.
- TRI (Tripoli): Libyan Arabic with "قداش" (how much), "نبي" (want), and negotiation-focused terms.
- FES (Fes): Moroccan Arabic marked by negotiation and politeness nuances.
- BAS (Basra): Southern Iraqi Arabic with a softer pronunciation, using "اكو" and "ماكو".
- MUS (Muscat): Omani Arabic featuring formal and polite phrases like "أبغا" (want) and "يصير" (can).
- TUN (Tunis): Tunisian Arabic with French influences and context-sensitive terms.
- JER (Jerusalem): Palestinian Arabic using "بدي" (want), melodic intonations, and social context markers.
- SAL (Salalah): Southern Omani Arabic using "قديش" (how much), and distinctive phrasing.
- AMM (Amman): Jordanian Arabic with more formal Levantine tones.
- ALX (Alexandria): Egyptian Arabic with humor-infused phrases and local twists.
- DAM (Damascus): Syrian Arabic using "بدك" (you want), formal phrasing, and softer intonations.
- DOH (Doha): Qatari Arabic using "بغيت" (want), and Gulf-inflected vocabulary.
- SAN (Sanaa): Yemeni Arabic with unique local references and vocabulary.

Options: SAN, ALX, JED, RIY, ALG, BAG, DAM, BEN, BEI, RAB, AMM, JER, MUS, SFX, TUN, MOS, FES, CAI, DOH, TRI, KHA, ALE, BAS, MSA, ASW, SAL.

Question: Given the unique features of each dialect, identify which one matches the sentence below.

Figure 2: Dialect markers generated by GPT-4 for different Arabic dialects based on vocabulary, pronunciation, grammar, and cultural context, intended to assist in dialect identification tasks.

A.7 Machine-Reading Comprehension (MRC)

```
Instruction:
Given a passage and a question, select the correct answer from the
    ↪ provided options. Read the passage carefully and choose the option
    ↪ that best answers the question based on the information given in
    ↪ the passage. Answer as concisely as possible in the same format as
    ↪ the examples below:

Passage: {flores_passage}
Question: {question}
Options:
1. {answer1}
2. {answer2}
3. {answer3}
4. {answer4}
Answer: <answer to predict>
```

B Clusters and Varieties

Table 6: Language clusters and varieties.

Lang-group	Variety	Count
albanian	albanian gheg albanian	2
anglic	philippine english english (a:scotland) southeast american english indian english (a:north) north american english australian english english southern african english nigerian english kenyan english new zealand english english (a:uk) indian english (a:south) singlish irish english	15
arabic	libyan arabic (a:ben) aleppo south levantine arabic (a:south-jer) arabian peninsula arabic (a:yemen) south levantine arabic (a:south-amm) ta'izzi-adeni arabic north mesopotamian arabic levantine arabic (a:north) najdi arabic north mesopotamian arabic (a:bas) gulf arabic (a:jed) south levantine arabic (a:south-sal) gulf arabic (a:mus) tunisian arabic standard arabic fez. meknes algerian arabic levantine arabic (a:north-dam) arabic (a:bahrain) egyptian arabic (a:kha) south levantine arabic tripolitanian arabic egyptian arabic (a:alx) arabic (a:saudi-arabia) sunni beiruti arabic moroccan arabic gulf arabic (a:doh) rabat-casablanca arabic tunisian arabic (a:tun) egyptian arabic sfax arabic (a:jordan) gilit mesopotamian arabic gulf arabic (a:riy) tunisian arabic (r:casual, o:latin) north mesopotamian arabic (a:mos) egyptian arabic (a:asw) north african arabic egyptian arabic (a:cai)	39
bengali	vanga (a:dhaka) vanga (a:west bengal)	2
common turkic	south azerbaijani central oghuz (m:spoken) north azerbaijani	3
eastern-western armenian	eastern armenian western armenian	2
gallo-italian	ligurian venetian lombard	3

Table continued on next page.

Table continued on next page.

