LAraBench: Benchmarking Arabic AI with Large Language Models

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs) have significantly influenced the landscape of language and speech research. Despite this progress, these models lack specific benchmarking against state-of-the-art (SOTA) models tailored to particular languages and tasks. *LAraBench* addresses this gap for Arabic Natural Language Processing (NLP) and Speech Processing tasks, including sequence tagging and content classification across different domains. We utilized models such as GPT-3.5-turbo, GPT4, BLOOMZ, Whisper, and USM, employing zero and few-shot learning techniques to tackle 33 distinct tasks across 61 publicly available datasets. This involved 98 experimental setups, encompassing ∼296K data points, ∼46 hours of speech, and 30 sentences for Text-to-Speech (TTS). This effort resulted in 330+ sets of experiments. Our analysis focused on measuring the performance gap between SOTA models and LLMs. The overarching trend observed was that SOTA models generally outperformed LLMs in zero-shot learning, with a few exceptions. Notably, larger computational models with few-shot learning techniques managed to reduce these performance gaps. Our findings provide valuable insights into the applicability of LLMs for Arabic NLP and speech processing tasks.

1 Introduction

Generative Pre-trained Transformer (GPT) models are examples of large language models (LLMs) trained on massive datasets and using hundreds of millions of parameters. Several LLMs have been recently released for use through APIs or pre-trained models and have demonstrated a high level of coherence in generating content in response to specific user tasks. However, quality assessments of released LLMs generally lack references to previous research and comparison with state-of-the-art (SOTA) methods that the research community has used for systematic evaluation and monitoring of scientific progress for various languages and tasks.

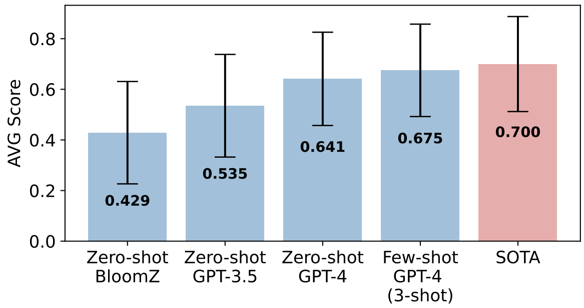

Figure 1: Average performance of the models as compared to SOTA across 21 unique NLP tasks and 31 testing setups.

Several research initiatives have evaluated these large models' performance on standard NLP and speech processing tasks. The HELM project (Liang et al., 2022) assessed English LLMs across various metrics and scenarios. BIG-Bench (Srivastava et al., 2022) introduced a large-scale evaluation with 214 tasks, including low-resource languages. GPT2.5 (Radford et al., 2019), ChatGPT (OpenAI, 2023), and BLOOM (Scao et al., 2022), were recently evaluated by Bang et al. (2023); Ahuja et al. (2023); Hendy et al. (2023); Khondaker et al. (2023). Large speech models such as Whisper (Radford et al., 2022) and Universal Speech Model (USM) (Zhang et al., 2023) were also explored for speech recognition and translation tasks. Initiatives such as SUPERB (Yang et al., 2021) were introduced to support benchmarking tools and leaderboards for several speech-related tasks. Bubeck et al. (2023) explored GPT-4's capabilities to determine if it surpasses mere memorization, possessing a profound and adaptable comprehension of concepts, skills, and domains. Their results indicate that GPT-4 demonstrates a higher level of general intelligence compared to its predecessors.

LAraBench study fulfills an important objective of assessing the LLMs capabilities for supporting

Arabic language processing tasks, for Modern Standard Arabic (MSA) and dialectal Arabic (DA), at the same level of depth and breadth as for English tasks. Our evaluation involves 61 publicly available datasets and 98 test setups used to perform and evaluate language processing tasks in both MSA and dialectal Arabic across various genres (e.g., news articles, tweets, meetings, telephony, and broadcast content). Our evaluation focuses on assessing the capabilities of GPT-3.5-turbo, GPT-4, and BLOOMZ (176B) for NLP tasks, and of Whisper (Large, 1.55B) and USM (2B) for Speech processing, in both zero and few-shot settings. We investigate: ***(i)*** *can LLMs effectively perform Arabic NLP and Speech processing tasks without prior task-specific knowledge (zero-shot)?* ***(ii)*** *how does performance vary across tasks with different complexities in zero- and few-shot settings?* ***(iii)*** *how do LLMs compare to current SOTA models, and are open LLMs as effective as the commercially available (closed) models?* Our investigation reveals unique insights about LLMs' performance on Arabic NLP and Speech tasks:

A. Zero-shot Multi-task Performer. GPT-4 outperforms GPT-3.5 and BLOOMZ in majority of the NLP tasks (see Figure 1). However, a large performance gap between GPT-4 and SOTA models remains due to the higher quality SOTA models. For speech tasks, USM outperforms all the Whisper variants and performs comparably with SOTA.

B. Few-shot and SOTA. GPT-4 reduces the performance gap with SOTA in the few-shot (only 3-shots) setting (see Figure 1). This significant performance gain is noticed for almost all tasks, particularly for more complex semantic and question-answering tasks compared to syntactic and segmentation tasks. Similarly, whisper models exhibit promising results in speech recognition with just 2 hours of speech examples in few-shot finetuning. Open models (BLOOMZ and Whisper) performed poorly w.r.t. to their commercially available counterparts. However, fine-tuning with more instructions may help these open models to achieve closer performance to SOTA and other closed LLMs.

C. MSA *vs* Dialect. The gaps in LLMs' performance between MSA and dialectal datasets (e.g., for machine translation (MT) and speech recognition task) are more pronounced, indicating ineffectiveness of LLMs for under-represented dialects.

D. Hallucination and Data Contamination. GPT models, specially GPT-3.5, suffer from the hallucination problem. We noticed the model insert extra information (e.g., for MT task with Bible dataset) from its parametric memory.

Benchmarking LLMs raises concerns about their exposure to existing datasets. In our study, we utilized datasets that were released after the cut-off date of GPT's training (September 2021). Moreover, we applied a prompt-based method with tailored instructions (Golchin and Surdeanu, 2023) on nine datasets using GPT-4, to determine if these datasets are present in the GPT-4 model. Our experiments yielded no examples from these datasets.

To the best of our knowledge, our *LAraBench* study is the *first* comprehensive effort that includes commercial (close) and open source LLMs and evaluates zero- and few-shot settings for a wide range of Arabic NLP and Speech tasks. It is the *first* to include the evaluation of Whisper and USM models for Arabic ASR and the *first* to report benchmarks for a standard Arabic TTS generative model. All resources and findings of the *LAraBench* study will be made publicly available to the community to scale up the effort.[1]

2 Tasks and Datasets

The LAraBench study was designed with an ambitious goal of empowering the research community and practitioners with the most comprehensive evaluation of LLMs for Arabic NLP and Speech tasks to date. It includes 61 publicly available datasets to support 9 task groups[2] discussed below. We briefly describe each task and refer to Appendix A for a comprehensive description of tasks and datasets.

Word Segmentation, Syntax and Information Extraction. We explore six sequence tagging tasks: i) word segmentation, ii) POS-tagging, iii) lemmatization, iv) diacritization, v) parsing, and vi) named-entity recognition (NER), using publicly available datasets. We also include a dialect identification task (e.g., Egyptian dialect) since vocabulary, pronunciation, grammar, and idiomatic expressions vary across dialects. For our benchmarking we used QADI (Abdelali et al., 2021) and ADI (in-house) datasets.

Machine Translation (MT). Machine translation of Arabic is challenging due to morphological complexity and dialectal variations. We experiment

[1] `http://anonymous.com/`

[2] Our task categorization is derived from the taxonomy outlined in the list of tracks established by ACL 2022.

with AraBench (Sajjad et al., 2020), an extensive suite of data from diverse genres and dialects.

Sentiment, Stylistic and Emotion Analysis. These tasks involve understanding and analyzing aspects of human expression and communication. We benchmark Sentiment Analysis, Emotion Recognition, Stance Detection, and Sarcasm Detection with datasets from Elmadany et al. (2018), Mohammad et al. (2018), Chouigui et al. (2017), and Abu Farha et al. (2021), respectively.

News Categorization. This task involves classification of news articles into pre-defined categories or topics (Sebastiani, 2002). We support benchmark evaluations using SANAD news article corpus (Einea et al., 2019) and ASND social media dataset (Chowdhury et al., 2020b).

Demographic Attributes. Demographic information, including gender, age, and country of origin, hold significant value across various applications such as population analysis. We include datasets that enable experimentation with tasks of identifying country, gender (Mubarak et al., 2022) and location (Mubarak and Hassan, 2021).

Ethics and NLP: Factuality, Disinformation and Harmful Content Detection. These tasks have emerged as critical areas within the field of NLP. We support benchmarking of several *detection* tasks, such as: i) Propaganda (Alam et al., 2022b), ii) Factuality using the datasets Baly et al. (2018a); Alam et al. (2021b); Khouja (2020), iii) Harmful content (Nakov et al., 2022b), iv) Offensive language (Zampieri et al., 2020), and v) Hate speech (Mubarak et al., 2021a).

Semantics. This task group includes Semantic Textual Similarity (STS) and Natural Language Inference (NLI). We benchmark STS using two datasets: SemEval-2017 STS task (Cer et al., 2017a) and similarity in Arabic question pairs, as explored by Seelawi et al. (2019). For the XNLI task, we used the translated version of Arabic from XNLI corpus (Conneau et al., 2018).

Question Answering (QA). For the QA task, we employed ARCD (Mozannar et al., 2019), MLQA (Lewis et al., 2019), TyDiQA (Clark et al., 2020), and XQuAD (Artetxe et al., 2020) datasets.

Speech Processing. We evaluate the large speech models on two tasks: speech recognition (ASR) and text-to-speech (TTS) synthesis. For ASR, we include datasets varying domain and dialects, e.g. MGB2 (Ali et al., 2016), QASR.CS (Mubarak et al., 2021b) and ESCWA.CS (Ali et al., 2021a). For TTS, we evaluated with in-house 30 test sentences, covering diverse topics (e.g., education, health).

