LAraBench

Benchmarking Arabic AI with Large Language Models

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine ElKheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, Firoj Alam

fialam@hbku.edu.qa

Study Design

Goal: Benchmark LLMs performance on Arabic AI and compare to SOTA models.

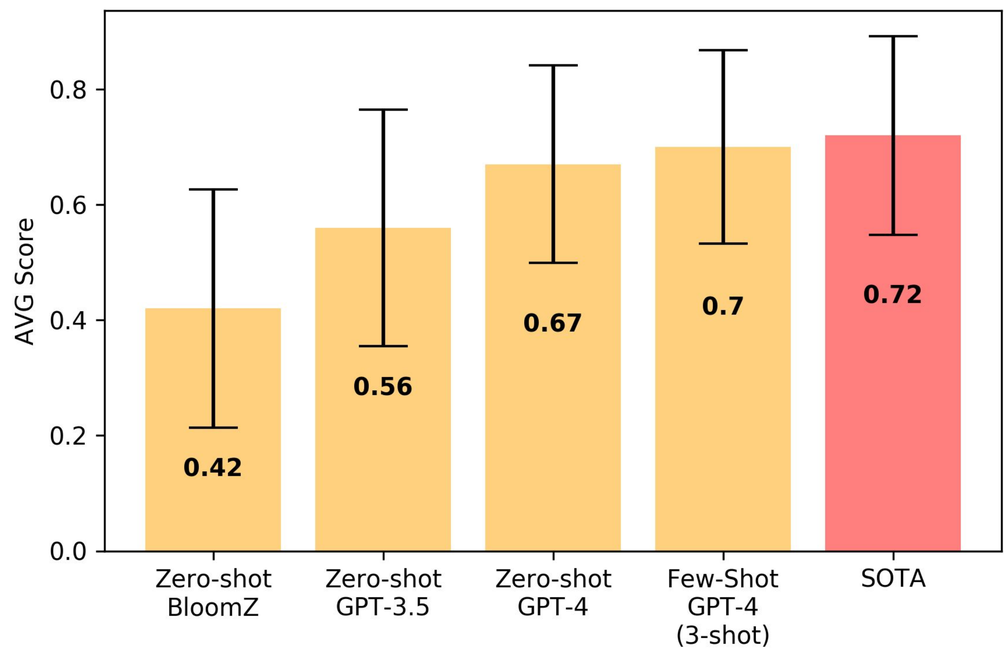
Modalities:

- **Speech Processing:** ASR, TTS
- **NLP tasks:** ranging from sequence tagging and content classification across different domains

TASKS	DATASETS	EVALUATION	MODELS
Word Segmentation, Syntax & Information Extraction (e.g., POS tagging)	XNLI	Accuracy	GPT-3.5
Factuality, Disinformation & Harmful Content Detection (e.g., Hate Speech & Propaganda Detection)	XGLUE	F1	GPT-4
Semantics (e.g., Semantic Textual Similarity and Natural Language Inference)	XQuAD	Macro-F1	BLOOMZ
Demographic & Protected Attributes (e.g., Gender and User Country Detection)	ASAD	Micro-F1	**LEARNING**
Sentiment, Stylistic & Emotion Analysis (e.g., Stance Detection, Sarcasm Detection)	Aqmar	Weighted-F1	Zero-shot
Machine Translation (e.g., English-Arabic and Arabic dialects)	SANAD	BLEU	Few-shot
News Categorization	MADAR	WER	
Question Answering	QASR	Pearson Correlation	
	WikiNews	Jaccard Similarity	
	Conll2006		
	ANERcorp		

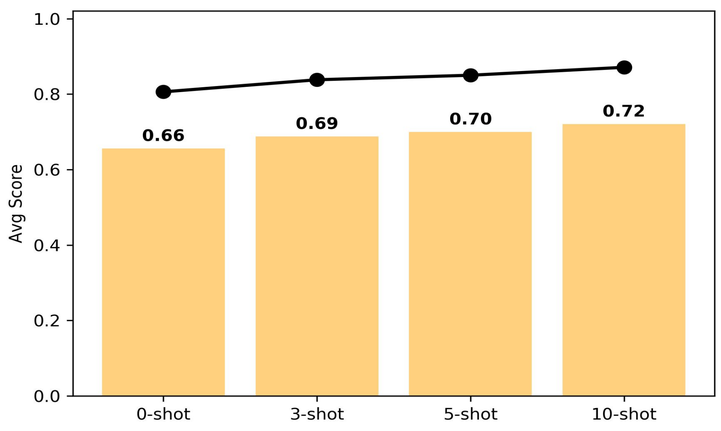
Findings

- GPT-4 outperforms other models in majority of the NLP tasks
- GPT-4 reduces performance gap with SOTA in the few-shot setting
- MSA vs Dialect: The gaps in LLMs' performance between MSA and dialectal datasets are more pronounced, indicating ineffectiveness of LLMs for under-represented dialects
- Patterns of errors in sequence tagging tasks like segmentation, POS tagging, and NER:
  - deviations in the output format
  - instances where responses included extra or omitted tokens
  - issues with generated output labels (Arabic instead of English)
- Models occasionally produced outputs that fell outside the predefined set of labels