Lang-group	**Variety**	**Count**
gallo-rhaetian	french (a:paris) friulian old french (842-ca. 1400) french	4
greek	cypriot greek (r:casual, m:written, i:twitter) modern greek (r:casual, m:written, i:twitter) cypriot greek (r:casual, m:written, i:other)	3
high german	luxemburgish central alemannic (a:bs) central alemannic (a:be) german central alemannic (a:zh) central alemannic (a:lu) limburgan	7
italian romance	italian (r:formal, m:written, i:essay) sicilian italian continental southern italian italian (r:casual, m:written, i:tweet)	5
komi	komi-zyrian (m:spoken) komi-zyrian (m:written) komi-permyak	3
korean	korean (a:south-eastern, m:spoken) seoul (m:spoken)	2
kurdish	central kurdish northern kurdish	2
latvian	latvian east latvian	2
neva	finnish estonian	2
norwegian	norwegian bokmål (m:written) norwegian nynorsk (m:written) norwegian nynorsk (m:written, i:old)	3
saami	skolt saami north saami	2
sinitic	mandarin chinese (a:mainland, o:simplified) mandarin chinese (a:taiwan, o:simplified) classical chinese classical-middle-modern sinitic (a:hongkong, o:traditional) classical-middle-modern sinitic (o:traditional) mandarin chinese (a:taiwan, o:traditional, i:synthetic) cantonese classical-middle-modern sinitic (o:simplified) mandarin chinese (a:mainland, o:traditional, i:synthetic)	9
sotho-tswana (s.30)	southern sotho northern sotho	2
southwestern shifted romance	portuguese (i:mix) spanish portuguese (m:written) occitan portuguese (a:european) spanish (a:europe) latin american spanish galician brazilian portuguese	9
swahili	swahili (a:tanzania) swahili (a:kenya)	2
tupi-guarani subgroup i.a	mbyá guaraní (a:paraguay) mbyá guaraní (a:brazil) old guarani	3
	Total	**126** varieties in **23** clusters

Table 6: Language clusters and varieties.

C Arabic Dialect Identification Grouped Classes

Group	Region/Influence	Dialects
Maghreb (North African Arabic)	Morocco, Algeria, Tunisia, Libya	RAB (Rabat), FES (Fes), ALG (Algiers), TUN (Tunis), SFX (Sfax), BEN (Benghazi), TRI (Tripoli)
Egyptian Arabic	Egypt	CAI (Cairo), ALX (Alexandria), ASW (Aswan)
Levantine Arabic	Lebanon, Palestine, Syria, Jordan	BEI (Beirut), JER (Jerusalem), DAM (Damascus), ALE (Aleppo), AMM (Amman)
Gulf Arabic	Arabian Peninsula	RIY (Riyadh), JED (Jeddah), DOH (Doha), MUS (Muscat), SAL (Salalah), SAN (Sanaa)
Iraqi Arabic	Iraq	BAG (Baghdad), BAS (Basra), MOS (Mosul)
Sudanese Arabic	Sudan	KHA (Khartoum)
Modern Standard Arabic (MSA)	Pan-Arab	MSA (Modern Standard Arabic)

Table 7: Grouped Regional Classes for Arabic Dialects Based on Linguistic and Cultural Similarities

For Arabic dialect identification, starting with an initial set of 26 city-level dialect labels, each representing a unique Arabic dialect from specific cities or regions, we aimed to simplify and organize these labels based on linguistic and cultural similarities. Recognizing that certain dialects share regional and linguistic traits, we grouped them into broader categories to provide a more manageable and insightful analysis as reported in Table 7. For instance, North African dialects like those in Morocco, Algeria, and Tunisia (RAB, ALG, TUN) share common influences, such as French loanwords and distinctive vocabulary, allowing us to consolidate them into a "Maghreb" category. Similarly, dialects from the Levant (Lebanon, Palestine, Syria, Jordan) and the Gulf region (Saudi Arabia, Oman, Qatar) exhibit shared linguistic features within their respective areas, making them natural groups.

D Task-Specific Results

D.1 Parts of Speech Tagging (POS)

The detailed results for the Parts of Speech tagging task, including performance metrics and analysis, are presented in Table 8.

D.2 Sentiment Analysis (SA)

The comprehensive results for the Sentiment Analysis task, showcasing model performance and evaluation, are provided in Table 9.

D.3 Dialect Identification (DID)

The results for the Dialect Identification task, highlighting key metrics and comparisons, can be found in Table 10.