3 Methodology

For benchmarking of Arabic NLP and Speech processing tasks, we use zero- and few-shot learning involving GPT-3.5-Turbo, GPT-4 and BLOOMZ for NLP, and Whisper (small, medium, and large), USM and Amazon Polly for Speech. We also compared LLM's performance with the respective SOTA models.

The use and evaluation of LLMs involve prompting and post-processing of output to extract the expected content. Therefore, for each task, we explored a number of prompts, guided by the same instruction and format as recommended in the Azure OpenAI Studio Chat playground, and PromptSource (Bach et al., 2022). After obtaining a reasonable prompt, we used it to complete the evaluation using the task and modality-specific API services, e.g., OpenAI API from Azure for NLP tasks and Google's USM API for Speech tasks. For BLOOMZ, we set up on-premises hosting and use.

We based our model selection on factors like performance, language support, and accessibility. For NLP tasks, we chose OpenAI models because they consistently outperformed others for English tasks. Initially, we used GPT-3.5 and later transitioned to GPT-4 when it became available. Limited budget and lack of Arabic language support led us to avoid other closed models. Among open models, we selected BLOOMZ because it's a large multilingual model, including 4% Arabic content.[3] For ASR, we chose Whisper and USM due to their excellent performance in recent studies.

3.1 Models and Prompts for NLP Tasks

Zero-shot Setup. For tasks with GPT-3.5-Turbo, GPT-4 and BLOOMZ, we use zero-shot prompting giving natural language instructions describing the task and specify the expected output. We allow the LLM itself to build context that narrows the inference space and produces more accurate output.

Few-shot Setup. In order to explore the maximum potential of specific LLMs, e.g., GPT-4

[3] New models such as JAIS (Sengupta et al., 2023) and AceGPT (Huang et al., 2023) have been released as we speak and we leave their benchmarking for the future.

model, we used available training data to select few-shot examples and provide context for the task. For a few tasks and datasets (e.g., location, name to country), training sets are either private or not available and therefore they could not be included in our few-shot experiments. We used maximal marginal relevance-based (MMR) selection to construct example sets that are deemed relevant and diverse (Carbonell and Goldstein, 1998), following the proven method by Ye et al. (2022). The MMR method computes the similarity between a test example and the example pool (e.g., training dataset) and selects m examples (shots). We apply MMR on top of embeddings of multilingual sentence-transformers (Reimers and Gurevych, 2019). In our few-shot investigation, we performed experiments on all tasks and datasets using only 3-shots to primarily reduce computational and API expenses. Additionally, we expanded our analysis to include 3, 5, and 10 instances across seven distinct datasets drawn from various task categories. More details are provided in Section F.2 of the Appendix.

Prompts Design. Prompt design is a complex and iterative process that present challenges due to the unknown representation of information within LLMs and a need for different types of outputs across tasks, e.g., token classification vs. sentence classification. The instructions expressed in our prompts were in English, including the content examples in Arabic. In Appendix C, we provide examples of prompts for different tasks. We also examined Arabic instructions in our study, to understand the effect of native language prompts. For this set of experiments we selected seven datasets from seven different task groups. More details can be found in Section F.3 (Appendix).

Post Processing. Outputs of LLMs are post-processed to enable automatic comparison with gold standard labels. Depending on the task, this may include mapping prefixes, or filtering tokens. For example, for POS tagging, the tags *'preposition'*, *'P'*, *'PRP'*, *'حرف جر'*, are mapped onto *PREP*. For NER, the model switches the tag of the prediction i.e., B-PER predicts as PER-B, and therefore requires remapping of the NER tags.

3.2 Models and Prompts for Speech Tasks

We use zero- and few-shot settings to benchmark large speech models. For ASR, we use three Whisper models (OpenAI) – small, medium, and large, and the USM model (Google). For the details of the models, see Table B.2 in Appendix. We compare these large models to SOTA: supervised KANARI[4] conformer-based (Chowdhury et al., 2021) offline and RNN-T based streaming ASR.[5] For the TTS task, we compare two public systems: Amazon Polly TTS engine[6] and KANARI TTS system.[7]

Zero-shot Setup. For zero-shot setup, we use the initial (or pre-trained) weights of Whisper and API of USM models with a goal to benchmark the performances of these LLMs in different domains, for different Arabic dialects, and for code-switching with no domain knowledge. As a prompt to the model, we passed only a language flag.

Few-shot Setup. Under this setup, we fine-tune Whisper (small and large) with 2 hours of domain-specific speech data and compare it with the SOTA models trained from scratch with 3K hours of speech.

ASR Post Processing. ASR is evaluated based on word error rate (WER) that aligns the model's output with reference transcription and penalizes the output based on insertion, deletion, and substitution errors. The measure is unable to disambiguate code-switching and minor formatting differences introduced by multilingual scripts or non-standardized orthography. Hence, post-processing is a crucial component. We normalized 'alif', 'ya' and ta-marbuta', and adapted a minimalist Global Mapping File (GLM) (Chowdhury et al., 2021) to transliterate common words and handle rendering mismatch. Thus keeping room for further improvement with more enhanced post-processing.

3.3 Random Baseline

We also calculated a random baseline for the NLP tasks (further details can be found in Appendix, Section F.1). The aim is to determine if the LLMs predictions are not merely the result of chance. It also serves as a lower limit to be expected for each task.

3.4 Evaluation Metrics

To measure the performance of each task, we followed current state-of-art references and used the metric reported in the respective work. This includes: Accuracy (Acc), F1 (macro, micro, and

[4]https://fenek.ai/
[5]https://arabicasr.kanari.ai/
[6]https://aws.amazon.com/polly/
[7]https://arabictts.kanari.ai/

weighted), word error rate (WER), Jaccard Similarity (JS), Pearson Correlation (PC), and mean opinion score (MOS) for naturalness, intelligibility and diacritization. We report average MOS (10-point Likert scale) from 3 native-annotators.

4 Results and Discussion

In Tables 1, 2, 3 and 4, we report the results of different NLP and Speech related tasks. In the below sections, we summarize the results and challenges specific to the task groups.

4.1 NLP Tasks

In Table 1, we report the random baseline, GPT-3.5, GPT-4 (zero-shot and few-shot), and BLOOMZ and compare them to SOTA.[8] In almost all tasks, models outperform random baseline, indicating that the predictions of the models are not by chance.

Word Segmentation, Syntax and Information Extraction. As Table 1 shows, for almost all tasks in this group, the performance is significantly below SOTA performance. For example, the difference between SOTA and GPT-4 (zero-shot) ranges from 19.4% (NER) to 73.8% (segmentation).

Machine Translation. Table 2 reports MT results by averaging them dialect-wise for different datasets. Appendix F.6 reports detailed results. The results indicate the short-coming of LLMs when explored with standard and dialectal Arabic.

Sentiment, Stylistic and Emotion Analysis. In the second group of Table 1, we report results for sentiment, emotion, stance and sarcasm detection mainly over tweets. We observe that performance gap significantly reduced between GPT-4 (best of zero- and few-shot) vs. SOTA compared to GPT-3.5 vs. SOTA, 4.77% vs 15.14%, respectively. For sarcasm detection task with ArSarcasm dataset, GPT-4 even outperformed SOTA by 4.41%.

News Categorization. Table 1 shows that performance gap reduced significantly ranging from 15.63% to 6.44% for GPT-3.5 to GPT-4, respectively. Low performance on tweet dataset (ASND) might be due to the higher number of class labels.

Demographic/Protected Attributes. Among the three tasks in this group, two of them ("name info" and "location" identification) demonstrate a significant performance improvement (by 3.62%) over the SOTA results, using GPT-4 model.

[8]Note that some results are missing either due to the unavailability of training data (marked with NA) or the incapability of the Bloomz model (marked with ‡).

Ethics and NLP: Factuality, Disinformation and Harmful Content Detection. Across eleven tasks, the performance gap significantly reduced with GPT-4 model, however in some tasks, model's performance is significantly lower than the SOTA. For example, for factuality with COVID-19 disinfo. dataset, GPT-4 model's performance is 33% lower than the SOTA, even though performances of GPT-4 significantly improved compared to GPT-3.5. This task is generally challenging requiring deep contextual analysis and reasoning abilities, and domain knowledge in many of the cases. With a few demonstrations (only 3-shots) may not be enough to determine the factuality of the content.

Semantics: The results for various semantic tasks reported in Table 1 indicate that the performance on three out of the four tasks surpasses the SOTA, with an overall improvement of 7%.

Question answering (QA): Results on four QA datasets (Table 1) show that for three of them, GPT-4 achieved higher performance than SOTA with an overall improvement of 4.42%.

4.2 Speech Recognition and Synthesis

In Table 3, we reported the performance of ASR using different datasets and models. We observed that USM outperforms Whisper in all datasets in both zero and few-shot setting. The USM model performs comparably to standard task- and domain-specific ASR systems and is better equipped to handle cross-language and dialectal code-switching data from unseen domains compared to the SOTAs and Whispers few-shot finetuned model.

Both the subjective and objective evaluations for the TTS are reported Table 4. The results show that KANARI models outperformed Amazon Polly significantly in objective evaluation (WER). Subjective scores show KANARI is better in naturalness and diacritization. With almost similar performance in intelligibility.