Number of Shots

Few-shot results across seven different datasets



Task Name	Dataset	Metric	0-shot	3-shot	5-shot	10-shot
NER	ANERcorp	M-F1	0.355	0.420	0.426	0.451
Sentiment	ArSAS	M-F1	0.569	0.598	0.619	0.639
News Cat.	ASND	M-F1	0.667	0.594	0.674	0.723
Gender	Arap-Tweet	M-F1	0.868	0.980	0.931	0.937
Subjectivity	In-house	M-F1	0.677	0.745	0.740	0.771
XNLI (Ar)	XNLI	Acc	0.753	0.774	0.789	0.809
QA	ARCD	F1/EM	0.705	0.704	0.718	0.716
Average			**0.656**	**0.688**	**0.700**	**0.721**

Semantic vs. Syntactic Task Differences

- The Gap between SOTA and the three LLMs for POS (a syntactic task) is considerably larger than for MT (a semantic task)
- The gap is much lower for semantic tasks compared to syntactic tasks, on average, across the three LLMs

	BLOOMZ	GPT-3.5	GPT-4	SOTA
Semantic				
MT	19.38	24.09	23.57	24.58
Semantics (STS, XNLI)	0.615	0.733	0.827	0.794
Syntactic				
POS	-	0.154	0.464	0.844
Parsing	-	0.239	0.504	0.796

Native Language Prompts

We observed increased performance (1%) in three out of seven datasets compared to their counterparts with English prompts

Task Name	Metric	English	Arabic
NER	Macro-F1	0.355	0.350
Sentiment	Macro-F1	0.569	0.547
News Cat.	Macro-F1	0.667	0.739
Gender	Macro-F1	0.868	0.892
Subjectivity	Macro-F1	0.677	0.725
XNLI (Arabic)	Acc	0.753	0.740
QA	F1 (exact match)	0.705	0.654
Average		**0.656**	**0.664**

Speech Tasks

- Performance is heavily dependent on the models' parameters
- USM model performs comparably with SOTA for MSA
- Both models show a performance gap when dealing with dialects
- Fine tuning with 2 hours of speech improves the performance significantly

Dataset *dom./dial.*	Models	Zero-Shot	N-Shot (*2hrs*)	SOTA
MGB2 *Broadcast/MSA*	W.S	46.70	36.8	O: **11.4** S:11.9
	W.M	33.00	-	
	W.Lv2	26.20	18.8	
	USM	15.70	*N/A*	
MGB3 *Broadcast/EGY*	W.S	83.20	77.5	O: **21.4** S: 26.70
	W.M	65.90	-	
	W.Lv2	55.60	44.6	
	USM	22.10	*N/A*	
MGB5 *Broadcast/MOR*	W.S	135.20	114.6	O: **44.1** S:49.20
	W.M	116.90	-	
	W.Lv2	89.40	85.5	
	USM	51.20	*N/A*	
QASR.CS *Broadcast/Mixed*	W.S	63.60	-	O: **23.4** S: 24.90
	W.M	48.90	-	
	W.Lv2	37.90	31.2+	
	USM	27.80	*N/A*	
DACS *Broadcast /MSA-EGY*	W.S	61.90	-	O: 15.9 S: 21.3
	W.M	48.70	-	
	W.Lv2	34.20	30.4+	
	USM	**14.30**	*N/A*	
ESCWA.CS *Meeting/Mixed*	W.S	101.50	-	O: 49.8 S:48.00
	W.M	69.30	-	
	W.Lv2	60.00	53.6+	
	USM	**45.70**	*N/A*	
CallHome *Telephony/EGY*	W.S	155.90	152.9	O: **45.8*** S: 50.90
	W.M	113.70	-	
	W.Lv2	78.70	64.6	
	USM	54.20	*N/A*	

Acknowledgments: The contributions of Maram Hasanain were funded by the NPRP grant 14C-0916-210015, which is provided by the Qatar National Research Fund (a member of Qatar Foundation).