Cluster	Variety	mBERT Eng FT	XLM-R Eng FT	GPT-4 Eng k-shot ICL	Aya-101 Eng k-shot ICL	$\Delta_{\text{LLM-enc}}$	$\Delta_{\text{closed-open}}$
albanian	albanian	75.80	84.41	0.00	9.51	-74.90	-9.51
	gheg albanian	48.96	55.84	56.37	11.36	0.53	45.01
anglic	english	96.41	97.16	87.76	22.86	-9.40	64.90
	singlish	76.27	77.55	78.91	24.16	1.35	54.75
arabic	south levantine arabic	51.99	61.84	74.61	20.36	12.77	54.26
	standard arabic	39.74	56.67	62.81	9.89	6.14	52.92
	north african arabic	28.30	26.01	24.03	16.62	-4.27	7.41
eastern-western armenian	eastern armenian	71.78	82.63	0.00	13.89	-68.75	-13.89
	western armenian	70.27	75.31	77.19	11.92	1.88	65.27
gallo-italian	ligurian	58.90	52.78	58.93	14.47	0.03	44.45
gallo-rhaetian	french	84.36	85.47	88.40	21.08	2.93	67.32
	french (a:paris)	81.37	82.77	87.69	15.07	4.92	72.61
	old french (842-ca. 1400)	64.70	59.41	72.93	21.85	8.23	51.07
high german	german	87.08	88.36	86.16	9.62	-2.20	76.53
	central alemannic (a:zh)	62.56	47.18	61.32	11.85	-1.24	49.47
italian romance	italian	81.09	83.12	0.00	11.61	-71.51	-11.61
	italian (r:formal, m:written, i:essay)	80.00	81.87	79.09	20.23	-2.79	58.86
	italian (r:casual, m:written, i:tweet)	73.71	76.45	76.89	20.93	0.45	55.96
	continental southern italian	30.00	57.14	76.19	0.00	19.05	76.19
komi	komi-zyrian (m:spoken)	41.25	46.66	49.17	13.37	2.51	35.80
	komi-permyak	29.52	43.67	47.16	15.87	3.49	31.29
	komi-zyrian (m:written)	20.40	35.12	37.55	13.37	2.43	24.18
neva	finnish	81.29	86.21	83.63	16.92	-2.57	66.71
	estonian	80.34	85.17	85.23	14.79	0.06	70.44
norwegian	norwegian bokmål (m:written)	88.53	89.55	88.12	21.85	-1.43	66.28
	norwegian nynorsk (m:written)	85.06	85.81	0.00	24.50	-61.32	-24.50
	norwegian nynorsk (m:written, i:old)	73.25	79.29	71.57	23.43	-7.73	48.13
saami	north saami	35.92	32.13	56.68	20.73	20.76	35.95
	skolt saami	20.26	34.15	41.95	12.11	7.80	29.84
sabellic	umbrian	11.90	5.44	0.00	3.44	-8.46	-3.44
sinitic	classical-middle-modern sinitic (a:hongkong, o:traditional)	68.99	35.49	78.19	20.78	9.20	57.41
	classical-middle-modern sinitic (o:simplified)	58.26	30.92	71.46	17.04	13.21	54.42
	classical chinese	35.80	20.85	40.33	30.73	4.53	9.59
southwestern shifted romance	portuguese (a:european)	80.08	81.38	80.36	19.30	-1.02	61.06
	brazilian portuguese	78.63	80.12	80.31	18.94	0.19	61.37
	portuguese (i:mix)	78.48	79.85	0.00	19.48	-60.37	-19.48
	portuguese (m:written)	76.19	78.76	78.53	11.43	-0.24	67.09
tupi-guarani subgroup i.a	mbyá guaraní (a:paraguay)	27.89	28.77	33.27	13.66	4.49	19.61
	old guarani	8.96	10.30	10.26	10.81	0.51	-0.55
	mbyá guaraní (a:brazil)	1.94	0.59	0.32	0.00	-1.61	0.32
west low german	west low german	69.65	54.93	75.94	10.07	6.29	65.87

Table 8: Comparison of **F1 scores** for Part-of-Speech (POS) tagging across various language clusters and varieties. We compare smaller, encoder-based models (mBERT and XLM-R) that were fine-tuned on English and evaluated on all available varieties, with closed-source LLM (GPT-4) and an open-weight multilingual LLM (Aya-101). For GPT-4 and Aya-101, we employed in-context learning with k=3 shots based on English examples.

Cluster	Variety	mBERT Combined FT	XLM-R Combined FT	GPT-4 Combined k-shot ICL	Aya-101 Combined k-shot ICL	$\Delta_{\text{LLM-enc}}$	$\Delta_{\text{closed-open}}$
arabic	tunisian arabic	94.55	94.61	86.95	77.66	-7.66	9.29
	algerian arabic	84.98	84.70	85.77	87.54	2.56	-1.77
	arabic (a:jordan)	82.96	89.07	91.30	92.41	3.34	-1.11
	arabic (a:saudi-arabia)	81.38	83.40	75.93	79.03	-4.37	-3.10
	tunisian arabic (r:casual, o:latin)	80.95	79.80	59.13	59.08	-21.82	0.05
	standard arabic	80.63	83.96	71.56	77.48	-6.48	-5.92
	moroccan arabic	78.08	77.41	61.65	71.10	-6.98	-9.45
	egyptian arabic	67.03	69.03	27.24	22.18	-41.79	5.06
	south levantine arabic	58.38	58.90	62.04	25.45	3.14	36.59
Average	Average	78.77	80.10	69.06	65.77	-8.90	3.29