5 Findings

NLP Model Performances. Our comprehensive study highlights the disparities in performance of LLMs – GPT-3.5 and GPT-4, as compared to SOTA models, in zero and few-shot settings. GPT-3.5

Task Name	Dataset	Metric	Random Baseline	BLOOMZ	Zero-shot GPT-3.5	Zero-shot GPT-4	Few-Shot GPT-4 (3-shot)	SOTA
Word Segmentation, Syntax and Information Extraction								
Segmentation	WikiNews	Acc	0.272	‡	0.195	0.252	0.927	**0.989** (Darwish and Mubarak, 2016)
Segmentation	Samih et al. (2017)	Acc_{AVG}	0.309	‡	0.283	0.372	0.850	**0.931** Samih et al. (2017)
Lemmatization	WikiNews	Acc	0.348	‡	0.471	0.397	NA	**0.973** (Mubarak, 2018)
Diacritization	WikiNews	WER	0.963	‡	0.308	0.420	0.237	**0.045** (Mubarak et al., 2019)
Diacritization	Darwish et al. (2018)	WER	0.999	‡	0.928	0.899	0.994	**0.031** (Darwish et al., 2018)
POS	WikiNews	Acc	0.030	‡	0.231	0.479	0.367	**0.953** (Darwish et al., 2017c)
POS	Samih et al. (2017)	Acc	0.036	‡	0.073	0.511	0.323	**0.892** Samih et al. (2017)
POS	‡GLUE (Arabic)	Acc	0.032	‡	0.159	0.402	0.524	**0.686** (Liang et al., 2020a)
Parsing	Conll2006	UAS	0.001	‡	0.239	0.504	0.551	**0.796** (Lei et al., 2014)
NER	ANERcorp	$\mathrm{F1}_{Macro}$	0.008	‡	0.210	0.355	0.420	**0.886** (Gridach, 2018)
NER	Aqmar	$\mathrm{F1}_{Macro}$	0.007	‡	0.230	0.365	0.390	**0.690** (Schneider et al., 2012)
NER	QASR	$\mathrm{F1}_{Macro}$	0.009	‡	0.208	0.504	NA	**0.698** (Mubarak et al., 2021b)
Dialect	QADI	$\mathrm{F1}_{Macro}$	0.052	0.067	0.149	0.243	NA	**0.600** (Abdelali et al., 2021)
Dialect	ADI	$\mathrm{F1}_{Macro}$	0.092	0.098	0.169	0.229	0.260	**0.26/0.57** (lexical/acoustic) (In-house)
Sentiment, Stylistic and Emotion Analysis								
Sentiment	ArSAS	$\mathrm{F1}_{Macro}$	0.222	0.251	0.550	0.569	0.598	**0.758** (Hassan et al., 2021)
Emotion	SemEval18-Task1	JS	0.167	0.142	0.395	0.373	0.489	**0.541** (Hassan et al., 2022)
Stance	Unified-FC	$\mathrm{F1}_{Macro}$	0.193	0.235	0.232	0.495	0.358	**0.558** (Baly et al., 2018b)
Stance	ANS	$\mathrm{F1}_{Macro}$	0.281	0.223	0.620	0.762	0.721	**0.767** (Khouja, 2020)
Sarcasm	ArSarcasm	$\mathrm{F1}_{(POS)}$	0.240	0.286	0.465	0.400	**0.504**	0.460 (Farha and Magdy, 2020)
Sarcasm	ArSarcasm-2	$\mathrm{F1}_{(POS)}$	0.333	0.436	0.537	0.573	0.537	**0.623** (Alharbi and Lee, 2021)
News Categorization								
News Cat.	ASND	$\mathrm{F1}_{Macro}$	0.048	0.371	0.512	0.667	0.594	**0.770** (Chowdhury et al., 2020b)
News Cat.	SANAD/Akhbarona	Acc	0.142	0.582	0.730	0.877	0.892	**0.940** (Elnagar et al., 2020)
News Cat.	SANAD/AlArabiya	Acc	0.144	0.716	0.922	0.921	0.925	**0.974** (Elnagar et al., 2020)
News Cat.	SANAD/AlKhaleej	Acc	0.142	0.738	0.864	0.911	0.899	**0.969** (Elnagar et al., 2020)
Demographic Attributes								
Name Info	ASAD	$\mathrm{F1}_{Weighted}$	0.014	‡	0.570	**0.629**	NA	0.530 (Under review)
Location	UL2C	$\mathrm{F1}_{Macro}$	0.027	0.118	0.339	0.735	NA	**0.881** (Mubarak and Hassan, 2021)
Gender	Arap-Tweet	$\mathrm{F1}_{Macro}$	0.521	0.532	0.883	0.868	**0.980**	0.821 (Mubarak et al., 2022)
Ethics and NLP: Factuality, Disinformation and Harmful Content Detection								
Offensive lang.	OffensEval2020	$\mathrm{F1}_{Macro}$	0.454	0.533	0.460	0.623	0.874	**0.905** (Mubarak et al., 2020b)
Hate Speech	OSACT2020	$\mathrm{F1}_{Macro}$	0.376	0.503	0.430	0.669	0.644	**0.823** (Mubarak et al., 2020c)
Adult Content	ASAD	$\mathrm{F1}_{Macro}$	0.421	0.513	0.460	0.727	0.832	**0.889** (Mubarak et al., 2021a)
Spam	ASAD	$\mathrm{F1}_{Macro}$	0.405	0.152	0.440	0.745	NA	**0.989** (Hassan et al., 2021)
Subjectivity	In-house	$\mathrm{F1}_{Macro}$	0.496	0.428	0.670	0.677	**0.745**	0.730 (In-house)
Propaganda	WANLP22	$\mathrm{F1}_{Micro}$	0.139	0.108	0.353	0.472	0.537	**0.649** (Samir et al., 2022)
Check-worthy	CT–CWT–22	$\mathrm{F1}_{(POS)}$	0.398	0.431	0.526	0.560	0.554	**0.628** (Du et al., 2022)
Factuality	COVID-19 Disinfo.	$\mathrm{F1}_{Weighted}$	0.582	**0.749**	0.393	0.485	0.491	**0.831** (Alam et al., 2021b)
Factuality	Unified-FC	$\mathrm{F1}_{Macro}$	0.464	0.460	0.306	0.581	**0.621**	In-house
Factuality	ANS	$\mathrm{F1}_{Macro}$	0.505	0.550	0.252	0.539	**0.704**	0.643 (Khouja, 2020)
Claim	CT–CWT–22	Acc	0.498	0.532	**0.703**	0.587	0.686	0.570 (Eyuboglu et al., 2022)
Harmful content	CT–CWT–22	$\mathrm{F1}_{(POS)}$	0.269	0.144	0.471	0.533	0.494	**0.557** (Bilel et al., 2022)
Attention-worthy	CT–CWT–22	$\mathrm{F1}_{Weighted}$	0.125	0.148	0.258	0.257	**0.412**	0.206 (Nakov et al., 2022a)
Semantics								
STS	STS2017-Track 1	PC	0.005	0.537	0.799	**0.813**	0.809	0.754 (Cer et al., 2017b)
STS	STS2017-Track 2	PC	-0.136	0.512	0.828	0.848	**0.857**	0.749 (Cer et al., 2017b)
STS QS (Q2Q)	Mawdoo3 Q2Q	$\mathrm{F1}_{Micro}$	0.491	0.910	0.816	0.895	0.935	**0.959** (Seelawi et al., 2019)
XNLI (Arabic)	XNLI	Acc	0.332	0.500	0.489	0.753	**0.774**	0.713 (Artetxe et al., 2020)
Question answering (QA)								
QA	ARCD	$\mathrm{F1}_{(EM)}$	0.085	0.368	0.502	**0.705**	0.704	0.613 (Mozannar et al., 2019)
QA	MLQA	$\mathrm{F1}_{(EM)}$	0.066	0.377	0.376	0.620	**0.653**	0.548 (Lewis et al., 2019)
QA	TyDi QA	$\mathrm{F1}_{(EM)}$	0.111	0.456	0.480	0.744	0.739	**0.820** (Clark et al., 2020)
QA	XQuAD	$\mathrm{F1}_{(EM)}$	0.047	0.367	0.442	**0.729**	0.722	0.665 (Artetxe et al., 2020)

Table 1: Results on NLP tasks. QS: Question similarity, PC: Pearson Correlation, JS: Jaccard Similarity, EM: Exact match, POS: positive class. Best result per row is **boldfaced**. **NA**: experiments could not be performed due to a lack of training data. BLOOMZ does not understand some tasks at all as marked with ‡ symbol.

exhibits a significant performance gap when compared to SOTA. However, GPT-4 manages to narrow this gap to some extent and even outperforms the SOTA models in high-level abstract tasks such as STS, QA, claim detection, news categorization, demographic attributes, and XNLI. Moreover, GPT-4 outperforms GPT-3.5 across all tasks. However, it remains a challenge for GPT-4 to surpass SOTA performance consistently in sequence tagging (especially syntactic and segmentation) tasks. The performance of BLOOMZ is significantly lower than SOTA and GPT models, and in some cases lower than random baseline. The performances of both open and close models are heavily dependent

Dataset	Dialect	#Sent.	BloomZ	Zero-shot GPT-3.5	Zero-shot GPT-4	SOTA
APT	LEV	1000	11.38	18.55	17.77	**21.90**
APT	Nile	1000	12.95	21.58	18.99	**22.60**
MADAR	Gulf	16000	32.34	34.60	**36.18**	32.46
MADAR	LEV	12000	31.36	33.42	**35.24**	32.45
MADAR	MGR	14000	23.59	23.91	**27.83**	23.14
MADAR	MSA	2000	42.33	37.55	37.67	**43.40**
MADAR	Nile	8000	34.87	36.97	**37.93**	35.15
MDC	LEV	3000	10.00	17.38	16.05	**17.63**
MDC	MGR	1000	8.28	14.46	**14.20**	13.90
MDC	MSA	1000	15.75	21.05	19.34	**20.40**
Media	Gulf	467	14.22	22.68	**22.76**	19.60
Media	LEV	250	7.54	17.65	16.65	**16.80**
Media	MGR	526	4.87	11.58	**10.20**	9.60
Media	MSA	1258	20.66	35.34	**33.57**	32.65
Bible	MGR	1200	17.09	16.72	15.29	**29.00**
Bible	MSA	1200	22.91	22.08	17.53	**31.20**

Table 2: BLEU score on MT using zero-shot prompts. #Sent: number of test set sentences. SOTA results are reported in (Sajjad et al., 2020).

on the *effective prompt* and implementing appropriate *post-processing techniques*. Overall, these findings indicate the potential of GPT-4 as a *multi-task model* without heavily relying on task-specific resources, particularly in zero/few-shot settings.

The *few-shot results* across *seven different datasets* show an average improvement of 0.656 (0-shot) to 0.721 (10-shot) indicating the promise of few-shot learning, as depicted in Figure 2 (in Appendix), with individual results are reported in Table 9 (in Appendix).

The use of *native language prompts* with GPT-4 in a zero-shot context highlighted the role played by the prompt language, as we observed increased performance in three out of seven datasets compared to their counterparts with English prompts while two underperformed, and one showed equivalent performance (see Table 10 in Appendix).