Table 9: Comparison of **accuracy scores** for sentiment analysis task across various language clusters and varieties. We compare smaller, encoder-based models (mBERT and XLM-R) that were fine-tuned on supervised classification task, with closed-source LLM (GPT-4) and an open-weight multilingual LLM (Aya-101). For GPT-4 and Aya-101, we employed in-context learning with k=3 shots example per class based on the specific variety of examples.

D.4 Natural Language Inference (NLI)

Detailed results for the Natural Language Inference task, including accuracy and other metrics, are outlined in Table 11.

D.5 Topic Classification (TC)

The results for the Topic Classification task, along with an evaluation summary, are presented in Table 12.

D.6 Extractive QA (EQA)

Comprehensive results for the Extractive QA task, covering key performance measures, are provided in Table 13.

D.7 Machine-Reading Comprehension (MRC)

The results for the Machine-Reading Comprehension task, including detailed analysis, are summarized in Table 14.

Cluster	Variety	Support	mBERT Combined FT	XLM-R Combined FT	GPT-4 Combined k-shot ICL	Aya-101 Combined k-shot ICL	$\Delta_{\text{LLM-enc}}$	$\Delta_{\text{closed-open}}$
anglic	english (a:uk)	249	90.00	79.58	79.84	77.33	-10.16	2.51
	north american english	349	88.05	85.01	83.85	82.31	-4.20	1.54
arabic	aleppo	200	59.50	52.94	7.87	2.94	-51.63	4.93
	algerian arabic	272	66.95	64.06	38.91	0.00	-28.04	38.91
	arabian peninsula arabic (a:yemen)	177	64.19	56.06	0.00	0.00	-64.19	0.00
	egyptian arabic (a:alx)	192	71.94	70.45	0.00	0.00	-71.94	0.00
	egyptian arabic (a:asw)	221	53.21	48.26	0.00	0.92	-52.29	-0.92
	egyptian arabic (a:cai)	130	43.03	48.50	26.32	6.36	-22.19	19.95
	egyptian arabic (a:kha)	244	57.21	49.12	7.33	6.75	-49.88	0.58
	fez. meknes	196	60.61	57.91	10.96	0.73	-49.65	10.23
	gilit mesopotamian arabic	203	57.07	48.47	35.69	4.79	-21.38	30.91
	gulf arabic (a:doh)	205	49.38	44.50	7.21	3.97	-42.17	3.25
	gulf arabic (a:jed)	196	58.59	43.29	11.22	1.47	-47.36	9.75
	gulf arabic (a:mus)	178	40.74	45.83	0.00	0.00	-45.83	0.00
	gulf arabic (a:riy)	311	48.53	45.38	4.84	2.48	-43.69	2.36
	levantine arabic (a:north-dam)	148	43.10	31.21	0.00	2.68	-40.43	-2.68
	libyan arabic (a:ben)	238	51.60	50.00	0.94	1.59	-50.00	-0.65
	north mesopotamian arabic (a:bas)	186	51.30	43.70	0.95	0.99	-50.31	-0.04
	north mesopotamian arabic (a:mos)	188	73.71	69.65	11.16	0.00	-62.55	11.16
	rabat-casablanca arabic	153	56.66	48.19	42.57	0.92	-14.09	41.65
	sfax	215	60.24	55.13	11.11	6.78	-49.13	4.33
	south levantine arabic (a:south-amm)	177	42.97	35.26	12.79	1.66	-30.18	11.13
	south levantine arabic (a:south-jer)	202	48.26	43.42	5.00	5.41	-42.85	-0.41
	south levantine arabic (a:south-sal)	167	50.14	62.59	0.00	0.00	-62.59	0.00
	standard arabic	244	67.57	96.79	39.09	1.86	-57.70	37.23
	sunni beiruti arabic	192	59.18	59.32	25.31	4.96	-34.01	20.34
	tripolitanian arabic	201	65.84	60.15	0.00	0.00	-65.84	0.00
	tunisian arabic (a:tun)	164	57.69	44.71	41.60	1.00	-16.09	40.61
greek	cypriot greek (r:casual, m:written, i:other)	81	61.87	67.59	60.87	38.99	-6.72	21.88
	cypriot greek (r:casual, m:written, i:twitter)	36	56.79	54.05	48.57	38.71	-8.22	9.86
	modern greek (r:casual, m:written, i:twitter)	94	69.28	69.41	44.16	3.33	-25.26	40.82
high german	central alemannic (a:be)	389	72.04	56.48	30.71	0.00	-41.33	30.71
	central alemannic (a:bs)	340	74.67	59.44	33.09	17.41	-41.58	15.68
	central alemannic (a:lu)	335	74.19	62.17	42.18	0.57	-32.01	41.61
	central alemannic (a:zh)	359	77.27	68.19	35.13	38.72	-38.56	-3.59
sinitic	mandarin chinese (a:mainland, o:simplified)	986	98.59	93.30	67.51	66.51	-31.08	1.00
	mandarin chinese (a:mainland, o:traditional, i:synthetic)	977	97.93	93.88	67.24	66.71	-30.69	0.53
	mandarin chinese (a:taiwan, o:simplified)	1014	98.61	92.89	11.03	1.77	-87.58	9.26
	mandarin chinese (a:taiwan, o:traditional, i:synthetic)	1023	97.97	94.11	11.31	1.19	-86.67	10.12
southwestern shifted romance	brazilian portuguese	627	93.83	88.51	82.29	55.50	-11.54	26.78
	latin american spanish	207	84.79	16.80	61.33	54.81	-23.46	6.52
	portuguese (a:european)	349	79.61	72.46	65.27	51.00	-14.34	14.28
	portuguese (m:written)	15	17.45	0.00	2.98	1.60	-14.47	1.38
	spanish	290	77.63	58.16	8.89	41.63	-36.00	-32.74
	spanish (a:europe)	492	86.32	81.05	79.40	43.86	-6.92	35.54