When evaluating these LLMs in *multi-dialectal* settings, the performance gap between MSA and dialectal test sets becomes more evident. For example, in both the GPT-models, we noticed a large discrepancy in the POS accuracy of 0.810 versus 0.379 on MSA and dialects respectively. Similarly, for the dialect identification we notice a significant difference between the SOTA acoustic and lexical model with respect to LLMs results.

From the average *performance gap between semantic and syntactic tasks*, as reported in Table 11 (in Appendix), we noticed the discrepancy in semantic tasks is much lower than in syntactic tasks, across the three LLMs. This suggests that these models might be better equipped at encoding and expressing semantic information than in pinpointing specific syntactic phenomena in their inputs.

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

Table 3: Reported WER (↓) on ASR in zero and few-shot setup and domain-specific ASR setup. W.S,M,Lv2 stands for OpenAI Whisper small, medium and Largev2 model. O: represent offline; S: streaming ASR; * represent the model's input is 8kHz sampling rate and Offline model was re-trained to accommodate telephony data. $^{+}$ represent model fine-tuned with 2hrs of MGB2-data.

Moreover, these performance gaps can also be linked to *undesirable hallucination*. In particular, during the MT for the Bible, results reveal an interesting phenomenon. It appears that the GPT models, particularly GPT-3.5-turbo, tend to hallucinate and insert additional content in their responses.

Is the data contaminated? We have used some datasets for evaluation that are released after the cut-off date of ChatGPT training, which include subjectivity, propaganda, check worthiness, factuality (CT-CWT-22), harmful content, and attention worthiness. Moreover, we experiment with nine datasets using the tailored instructions approach proposed by Golchin and Surdeanu (2023) revealing that GPT-4 could not produce any example from these datasets. Thus, we can confirm that the models have not been contaminated with such datasets. More details in Appendix F.5.

Speech Model Performances: We observed the performance of these models is heavily depen-

	Subjective (MOS) ↑			Objective ↓	
Model	Diac.	Natur.	Intel.	WER	CER
Amazon	8.2	8.3	9.8	5.2	**1.0**
KANARI	**9.5**	8.6	9.8	**3.7**	1.2

Table 4: Evaluation for Arabic TTS. Diac.: Diacritization, Natur.: Naturalness, Intel.: Intelligibility.

dent on the architecture parameters. USM model performs comparably with SOTA for MSA. Both Whisper (and its variants) and USM show a performance gap when dealing with dialects specially Moroccan dialect. Fine-tuning the open model (Whisper Largev2) with only 2 hours of speech data bridges the performance gap significantly, indicating the potential to be a robust and strong foundation model. Our observation also suggests that USM model is better equipped to handle code-switching phenomena in spoken utterance than the supervised large transformer models.

6 Related Work

Models for NLP: Since the inception of the transformer architecture (Vaswani et al., 2017), there have been efforts to develop larger models with its variants such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020), GPT models (Radford et al., 2018, 2019; Ouyang et al., 2022) among others.

Such advancements have led to the development LLMs with parameter sizes exceeding 100 billion, which are pre-trained on massive datasets. Examples of LLMs include Megatron (Shoeybi et al., 2019), GPT-3 (Brown et al., 2020), GPT-Jurassic (Lieber et al., 2021), OPT-175B (Zhang et al., 2022), and Bloom (Scao et al., 2022). This unprecedented scale enabled new capabilities that address the zero-shot and multilingual tasks learning. ChatGPT (GPT-3.5) and its subsequent model GPT-4 is the latest development in NLP that have addressed many limitations of prior LLMs and enabled us to perform diverse tasks (OpenAI, 2023). The ability of LLMs to solve various tasks can be attributed to the meticulous design of prompts, which enable the generation of desired responses (Wei et al., 2022; Shin et al., 2020).

Models for Speech Processing: When handling complex audio/speech data, LLMs face significant challenges. However, with the advent of self-supervised learning, models like Wav2vec, WavLM, and Whisper have been leading in addressing these challenges (Baevski et al., 2019, 2020; Chen et al., 2022; Radford et al., 2022). More recent developments like the Universal Speech Model (USM) and VALL-E have demonstrated superior capabilities in ASR and zero-shot TTS tasks, respectively (Zhang et al., 2023; Wang et al., 2023).

LLMs Benchmarking: Since the release of ChatGPT, there have been efforts to evaluate the performance of LLMs on standard NLP tasks (Bubeck et al., 2023; Bang et al., 2023; Ahuja et al., 2023; Hendy et al., 2023). Liang et al. (2022) conducted a comprehensive assessment of LLMs for English. It encompassed various metrics such as accuracy, calibration, toxicity, and efficiency, along with 42 scenarios involving 30 prominent language models.

Benchmarks on Arabic: The complexity and linguistic diversity of Arabic have led to a limited number of benchmarks for language tasks, such as ORCA (Elmadany et al., 2022), ALUE (Seelawi et al., 2021), ArBERT (Abdul-Mageed et al., 2021), and AraBench (Sajjad et al., 2020).

LAraBench: To the best of our knowledge, our study represents the first comprehensive Arabic language benchmarking effort exploring GPT-3.5 (zero-shot), GPT-4 (zero- and few-shot), BLOOMZ (zero-shot), and Speech models like Whisper and USM. Our evaluation spans a broad array of LLMs, tasks, and datasets, distinguishing it from prior benchmarks in terms of task and dataset diversity, test setup, modalities (text, speech), and state-of-the-art comparisons. Table 8 (Appendix E), provides a detailed comparison.

7 Conclusion and Future Studies

This study is the *first* large-scale benchmark that brings together both Arabic Speech and NLP tasks under the same study. We report the performance of LLMs for a variety of tasks covering different domains and dialects. Our study also considers tasks with a wide range of complexity ranging from token to text classification, different application settings, NER to sentiment, factuality and disinformation, ASR, and TTS among others. We evaluate 33 tasks and 61 datasets with 98 test setups, which are very prominent for Arabic AI. We compare and report the performance of each task and dataset with SOTA, which will enable the community and practitioners of large language models to decide on their uses of these models. Future work aims to investigate open models and explore ways to reduce the performance gap with SOTA; enhance prompts for better performance; and expand datasets and tasks studied.

Limitations

The main focus of this study was to benchmark large language models for Arabic NLP and Speech tasks. Given that this is a work in progress, there are currently some limitations. In this edition, we evaluated several large models: ChatGPT, USM, and Whisper models and compared them to SOTA. We plan to extend our study by adding other models such Bard, Claude, MMS, and other open multilingual models that have Arabic. In this work, we benchmarked 61 datasets with 98 test setups for 33 tasks. However, we did not benchmark all available data sets. For example, the study reported in (Elmadany et al., 2022) benchmarked 19 sentiment datasets, whereas we only covered one. It is also possible that we missed many other Arabic NLP and Speech tasks, which we will attempt to cover in the future. Our current results are highly dependent on prompt design. Additional efforts on prompt engineering could potentially improve the results.

In addition, performance may vary depending on the version of the models we used.[9] For GPTs, we utilized gpt-3.5-turbo-0301 and gpt-4-0314 versions for our NLP tasks. To ensure transparency and reproducibility, we are committed to sharing all our experimental resources, including prompts and parameter details. This will facilitate the easy replication of our results using the provided pipeline and the fixed model versions. The same principle extends to our speech models. We have taken steps to maintain versioning not only for the models themselves but also for the prompts used. This ensures that our work remains reproducible for future researchers in the field.

Potential Risk We do not oversee any potential risk that can result from our study.

Ethics Statement

We used publicly available and in-house developed datasets in our study. Any biases are unintended.

References

Ahmed Abdelali, Mohammed Attia, Younes Samih, Kareem Darwish, and Hamdy Mubarak. 2019. Diacritization of maghrebi Arabic sub-dialects. *arXiv preprint arXiv:1810.06619*.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic dialect identification in the wild. *arXiv preprint arXiv:2005.06557*.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021. Arbert & marbert: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the wanlp 2021 shared task on sarcasm and sentiment detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. MEGA: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022a. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, COLING '22, pages 6625–6643, Gyeongju, Republic of Korea.

Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. 2021a. Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '21, pages 913–922.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022b. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. pages 108–118.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021b. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the*

[9] https://platform.openai.com/docs/models/overview

Association for Computational Linguistics: EMNLP 2021, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Abdullah I Alharbi and Mark Lee. 2021. Multi-task learning using a combination of contextualised and static word embeddings for arabic sarcasm detection and sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 318–322.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Ahmed Ali, Shammur Chowdhury, Amir Hussein, and Yasser Hifny. 2021a. Arabic code-switching speech recognition using monolingual data. *arXiv preprint arXiv:2107.01573*.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019. The MGB-5 challenge: Recognition and dialect identification of dialectal Arabic speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic MGB-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021b. Arafacts: the first large Arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236.

Francesco Antici, Luca Bolognini, Matteo Antonio Inajetovic, Bogdan Ivasiuk, Andrea Galassi, and Federico Ruggeri. 2021. Subjectivita: An italian corpus for subjectivity detection in newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 40–52, Cham. Springer International Publishing.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Févry, et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018a. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Giovanni Da San Martino, Tamer Elsayed, Andrea Galassi, Fatima Haouari, Federico Ruggeri, Julia Maria Struß, Rabindra Nath Nandi, et al. 2023. The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, pages 506–517. Springer.

Yassine Benajiba and Paolo Rosso. 2007. ANERsys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with pos-tag information. In *IICAI*, pages 1814–1823.

Yassine Benajiba, Paolo Rosso, and José Miguel BenedíRuiz. 2007. ANERsys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing*, pages 143–153, Berlin, Heidelberg. Springer Berlin Heidelberg.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in english. *Discourse processes*, 11(1):1–34.