Table 10: Results for the dialect identification task (**F1 scores**) across various language clusters and dialect varieties. The encoder-based models (mBERT and XLM-R) were fine-tuned separately on supervised classification tasks for each language cluster. In contrast, the closed-weight LLM (GPT-4) and the open-weight multilingual LLM (Aya-101) were evaluated using in-context learning with k=3 shot examples per class (with an exception of k=1 for Arabic clusters due to the larger number of varieties).

Cluster	Variety	mBERT Eng FT	XLM-R Eng FT	GPT-4 Eng k-shot ICL	Aya-101 Eng k-shot ICL	$\Delta_{\text{LLM-enc}}$	$\Delta_{\text{closed-open}}$
anglic	english	81.95	83.43	88.17	70.07	4.74	18.10
arabic	standard arabic	65.57	73.85	78.27	66.43	4.42	11.83
	najdi arabic	59.14	68.94	78.99	69.48	10.05	9.51
	ta'izzi-adeni arabic	58.64	68.62	74.26	66.51	5.64	7.75
	moroccan arabic	54.61	58.14	72.15	63.66	14.01	8.49
	egyptian arabic	53.86	65.70	77.94	63.78	12.24	14.16
	south levantine arabic	53.42	63.81	74.80	64.89	10.99	9.91
	north mesopotamian arabic	52.84	58.75	71.84	62.45	13.09	9.38
	levantine arabic (a:north)	51.40	61.31	75.55	64.14	14.24	11.42
	tunisian arabic	47.42	50.20	57.17	57.26	7.06	-0.09
common turkic	north azerbaijani	59.20	73.17	72.00	63.81	-1.17	8.20
	central oghuz (m:spoken)	58.37	74.52	78.78	65.59	4.25	13.19
	south azerbaijani	44.58	39.24	47.03	57.40	12.82	-10.36
gallo-italian	venetian	64.99	68.55	70.97	64.32	2.42	6.65
	lombard	59.34	56.16	66.77	63.60	7.44	3.18
	ligurian	56.70	57.16	53.39	61.73	4.57	-8.34
gallo-rhaetian	friulian	54.01	54.56	53.48	60.15	5.59	-6.67
high german	luxemburgish	60.01	46.21	69.21	66.34	9.20	2.86
	limburgan	50.31	59.75	65.44	56.44	5.69	9.00
italian romance	italian	73.71	78.19	76.06	69.06	-2.13	7.00
	sicilian	62.66	55.82	71.45	63.30	8.79	8.15
kurdish	central kurdish	37.40	39.59	57.35	63.37	23.78	-6.02
	northern kurdish	33.93	63.26	60.33	62.77	-0.49	-2.44
latvian	latvian	59.95	73.63	73.93	66.19	0.30	7.75
	east latvian	47.02	53.54	37.31	54.05	0.51	-16.74
modern dutch	dutch	71.77	76.45	81.95	68.20	5.50	13.75
norwegian	norwegian bokmål (m:written)	72.45	79.51	83.11	69.12	3.60	13.99
	norwegian nynorsk (m:written)	68.10	71.06	70.28	64.97	-0.78	5.31
sardo-corsican	sardinian	56.63	58.32	58.36	62.05	3.73	-3.69
sinitic	classical-middle-modern sinitic (o:simplified)	68.54	72.57	72.00	65.10	-0.57	6.90
	classical-middle-modern sinitic (o:traditional)	61.48	64.49	62.40	56.68	-2.10	5.72
	cantonese	60.27	67.41	64.08	63.50	-3.33	0.58
sotho-tswana (s.30)	northern sotho	35.06	35.98	55.33	60.11	24.13	-4.78
	southern sotho	34.62	34.16	48.44	61.31	26.69	-12.87
southwestern shifted romance	spanish	75.15	79.09	84.25	66.64	5.16	17.61
	portuguese (a:european)	73.73	79.22	84.95	64.53	5.73	20.42
	galician	73.39	78.55	78.48	68.50	-0.06	9.99
	occitan	68.47	62.96	73.15	57.28	4.68	15.87
Average	Average	58.44	63.31	68.93	63.55	6.59	5.39

Table 11: Results for the natural language inference (NLI) task. We compute **F1 scores** across various language clusters and dialect varieties. The encoder-based models (mBERT and XLM-R) were fine-tuned in Standard English and evaluated on all available varieties. In contrast, the closed-weight LLM (GPT-4) and the open-weight multilingual LLM (Aya-101) were evaluated using in-context learning with k=3 shot English examples.

Cluster	Variety	mBERT Eng FT	XLM-R Eng FT	mBERT Cluster-rep FT	XLM-R Cluster-rep FT	GPT-4 Eng k-shot ICL	GPT-4 Cluster-rep k-shot ICL	Aya-101 Eng k-shot ICL	Aya-101 Cluster-rep k-shot ICL	Δ_{LLM} -enc	Δ_{closed} -open
anglic	english	89.74	89.21	89.74	89.21	86.67	83.05	77.84	77.59	-3.07	8.83
arabic	standard arabic	85.25	83.96	86.71	82.27	87.40	88.73	79.17	78.57	2.01	9.55
	ta'izzi-adeni arabic	84.96	82.05	86.44	81.98	86.03	82.80	78.22	81.22	-0.41	4.81
	najdi arabic	84.80	84.39	87.41	83.33	85.35	85.51	80.53	80.44	-1.90	4.97
	north mesopotamian arabic	82.97	80.95	84.77	80.36	86.15	87.42	79.55	79.61	2.65	7.81
	south levantine arabic	81.82	80.16	84.16	79.05	86.67	83.53	80.81	80.59	2.50	5.86
	levantine arabic (a:north)	81.59	80.15	83.76	79.88	87.47	86.41	76.63	80.25	3.71	7.22
	egyptian arabic	81.02	76.38	84.43	81.03	87.34	83.09	82.53	78.93	2.91	4.81
	tunisian arabic	79.45	72.88	83.97	77.33	85.14	81.46	78.87	79.04	1.17	6.10
	moroccan arabic	73.87	79.14	78.76	78.55	87.58	87.70	80.68	79.95	8.56	7.02
common turkic	north azerbaijani	80.46	79.87	82.00	79.55	86.78	82.96	81.24	82.34	4.78	4.44
	central oghuz (m:spoken)	79.10	84.41	80.61	79.51	87.97	86.41	81.87	79.26	3.56	6.10
	south azerbaijani	65.90	67.08	69.71	68.37	77.86	74.65	74.23	83.27	13.56	-5.41
gallo-italian	venetian	76.72	70.68	75.07	74.28	85.98	81.70	77.50	77.09	9.26	8.47
	lombard	69.92	59.90	70.65	64.56	86.45	82.96	77.67	78.46	15.80	7.99
	ligurian	66.81	63.42	74.03	57.78	80.08	76.96	76.76	77.25	6.05	2.83
gallo-rhaetian	friulian	68.79	64.66	67.69	63.14	86.32	77.05	79.40	76.90	17.52	6.92
high german	luxemburgish	74.74	58.50	77.86	64.83	86.33	83.37	77.15	79.83	8.47	6.50
	limburgan	71.09	65.83	71.12	65.73	86.06	80.47	79.55	75.59	14.95	6.52