Taboubi Bilel, Ben Nessir Mohamed Aziz, and Hatem Haddad. 2022. iCompass at CheckThat! 2022: ARBERT and AraBERT for Arabic checkworthy tweet identification. In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, CLEF ’2022, Bologna, Italy.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *LREC*, pages 1240–1245.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The MADAR Arabic Dialect Corpus and Lexicon. In *LREC*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. 2018. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American journal of public health*, 108(10):1378–1384.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. Technical report, Microsoft Research.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017a. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko E. Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017b. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, SemEval-2017, pages 1–14, Vancouver, Canada.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2017. ANT corpus: An Arabic news text collection for textual classification. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 135–142. IEEE.

Shammur A Chowdhury, Younes Samih, Mohamed Eldesouki, and Ahmed Ali. 2020a. Effects of dialectal code-switching on speech modules: A study using egyptian Arabic broadcast speech. *Proc. Interspeech*.

Shammur Absar Chowdhury, Ahmed Abdelali, Kareem Darwish, Jung Soon-Gyo, Joni Salminen, and Bernard J Jansen. 2020b. Improving Arabic text categorization using transformer training diversification. In *Proceedings of the fifth arabic natural language processing workshop*, pages 226–236.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching Arabic asr. *Proc. Interspeech*.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. 2020c. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6203–6212.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

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*, ACL '20, pages 8440–8451, Online. Association for Computational Linguistics.

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*, EMNLP '18, pages 2475–2485.

Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, Younes Samih, and Mohammed Attia. 2018. Diacritization of moroccan and tunisian Arabic dialects: A crf approach. *OSACT*, 3:62.

Kareem Darwish, Dimitar Alexandrov, Preslav Nakov, and Yelena Mejova. 2017a. Seminar users in the Arabic twitter sphere. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*, pages 91–108. Springer.

Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A new fast and accurate Arabic word segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1070–1074.

Kareem Darwish, Hamdy Mubarak, and Ahmed Abdelali. 2017b. Arabic diacritization: Stats, rules, and hacks. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 9–17, Valencia, Spain. Association for Computational Linguistics.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, and Mohamed Eldesouki. 2017c. Arabic POS tagging: Don't abandon feature engineering just yet. In *Proceedings of the third arabic natural language processing workshop*, pages 130–137.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. 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.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Mingzhe Du, Sujatha Das Gollapalli, and See-Kiong Ng. 2022. Nus-ids at checkthat! 2022: identifying check-worthiness of tweets using checkthat5. *Working Notes of CLEF*.

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. 2019. SANAD: Single-label Arabic news articles dataset for automatic text categorization. *Data in brief*, 25:104076.

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2022. Orca: A challenging benchmark for Arabic language understanding. *arXiv preprint arXiv:2212.10758*.

AbdelRahim A. Elmadany, Hamdy Mubarak, and Walid Magdy. 2018. ArSAS: An Arabic speech-act and sentiment corpus of tweets. *OSACT*, 3:20.

Ashraf Elnagar, Ridhwan Al-Debsi, and Omar Einea. 2020. Arabic text classification using deep learning models. *Information Processing & Management*, 57(1):102121.

A. Etman and A. A. Louis Beex. 2015. Language and dialect identification: A survey. In *2015 SAI Intelligent Systems Conference (IntelliSys)*, pages 220–231.

Ahmet Bahadir Eyuboglu, Mustafa Bora Arslan, Ekrem Sonmezer, and Mucahid Kutlu. 2022. TOBB ETU at CheckThat! 2022: detecting attention-worthy and harmful tweets and check-worthy claims. In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, CLEF '2022, Bologna, Italy.

Ibrahim Abu Farha and Walid Magdy. 2020. From arabic sentiment analysis to sarcasm detection: The arsarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39.

Hongyu Gao, Yan Chen, Kathy Lee, Diana Palsetia, and Alok N. Choudhary. 2012. Towards online spam filtering in social networks. In *Network and Distributed System Security Symposium*, NDSS '12, pages 1–16.

Razan Ghanem, Hasan Erbay, and Khaled Bakour. 2023. Contents-based spam detection on social networks using roberta embedding and stacked blstm. *SN Computer Science*, 4(4):380.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Mourad Gridach. 2018. Deep learning approach for arabic named entity recognition. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part I 17*, pages 439–451. Springer.

Jan Hajic, Otakar Smrz, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. 2004. Prague Arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, volume 1.

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. Asad: Arabic social media analytics and understanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118.

Sabit Hassan, Shaden Shaar, and Kareem Darwish. 2022. Cross-lingual emotion detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6948–6958.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, et al. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the Arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv preprint arXiv:2305.14976*.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, 2(2).

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Gaurav Kumar, Yuan Cao, Ryan Cotterell, Chris Callison-Burch, Daniel Povey, and Sanjeev Khudanpur. 2014. Translations of the callhome Egyptian Arabic corpus for conversational speech translation. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Papers*, pages 244–248, Lake Tahoe, California.

Tao Lei, Yu Xin, Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2014. Low-rank tensors for scoring dependency structures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1381–1391.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020a. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020b. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 1.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Clyde R. Miller. 1939. The Techniques of Propaganda. From "How to Detect and Analyze Propaganda," an address given at Town Hall. The Center for learning.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural Arabic question answering. *arXiv preprint arXiv:1906.05394*.

Hamdy Mubarak. 2018. Build fast and accurate lemmatization for Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hamdy Mubarak, Ahmed Abdelali, Sabit Hassan, and Kareem Darwish. 2020a. Spam detection on Arabic twitter. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 237–251. Springer.

Hamdy Mubarak, Ahmed Abdelali, Hassan Sajjad, Younes Samih, and Kareem Darwish. 2019. Highly effective Arabic diacritization using sequence to sequence modeling. 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 2390–2395.

Hamdy Mubarak, Shammur Absar Chowdhury, and Firoj Alam. 2022. ArabGend: Gender analysis and inference on Arabic Twitter. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 124–135, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020b. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resource Association.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020c. Overview of osact4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection*, pages 48–52.

Hamdy Mubarak and Sabit Hassan. 2021. Ul2c: Mapping user locations to countries on Arabic twitter. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 145–153.

Hamdy Mubarak, Sabit Hassan, and Ahmed Abdelali. 2021a. Adult content detection on Arabic twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 136–144.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021b. QASR: QCRI Aljazeera speech resource–a large scale annotated Arabic speech corpus. *arXiv preprint arXiv:2106.13000*.

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.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak, Alex Nikolov, Yavuz Selim Kartal, and Javier Beltrán. 2022a. Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In *Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum*, CLEF '2022.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, et al. 2022b. Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, pages 495–520. Springer.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, IJCAI '21, pages 4551–4558.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving Language Understanding by Generative Pre-Training". Technical report, Open AI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed Elsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, et al. 2022. Newsclaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Pro-*

ceedings of the 28th International Conference on Computational Linguistics, pages 5094–5107.

Younes Samih, Mohamed Eldesouki, Mohammed Attia, Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, and Laura Kallmeyer. 2017. Learning from relatives: Unified dialectal Arabic segmentation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 432–441.

Ahmed Samir, Abu Bakr Soliman, Mohamed Ibrahim, Laila Hesham, and Samhaa R El-Beltagy. 2022. Ngu_cnlp at wanlp 2022 shared task: Propaganda detection in arabic. *WANLP 2022*, page 545.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A Smith. 2012. Coarse lexical semantic annotation with supersenses: an arabic case study. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 253–258.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Haitham Seelawi, Ahmad Mustafa, Hesham Al-Bataineh, Wael Farhan, and Hussein T Al-Natsheh. 2019. Nsurl-2019 task 8: Semantic question similarity in Arabic. In *Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers*, pages 1–8.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. Alue: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Shaden Shaar, Maram Hasanain, Bayan Hamdan, Zien Sheikh Ali, Fatima Haouari, Alex Nikolov, Mucahid Kutlu, Yavuz Selim Kartal, Firoj Alam, Giovanni Da San Martino, Alberto Barrón-Cedeño, Rubén Míguez, Javier Beltrán, Tamer Elsayed, and Preslav Nakov. 2021. Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates. In *2021 Working Notes of CLEF - Conference and Labs of the Evaluation Forum*.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI '22, pages 5597–5606, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Md Tawkat, Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv e-prints*, pages arXiv–2305.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *arXiv preprint arXiv:2211.13892*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, and R Ziane. 2020. Universal dependencies 2.5. *LINDAT/CLARIAHCZ digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University. url: http://hdl. handle. net/11234/1-3226*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Appendix

A Tasks and Datasets

In this section, we discuss the tasks and the associated datasets by grouping them based on ACL-2022 track.[10] In Tables 5 and 6, we provide a summarized description of the test sets used for evaluating textual and speech processing tasks, respectively.

A.1 Word Segmentation, Syntax and Information Extraction

A.1.1 Segmentation

Segmentation is an important problem for language like Arabic, which is rich with bound morphemes that change the tense of verbs, or represent pronouns and prepositions in nouns. It is a building block for NLP tasks such as search, part-of-speech tagging, parsing, and machine translation. The idea is segmenting Arabic words into prefixes, stems, and suffixes, which can facilitate many other tasks.

Datasets

WikiNews For modern standard Arabic (MSA), we used the WikiNews dataset of (Darwish and Mubarak, 2016) which comprises 70 news articles in politics, economics, health, science and technology, sports, arts, and culture. The dataset has 400 sentences (18,271 words) in total.

Tweets For the dialectal Arabic, we used the dataset in (Samih et al., 2017), which provides 1400 tweets in Egyptian, Gulf, Levantine, and Maghrebi dialects for a total of 25,708 annotated words .

A.1.2 Part-Of-Speech (POS) Tagging

Part-of-speech (POS) is one of the fundamental components in the NLP pipeline. It helps in extracting higher-level information such as named entities, discourse, and syntactic parsing.

Datasets

WikiNews We used for this task the WikiNews dataset tagged for POS (Darwish et al., 2017c) for modern standard Arabic.

Tweets For POS tagging with noisy texts and different dialects we used the same dataset reported in (Samih et al., 2017) (see §A.1.1).

XGLUE We also used the Arabic part of XGLUE benchmark (Liang et al., 2020b) for POS tagging, which uses a subset of Universal Dependencies Treebanks (v2.5) (Zeman et al., 2020).