italian romance	italian	87.67	84.92	86.68	85.83	89.39	85.87	84.05	81.32	1.73	5.35
	sicilian	75.22	59.71	72.70	59.47	88.30	80.20	79.73	80.02	13.08	8.28
kurdish	northern kurdish	33.23	68.21	10.45	5.71	86.13	74.18	79.25	75.02	17.91	6.87
	central kurdish	13.10	19.37	16.86	12.38	75.54	78.22	76.37	77.61	58.85	0.61
latvian	latvian	76.35	83.75	80.63	82.80	87.15	86.46	76.95	81.52	3.40	5.64
	east latvian	55.67	65.02	63.69	67.42	79.68	72.95	78.05	75.60	12.26	1.63
modern dutch	dutch	88.97	83.37	89.55	84.51	85.99	85.05	79.89	81.11	-3.56	4.88
norwegian	norwegian nynorsk (m:written)	85.66	79.94	89.20	79.06	87.30	85.24	79.47	79.70	-1.90	7.60
	norwegian bokmål (m:written)	83.81	82.90	83.82	84.14	86.70	81.21	78.17	79.74	2.56	6.96
sardo-corsican	sardinian	71.03	66.89	69.65	62.49	84.40	79.15	79.72	81.22	13.37	3.19
sinitic	classical-middle-modern sinitic (o:traditional)	89.82	86.80	89.02	86.39	84.91	85.41	79.78	78.23	-4.41	5.63
	cantonese	89.45	86.46	88.71	87.64	85.46	83.99	77.90	79.63	-4.00	5.82
	classical-middle-modern sinitic (o:simplified)	88.74	86.38	88.86	89.15	85.64	84.36	74.74	80.21	-3.51	5.43
sotho-tswana (s.30)	northern sotho	35.62	28.16	34.86	13.55	72.19	70.28	78.87	79.01	43.39	-6.81
	southern sotho	32.55	32.31	39.93	19.08	72.23	70.45	74.79	75.15	35.22	-2.92
swe. shift. romance	portuguese (a:european)	88.13	89.10	88.10	87.74	86.31	84.97	77.94	81.35	-2.79	4.96
	galician	86.99	89.00	86.93	87.83	87.82	87.27	79.59	80.78	-1.19	7.04
	spanish	86.74	85.93	84.87	86.55	86.95	85.74	80.23	77.86	0.21	6.72
	occitan	84.12	74.80	78.53	62.56	84.12	80.51	79.34	77.80	-0.00	4.79
Average	Average	74.52	73.07	75.31	70.40	84.89	82.05	78.82	79.19	7.70	5.08

Table 12: Topic Classification (TC) task results, displaying **F1 scores** across different language clusters and dialect varieties. Encoder-based models (mBERT and XLM-R) were fine-tuned in either Standard English or a representative language of the target cluster and evaluated on all available varieties. In contrast, the closed-weight LLM (GPT-4) and open-weight multilingual LLM (Aya-101) were evaluated through in-context learning with 3-shot examples, either in English or the target variety.