[10] https://www.2022.aclweb.org/callpapers

A.1.3 Lemmatization

Lemmatization is another component in the NLP pipeline, which reduces words to their base or root form, known as a lemma. It takes into consideration the morphological analysis of the words, which uses the context and POS to convert a word to its simplest form. This task differs from segmentation which only separates a word stem from prefixes and suffixes. In contrast, lemmatization requires returning the lexicon entry for a certain word, which may depend on POS tagging.

Dataset We used WikiNews dataset tagged for lemmas (Mubarak, 2018) (see §A.1.1 for the details of the dataset).

A.1.4 Diacritization

Diacritization involves assigning the diacritics to each letter in an Arabic word within a sentence. Diacritical marks indicate the correct pronunciation and meaning of the written Arabic words. For example, different word diacretizations could transform a noun into a verb or vice versa.

Datasets

WikiNews We use a dataset of modern standard Arabic from (Mubarak et al., 2019) that comprises fully diacritized WikiNews corpus (Darwish et al., 2017b).

Bibles This dataset includes translations of the New Testament into two Maghrebi sub-dialects: Moroccan and Tunisian (Darwish et al., 2018; Abdelali et al., 2019).

A.1.5 Parsing

Dependency parsing is the task of identifying syntactical and grammatical relations among the words in a sentence. These dependencies result in a hierarchical tree representation that captures the structure of the sentence at different levels.

Dataset For this task we used the Arabic part of CoNLL-X 2006 shared tasks on dependency parsing (Buchholz and Marsi, 2006), which has 4,990 scoring tokens and uses the Prague Arabic Dependency Treebank (Hajic et al., 2004).

A.1.6 Named-Entity Recognition (NER)

This task involves identifying and classifying the words in a sentence that are proper names, names of places, entities like organizations or products, amongst other things. This depends on understanding the context and the relations of a word or a

Dataset	Task	Domain	Test Set Size
Word Segmentation, Syntax and Information Extraction			
WikiNews	Segmentation	News articles (MSA)	400 sentences
Samih et al. (2017)	Segmentation	Tweets (Dialects: EGY, LEV, GLF, MGR)	70 X 4 dialects
WikiNews	Lemmatization	News articles (MSA)	400 sentences
WikiNews	Diacritization	News articles (MSA)	400 sentences
Darwish et al. (2018)	Diacritization	Sentences (Dialects: Moroccan, Tunisian)	1,640 X 2 dialects
WikiNews	POS	News articles (MSA)	400 sentences
Samih et al. (2017)	POS	Tweets (Dialects: EGY, LEV, GLF, MGR)	70 X 4 dialects
XGLUE (Arabic)	POS	Web, Wikipedia	680 sentences
Conll2006	Parsing	MSA	146 sentences
ANERcorp	NER	News articles	924 sentences
AQMAR	NER	Wikipidia	1,976 sentences
QASR	NER	Transcripts	7,906 segments
QADI	Dialect	Tweets	3,797
ADI	Dialect	Transcripts	751
Sentiment, Stylistic and Emotion Analysis			
ArSAS	Sentiment	Tweets	4,213
SemEval2018-Task1	Emotion	Tweets (Dialectal)	1,518
Unified-FC	Stance	News articles	3,042 claim-article pairs
ANS	Stance	News articles	379 headline pairs
ArSarcasm	Sarcasm	Tweets	2,110
ArSarcasm-2	Sarcasm	Tweets	3,000
News Categorization			
ASND	News Cat.	Posts∗	1,103
SANAD/Akhbarona	News Cat.	News articles	7,843
SANAD/AlArabiya	News Cat.	News articles	7,125
SANAD/AlKhaleej	News Cat.	News articles	4,550
Demographic Attributes			
ASAD	Name Info	Wikidata	80,130
UL2C	Location	User loc. (Twitter)	28,317
Arap-Tweet	Gender	Usernames (Twitter)	640
Ethics in NLP: Factuality, Disinformation and Harmful Content Detection			
OffensEval2020	Offensive lang.	Tweets (Dialectal)	2,000
OSACT2020	Hate Speech	Tweets (Dialectal)	2,000
ASAD	Adult Content	Tweets (Dialectal)	10,000
ASAD	Spam	Tweets (Dialectal)	28,383
In-house	Subjectivity	News articles	297 sentences
WANLP23	Propaganda	Tweets	323
CT–CWT–22	Checkworthiness	Tweets (COVID19)	680
COVID19 Disinfo.	Factuality	Tweets	996
Unified-FC	Factuality	News articles	422 claims
ANS	Factuality	News articles	456 headlines
CT–CWT–22	Claim	Tweets (COVID19)	1,248
CT–CWT–22	Harmful content	Tweets (COVID19)	1,201
CT–CWT–22	Attention-worthy	Tweets (COVID19)	1,186
Semantic Textual Similarity (STS)			
STS2017-Track 1	STS	Image captions	250 sentence pairs
STS2017-Track 2	STS	Image captions	250 sentence pairs
Mawdoo3 Q2Q	STS QS (Q2Q)	Questions	3,715 question pairs
XNLI	XNLI	ANC	5,010 sentence pairs
Question Answering (QA)			
ARCD	QA	Wikipedia	702 questions
MLQA	QA	Wikipedia	5,335 questions
TyDi QA	QA	Wikipedia	921 questions
XQuAD	QA	Wikipedia	1,190 questions

Table 5: Summary on test sets and their sizes used in evaluation for the different textual tasks. **ANC**: American National Corpus. **Posts**∗: posts from Twitter, Youtube and Facebook. **News Cat.**: News Categorization

collection of words in a sentence, and is key to tasks such as question answering.

Datasets

ANERCorp We used the test corpus of the ANERCorp dataset (Benajiba et al., 2007; Benajiba and Rosso, 2007), which contains 316 articles, 150,286 tokens and 32,114 types, and classifies words into one of four classes (organization, location, person and miscellaneous), we used the test split of the dataset for our evaluation.

AQMAR The dataset is developed as an evaluation suite for the named entity recognition task in Arabic. It consists of a collection of 28 Wikipedia articles with 74,000 tokens. We consider the articles corresponding to the test split for our evaluation. (Schneider et al., 2012).

QASR The QASR dataset consists of 70k words extracted from 2,000 hours of transcribed Arabic speech (Mubarak et al., 2021b).

A.2 Machine Translation (MT)

The machine translation evaluation set is a rich set that covers a variety of Arabic in addition to the Modern Standard Arabic (MSA). The genera of the evaluation set also cover formal, informal, speech, and other modalities. These types and varieties allowed us to assess the system and reveal its potential and limitations. For this study, we focused on translating Arabic to English and used the datasets discussed below.

Datasets

MADAR Corpus This dataset consists of 2,000 sentences from the BTEC corpus translated to modern standard Arabic and four major dialects from 15 countries (Bouamor et al., 2018).

(Zbib et al., 2012) It is collected from the Arabic-Dialect/English Parallel Text (APT), which consists of 2,000 sentences with 3.5 million tokens of translated dialectal Arabic (Zbib et al., 2012).

Multi-dialectal Parallel Corpus of Arabic (MDC) This dataset also consists of 2,000 sentences in Egyptian, Palestinian, Syrian, Jordanian, and Tunisian dialects and their English counterparts (Bouamor et al., 2014).

The Bible It consists of 8.2k parallel sentences translated into modern standard Arabic, and to Moroccan[11] and Tunisian[12] dialects (Abdelali et al., 2019).

Media Dataset The dataset consists of 7.5 hours of recordings collected from five public broadcasting channels that cover programs with Maghrebi, Lebanese, Omani dialects, and MSA with genres involving movies, news reports, and cultural programs. The recordings were transcribed and translated by a professional translation house (Sajjad et al., 2020).

A.3 Dialect Identification

Dialect is defined as the speaker's grammatical, lexical, and phonological variation in pronunciation (Etman and Beex, 2015). Automatic Dialect Identification (ADI) has became an important research area in order to improve certain applications and services, such as ASR and many downstream NLP tasks.

Dataset For this task, we used the QADI dataset containing a wide range of country-level Arabic dialects covering 18 different countries in the Middle East and North Africa region (Abdelali et al., 2020). It consists of 540,590 tweets from 2,525 users.

A.4 Sentiment, Stylistic and Emotion Analysis

A.4.1 Sentiment Analysis

Sentiment analysis has been an active research area and aims to analyze people's sentiment or opinion toward entities such as topics, events, individuals, issues, services, products, organizations, and their attributes (Liu and Zhang, 2012; Zhang et al., 2018). This task involves classifying the content into sentiment labels such as positive, neutral, and negative.

Dataset ArSAS dataset consists of 21k Arabic tweets covering multiple topics that were collected, prepared, and annotated for six different classes of speech-act labels and four sentiment classes (Elmadany et al., 2018). For the experiments, we used only sentiment labels from this dataset.

A.4.2 Emotion Recognition

Emotion recognition is the task of categorizing different types of content (e.g., text, speech, and visual) in different emotion labels (six basic emo-

[11]The Morocco Bible Society https://www.biblesociety.ma
[12]The United Bible Societies https://www.bible.com

tions (Ekman, 1971) or more fine-grained categories (Demszky et al., 2020)).

Dataset For the emotion recognition tasks we used SemEval-2018 Task 1: Affect in Tweets (Mohammad et al., 2018). The task is defined as classifying a tweet as one or more of the eleven emotion labels, which is annotated as a multilabel (presence/absence of 11 emotions) annotation setting.

A.4.3 Stance Detection

Stance is defined as the expression of the speaker's view and judgment toward a given argument or statement (Biber and Finegan, 1988). Given that the social media platforms allow users to consume and disseminate information by expressing their views, enabling them to obtain instant feedback and explore others' views, it is important to characterize a stance expressed in a given content. Automatic stance detection also allows for assessing public opinion on social media, particularly on different social and political issues such as abortion, climate change, and feminism, on which people express supportive or opposing opinions (ALDayel and Magdy, 2021; Küçük and Can, 2020). The task involves "classification as the stance of the producer of a piece of text, towards a target as either one of the three classes: {support, against, neither} or {agree, disagree, discuss, or unrelated}" (Küçük and Can, 2020).