Cluster	Variety	mBERT Combined FT	XLM-R Combined FT	mBERT Eng FT	XLM-R Eng FT	GPT-4 Combined k-shot ICL	Aya-101 Combined k-shot ICL	GPT-4 Eng k-shot ICL	Aya-101 Eng k-shot ICL	$\Delta_{\text{LLM-enc}}$	$\Delta_{\text{closed-open}}$
anglic	english (a:scotland)	76.38	70.34	71.82	63.15	56.94	74.23	64.11	72.07	-2.15	-10.12
	southern african english	76.66	71.18	71.49	63.87	59.66	73.40	60.89	73.65	-3.01	-12.76
	new zealand english	76.71	71.39	71.22	63.69	53.90	76.95	66.03	75.49	0.24	-10.92
	australian english	75.66	70.89	71.20	62.28	61.22	73.73	57.86	72.47	-1.93	-12.52
	southeast american english	77.26	71.50	71.17	63.71	63.35	76.46	62.46	76.31	-0.80	-13.10
	irish english	75.52	70.73	70.92	62.15	57.71	73.28	59.30	70.87	-2.24	-13.98
	philippine english	76.37	70.64	70.47	62.22	64.94	73.56	58.55	72.35	-2.81	-8.62
	nigerian english	73.61	68.33	69.10	61.27	59.01	67.68	57.63	67.04	-5.93	-8.67
	indian english (a:north)	74.62	68.03	68.84	61.25	54.62	68.13	60.46	69.24	-5.38	-8.78
	kenyan english	72.59	66.68	68.72	58.64	53.86	67.60	46.55	68.13	-4.46	-14.28
	indian english (a:south)	71.93	66.88	66.49	60.36	56.03	65.05	51.03	64.87	-6.88	-9.02
arabic	arabic (a:bahrain)	77.52	72.11	53.25	53.28	44.72	76.58	49.31	74.39	-0.94	-27.28
	arabic (a:jordan)	77.35	71.29	52.72	53.72	48.15	73.75	44.81	74.37	-2.98	-26.22
	arabic (a:saudi-arabia)	77.88	72.11	52.72	53.24	47.66	75.68	45.36	74.56	-2.20	-28.02
	algerian arabic	77.85	72.34	52.56	53.52	44.05	74.67	48.77	74.69	-3.16	-25.92
	tunisian arabic	76.72	71.64	52.28	52.94	42.52	73.67	54.13	73.09	-3.05	-19.54
	moroccan arabic	76.73	71.57	51.86	52.17	46.67	74.57	50.74	71.89	-2.16	-23.83
	egyptian arabic	76.53	70.75	51.80	51.99	44.10	72.93	41.43	73.32	-3.21	-29.22
bengali	vanga (a:west bengal)	68.62	73.27	32.30	36.39	54.69	87.44	49.66	85.58	14.17	-32.75
	vanga (a:dhaka)	67.37	74.24	31.79	35.52	55.13	84.99	59.58	84.64	10.75	-25.41
korean	seoul (m:spoken)	10.15	31.91	7.26	19.62	60.74	76.13	58.36	76.14	44.23	-15.40
	korean (a:south-eastern, m:spoken)	9.92	31.01	7.22	20.08	64.43	68.08	61.91	78.46	47.45	-14.03
swahili	swahili (a:tanzania)	63.54	62.30	38.24	39.38	48.19	59.30	38.64	56.85	-4.24	-11.10
	swahili (a:kenya)	72.25	70.53	37.97	41.59	49.88	67.42	39.46	66.76	-4.83	-17.55
Average	Average	69.16	67.15	53.89	51.92	53.84	73.14	53.63	72.80	2.27	-17.46

Table 13: Results for the Extractive Question Answering (EQA) task, showing **F1 scores** across various language clusters and dialect varieties. Encoder-based models (mBERT and XLM-R) were fine-tuned on Standard English or combined training data and evaluated on all available varieties. In contrast, the closed-weight LLM (GPT-4) and open-weight multilingual LLM (Aya-101) were assessed using in-context learning with 3-shot examples from English or the combined training data.

Cluster	Variety	mBERT Combined FT	XLM-R Combined FT	GPT-4 Combined k-shot ICL	Aya-101 Combined k-shot ICL	$\Delta_{\text{LLM-enc}}$	$\Delta_{\text{closed-open}}$
anglic	english	51.97	53.44	95.65	84.34	42.20	11.31
arabic	standard arabic	39.01	43.78	93.04	78.31	49.26	14.74
	levantine arabic (a:north)	38.64	40.71	81.02	71.04	40.32	9.98
	north mesopotamian arabic	37.99	41.35	78.55	63.72	37.20	14.83
	moroccan arabic	36.94	37.61	80.52	66.02	42.91	14.50
	egyptian arabic	36.21	37.98	88.59	70.38	50.61	18.21
	najdi arabic	36.05	38.16	85.12	71.47	46.96	13.66
sinitic	classical-middle-modern sinitic (o:simplified)	49.79	47.10	93.88	80.66	44.10	13.23
	classical-middle-modern sinitic (o:traditional)	46.88	44.76	93.07	76.89	46.19	16.19
sotho-tswana (s.30)	northern sotho	31.18	29.72	47.34	62.18	31.00	-14.85
	southern sotho	28.52	29.00	52.40	63.62	34.62	-11.21
Average	Average	39.38	40.33	80.84	71.69	42.31	9.14

Table 14: Results for the Machine Reading Comprehension (MRC) task, showing **F1 scores** across various language clusters and dialect varieties. Encoder-based models (mBERT and XLM-R) were fine-tuned on the combined training data and evaluated on all available varieties. Whereas, the closed-weight LLM (GPT-4) and open-weight multilingual LLM (Aya-101) were assessed using in-context learning with 3-shot examples drawn from similar data.