Datasets

Unified-FC dataset consists of claims collected from Verify.sy (false claims) and Reuters (true claims), which resulted in 422 claims. Based on these claims documents are collected using Google custom search API and filtered by computing claim-documents similarity (Baly et al., 2018b). This approach resulted in 3,042 claim-documents pairs, which are then annotated for stance (agree, disagree, discuss, unrelated) by Appen crowd-sourcing platform.

ANS Khouja (2020) developed a dataset by first sampling news titles from Arabic News Texts (ANT) corpus (Chouigui et al., 2017) and then generating true and false claims. From these claims stance (three classes – agree, disagree, other) is annotated from a pair of sentences using Amazon Mechanical Turk and Upwork. The dataset consists of 3,786 claim-reference pairs.

ArSarcasm Abu Farha and Magdy (2020) developed an Arabic sarcasm detection dataset. The dataset was created using previously available Arabic sentiment analysis datasets (Rosenthal et al., 2017; Nabil et al., 2015) and adds sarcasm and dialect labels to them. The dataset contains 10,547 tweets, 1,682 of which are sarcastic. The training set contains 8,437 tweets, while the test set contains 2,110 tweets.

ArSarcasm-v2 This dataset is an extension of the original ArSarcasm dataset published along with the paper (Farha and Magdy, 2020). ArSarcasm-v2 conisists of ArSarcasm along with portions of DAICT corpus and some new tweets. Each tweet was annotated for sarcasm, sentiment and dialect. The final dataset consists of 15,548 tweets divided into 12,548 training tweets and 3,000 testing tweets. ArSarcasm-v2 was used and released as a part of the shared task on sarcasm detection and sentiment analysis in Arabic.

A.5 News Categorization

News text categorization was a popular task in the earlier days of NLP research (Sebastiani, 2002). The idea of to assign a category $C = \{c_1, ...c_n\}$ to a document $D = \{d_1, ...d_n\}$. For the news categorization the D is a set of news articles and C is a set of predefined categories. Most often a news article can be categorized into more than one category and the models are trained in a multilabel setting. While earlier work mostly focused on news article, however, lately it has been used for the categorization of tweets in which news articles are shared as a part of a tweet.

Datasets

Social Media Posts ASND is a News Tweets dataset (Chowdhury et al., 2020b), collected from Aljazeera news channel accounts on Twitter, Facebook, and YouTube. The dataset consists of twelve categories such as art-and-entertainment, business-and-economy, crime-war-conflict, education, environment, health, human-rights-press-freedom, politics, science-and-technology, spiritual, sports, and (xii) others. We used the test split from each dataset for the evaluation.

Arabic News SANAD corpus is a large collection of Arabic news articles collected from Akhbarona, AlKhaleej, and AlArabiya (Einea et al., 2019). The dataset has separate collections gathered from different news media, each of which has six news categories; namely culture, finance, medical, politics, sports and technology.

A.6 Demographic/Protected Attributes

Demographic information (e.g., gender, age, country of origin) are useful in many different applications such as understanding population characteristics, personalized advertising, socio-cultural studies, etc. Demographic information helps governments, businesses, and organizations understand their target audiences, and plan accordingly.

A.6.1 Gender

Gender analysis can reveal important differences between male and female users such as topics of interest, gender gap, preferences, etc.

Dataset We used the ArabGend test set, which contains 1,000 names collected from Twitter (divided equally between males and females) (Mubarak et al., 2022).

A.6.2 Location

Identifying user locations is useful for many applications such as author profiling, dialect identification, recommendation systems, etc. Often, users on social media platforms, such as Twitter, declare their locations in noisy ways, and mapping these locations to countries is a challenging task.

Dataset We used the UL2C dataset, which contains 28K unique locations, as written by Arabic Twitter users, and their mappings to Arab countries (Mubarak and Hassan, 2021).

A.6.3 Name Info

Names contain important information about our identities and demographic characteristics, including factors like gender, nationality, and ethnicity. The purpose of this task is to predict the country of origin of a person name giving only their names.

Dataset We used an in-house dataset for mapping person names to World countries extracted from Wikipedia.[13]

A.7 Ethics and NLP: Factuality, Disinformation and Harmful content detection

A.7.1 Subjectivity Identification

A sentence is considered subjective when it is based on – or influenced by – personal feelings, tastes, or opinions. Otherwise, the sentence is considered objective (Antici et al., 2021). Given that the identification of subjectivity is subjective itself, therefore, it poses challenges in the annotation process by the annotator. The complexity lies due to the different levels of expertise by the annotators, different interpretations and their conscious and unconscious bias towards the content they annotate. The content can be text (e.g., sentence, article), image or multimodal content, consisting of opinionated, factual or non-factual content. The annotation typically has been done using two labels, objective (OBJ) and subjective (SUBJ).

Dataset The dataset consists of sentences curated from news articles. The dataset has been developed based on the existing AraFacts dataset (Ali et al., 2021b) that contains claims verified by Arabic fact-checking websites, and each claim is associated with web pages propagating or negating the claim. The news articles are collected from different news media. News articles were automatically parsed, split into sentences and filtered poorly-formatted sentences using a rule-based approach. The dataset has been released as a part of Task 2 of CLEF2023 CheckThat Lab (Barrón-Cedeño et al., 2023).

A.7.2 Propaganda Detection

Propaganda can be defined as a form of communication that aims to influence the opinions or the actions of people towards a specific goal; this is achieved utilizing well-defined rhetorical and psychological devices (Dimitrov et al., 2021). In different communication channels, propaganda (persuasion techniques) is conveyed through the use of diverse techniques (Miller, 1939), which range from leveraging the emotions of the audience, such as using *emotional technique* or logical fallacies such as *straw man* (misrepresenting someone's opinion), hidden *ad-hominem fallacies*, and *red herring* (presenting irrelevant data).

Dataset The dataset used for this study consists of Arabic tweets (Alam et al., 2022b) posted by different news media from Arab countries such as Al Arabiya and Sky News Arabia from UAE, Al Jazeera, and Al Sharq from Qatar, and from five international Arabic news sources Al-Hurra News, BBC Arabic, CNN Arabic, France 24, and Russia Today. The final annotated dataset consists of 930 tweets. Alam et al. (2022b) formulated the task as a multilabel and multiclass span level classification task. For this study, we used the multilabel setup.

[13] Paper is under revision.

A.7.3 Check-worthiness Detection

Fact-checking is a time-consuming and complex process, and it often takes effort to determine whether a claim is important to check, irrespective of its potential to be misleading or not. Check-worthiness detection is the first step and a critical component of fact-checking systems (Nakov et al., 2021) and the aim is to facilitate manual fact-checking efforts by prioritizing the claims for the fact-checkers. Research on check-worthiness includes check-worthiness detection/ranking from political speeches, debates, and social media posts (Nakov et al., 2022a; Shaar et al., 2021). A check-worthy claim is usually defined by its importance to the public and journalists, and whether it can cause harm to an individual, organization, and/or society.

Dataset For this study, we used the Arabic subset of the dataset released with Task 1A (Arabic) of the CLEF2022 CheckThat Lab (Nakov et al., 2022b) The dataset consists of 4,121 annotated tweets. The Arabic tweets were collected using keywords related to COVID-19, vaccines, and politics.

A.7.4 Claim Detection

Information shared in the mainstream and social media often contains misleading content. Claim detection has become an important problem in order to mitigate misinformation and disinformation in those media channels. A factual (verifiable) claim is a sentence claiming that something is true, and this can be verified using factually verifiable information such as statistics, specific examples, or personal testimony (Konstantinovskiy et al., 2021). Research on claim detection includes social media posts – text modality (Alam et al., 2021b), multimodality (Cheema et al., 2022) and news (Reddy et al., 2022).

Datasets

CT-CWT-22-Claim We used the Arabic subset of the dataset released with Task 1B of the CLEF2022 CheckThat Lab (Nakov et al., 2022a). The dataset has been annotated using a multi-question annotation schema (Alam et al., 2021a), which consists of tweets collected using COVID-19 related keywords. The dataset contains 6,214 tweets (Nakov et al., 2022b).

ANS (Khouja, 2020) This dataset consists of 4,547 true and false claims, which was developed based on Arabic News Texts (ANT) corpus. A sample of articles was modified to generate true and false claims using crowdsourcing.

A.7.5 Attention-worthiness Detection

In social media most often people tweet by blaming authorities, providing advice, and/or call for action. It might be important for the policy makers to respond to those posts. The purpose of this task is to categorize such information into one of the following categories: *not interesting, not sure, harmfullness, other, blames authorities, contains advice, calls for action, discusses action taken, discusses cure, asks a question*.

Dataset For this task, we used a subset of the dataset Task 1D of the CLEF2022 CheckThat Lab (Nakov et al., 2022a), which contains 6,140 annotated tweets.

A.7.6 Factuality Detection

Fact-checking has emerged as an important research topic due to a large amount of fake news, rumors, and conspiracy theories that are spreading in different social media channels to manipulate people's opinions or to influence the outcome of major events such as political elections (Darwish et al., 2017a; Baly et al., 2018b). While fact-checking has largely been done by manual fact-checker due to the reliability, however, that does not scale well as the enormous amount of information shared online every day. Therefore, an automatic fact-checking system is important and it has been used for facilitating human fact-checker (Nakov et al., 2021). The task typically involves assessing the level of factual correctness in a news article, media outlets, or social media posts. The content is generally judged to be of high, low, or mixed factual correctness, seven-point Likert scale[14,15] or just binary labels {yes, no} (Baly et al., 2018a; Alam et al., 2021b).

Datasets

News Articles We used the dataset developed by Baly et al. (2018a) in which false claims are extracted from verify-sy[16] and true claims are extracted from `http://ara.reuters.com`. The dataset consists of 3,042 documents.

Tweets For the claim detection from tweets, we used the same dataset (Alam et al., 2021b) discussed in A.7.4. As mentioned earlier, this dataset

[14] `https://mediabiasfactcheck.com`
[15] `https://allsides.com`
[16] `http://www.verify-sy.com`

was annotated using a multi-questions annotation schema in which one of the questions was "does the tweet appear to contain false information?". Based on the answer to this question factuality label of the tweet has been defined. The Arabic dataset contains a total of 4,966 tweets.

A.7.7 Harmful Content Detection

For the harmful content detection we adopted the task proposed in (Alam et al., 2021b; Nakov et al., 2022b) though the research on harmful content detection also include identifying or detecting offensive, hate-speech, cyberbullying, violence, racist, misogynistic and sexist content (Sharma et al., 2022; Alam et al., 2022a). For some of the those harmful content detection tasks we addressed them separately and discussed in the below sections. Alam et al. (2021b); Nakov et al. (2022b) proposed the as in the context of tweets and idea was to detect whether the content of the tweet aims to and can negatively affect society as a whole, specific person(s), company(s), product(s), or spread rumors about them. The content intends to harm or *weaponize the information*[17] (Broniatowski et al., 2018).

Dataset We used the Arabic dataset proposed in (Nakov et al., 2022b), which consists of a total of 6,155 annotated tweets.

A.7.8 Offensive Language Detection

The use of offensive language in social media has became a major problem, which can lead to real-world violence (Husain and Uzuner, 2021; Sap et al., 2019). This literature for offensive language detection mainly focused on social media content and addressing for variety of languages. The task is mainly defined as whether the content (e.g., text, image, or multimodal) is offensive or not (Chowdhury et al., 2020c).

Dataset For this task, we used the dataset from the SemEval-2020 Task 12 (OffensEval 2020) (Zampieri et al., 2020), which consists of 10,000 tweets, collected from a set of 660k Arabic tweets containing the vocative particle ("yA" – O) from April 15 to May 6, 2019.

A.7.9 Hate Speech Detection

Davidson et al. (2017) defined hate speech as "as language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group". The literature for hate speech detection defined the task as detecting hate vs. non-hate from different types of content such as text, image and multimodal (Schmidt and Wiegand, 2017; Kiela et al., 2020; Gomez et al., 2020).

Dataset For this task, we also used the OSACT 2020 dataset (Mubarak et al., 2020c), which consists of 10,000 tweets with annotated label hate-speech, not-hate-speech.

A.7.10 Adult Content Detection

Identifying this type of content is important for social media platforms to make a safe place for users. Especially this type of content poses a serious threat to other vulnerable groups (e.g., younger age groups). The task typically involves detecting and identifying whether the textual content contains sensitive/adult content or account that share such content.

Dataset We used the dataset discussed in (Mubarak et al., 2021a), which contains 10,000 tweets collected by first identifying Twitter accounts that post adult content. Tweets are manually annotated as adult and not-adult.

A.7.11 Spam Detection

Spam content in social media includes ads, malicious content, and any low-quality content (Ghanem et al., 2023). Spam detection is another important problem as such content may often annoy and mislead the users (Gao et al., 2012).

Dataset We used the dataset discussed in (Mubarak et al., 2020a) for Arabic spam detection which contains 28K tweets manually labeled as spam and not-spam.

A.8 Semantic textual similarity

A.8.1 Textual Similarity

Semantic textual similarity is a measure used to determine if two sentences are semantically equivalent. The task involves generating numerical similarity scores for pairs of sentences, with performance evaluated based on the Pearson correlation between machine-generated scores and human judgments (Cer et al., 2017a). Two tasks were conducted to gauge the similarity between 250 pairs of Arabic sentences, as well as Arabic-English sentence pairs.

[17] The use of information as a weapon to spread misinformation and mislead people.

Dataset We used SemEval-2017 Task 1 (Track 1: ar-ar and Track 2: ar-en) dataset (Cer et al., 2017a), which is a translated version (machine translation followed by post-editing by human) of SNLI dataset (Bowman et al., 2015).

A.8.2 Semantic Question Similarity

The idea of this task is to determine how similar two questions are in terms of their meaning.

Dataset We used Mawdoo3 Q2Q dataset (NSURL-2019 task 8: Semantic question similarity in Arabic), which consists of 15,712 annotated pairs of questions. Each pair is labeled as *no semantic similarity (0)* or *semantically similar(1)* (Seelawi et al., 2019).

A.8.3 Natural Language Inference (NLI)

The XNLI task, known as Cross-lingual Natural Language Inference (Conneau et al., 2018), is a widely used benchmark in the field of natural language processing (NLP). It involves determining the logical relationship between pairs of sentences written in different languages. Specifically, the task requires NLP models to determine whether a given hypothesis sentence is entailed, contradicted, or neutral in relation to a given premise sentence, across multiple languages. The XNLI task serves as a rigorous evaluation of the cross-lingual transfer capabilities of NLP models, assessing their ability to understand and reason in different languages within a multilingual context.

Dataset The dataset we used for this study is the translated version of Arabic from XNLI corpus (Conneau et al., 2018). For the annotation, 250 English sentences were selected from ten different sources and then asked the annotators to produce three hypotheses per sentence premise. The resulting premises and hypotheses are then translated into 15 languages and we used the Arabic version for this study.

A.9 Question Answering (QA)

This task involves answering questions in Arabic based on a given text[18]. For this task, we use four different datasets consisting of (passage, question, and answer) pairs.

[18]This task is also referred to as machine reading comprehension where the model is tested on its ability to extract answers from the given text

Datasets

ARCD consists of 1,395 Arabic MSA questions posed by crowd-sourced workers along with the text segments from Arabic Wikipedia. We use the test set only for our evaluation. The test set consists of 78 articles, 234 paragraphs, and 702 questions (Mozannar et al., 2019).

MLQA comprises multilingual question-answer instances in 7 languages, *English*, *Arabic*, *Simplified Chinese*, *Hindi*, *German*, *Vietnamese* and *Spanish*. We used the Arabic QA pairs from this dataset, which consist of 2389 articles, 4646 paragraphs, and 5335 questions (Lewis et al., 2019).

TyDi QA comprises 11 languages with 204K question-answer pairs. We used the data provided for the *Gold Passage task* in which a passage that contains the answer is provided and the task is to predict the span that contains the answer. We used the Arabic split of the data which contains 921 articles, 921 paragraphs and 921 questions (Artetxe et al., 2020).

XQuAD comprises 240 paragraphs and 1190 question-answers pairs from the development set of SQuAD v1.1 with their professional translations into ten languages. *Hindi*, *Turkish*, *Arabic*, *Vietnamese*, *Thai*, *German*, *Greek*, *Russian*, *Spanish* and *Chinese*. We use the the Arabic split of the data which consists of 48 articles, 240 paragraphs, and 1190 questions (Artetxe et al., 2020). We used the sQuad version of all datasets along with the official squad evaluation script.

A.10 Speech Processing

For this study, we address the speech modalities in the context of large foundation models, and we evaluate the following two tasks in this edition: (*i*) automatic speech recognition (ASR); and (*ii*) text to speech (TTS) models. In future, we will scale the speech benchmark with speech translation (ST) and spoken Arabic dialect identification spoken (ADI).

A.10.1 Speech Recognition

The primary objective of an ASR system is to transform spoken language into written text. The task itself is challenging due to the presence of variability in human speech, which can be affected by factors such as accent, speaking style, code-switching, environmental factors like channels, and background noise among others. Furthermore, the

Dataset	Task	Domain	Size
MGB2	ASR	Broadcast (MSA)	9.57 hrs
MGB3	ASR	Broadcast (EGY)	5.78 hrs
MGB5	ASR	Broadcast (MOR)	1.40 hrs
QASR.CS	ASR	Broadcast (Mixed) → Code-switching	5.90 hrs
DACS	ASR	Broadcast (MSA-EGY) → Code-switching	1.50 hrs
ESCWA.CS	ASR	Meeting (Mixed DA - ENG) → Code-switching	2.80 hrs
CallHome	ASR	Telephony (EGY)	20 phone conversations
In-house	TTS	Mixed Topics (education, health, etc)	20 sentences

Table 6: Summary on test sets and their sizes used in evaluation for the speech processing tasks.

presence of language-related challenges, including complex morphology, unstandardized orthography, and a wide array of dialects as a primary mode of communication, adds a layer of complexity to the task. Therefore to properly benchmark Arabic ASR, we covered a wide range of domains encapsulating different speaking styles, dialects, and environments. For our study, we considered broadcast news, telephony, and meeting data for MSA, Egyptian, Moroccan Arabic, etc., in both monolingual and code-switching setups.

Datasets

MGB2 consists of 9.57 hours of multi-dialect speech data that was collected from Aljazeera TV programs and manually transcribed. The data consists of a mix of Modern Standard Arabic (MSA) and various dialects, including Egyptian, Levantine, Gulf, and North African (Ali et al., 2016).[19]

MGB3 is a collection of 5.78 hours of multi-genre speech data in Egyptian dialect. The data was collected from YouTube videos and manually transcribed (Ali et al., 2017).[20]

MGB5 is a collection of 1.4 hours of speech data in Moroccan dialect. The data was collected from YouTube videos and manually transcribed (Ali et al., 2019).[21]

ESCWA.CS is a collection of 2.8 hours of speech code-switching corpus collected over two days of meetings of the United Nations Economic and Social Commission for West Asia (ESCWA) in 2019 (Chowdhury et al., 2021).[22]

QASR.CS is a collection of 5.9 hours of code-switching extracted from the Arabic broadcast news data (QASR) to test the system for code-switching. The dataset also includes some instances where the switch is between Arabic and French, however, this type of instance are very rare occurrence (Mubarak et al., 2021b).[23]

DACS is a collection of ≈ 1.5 hours of broadcast speech designed to evaluate the performance of ASR for code-switching between MSA to Egyptian dialect and vice versa (Chowdhury et al., 2020a).[24]

CallHome Egyptian is a speech corpus of telephone conversations between native speakers of Egyptian Arabic. It consists of 20 unscripted telephone conversations, each of which lasts between 5-30 minutes (Kumar et al., 2014).[25]

A.10.2 Text to Speech

Speech Synthesis a.k.a text to speech (TTS) helps users to get the written output easier and in some cases faster. Most state-of-the-art end-to-end TTS systems comprise three modules: text front-end, acoustic model, and vocoder. However, there is ongoing research to combine acoustic models and vocoder in a single neural network. Text front-end module normalizes input text by converting digits, symbols, abbreviations, and acronyms into full words, processing words with special sounds, borrowed words, etc. This task is challenging in Arabic due to missing diacritics in modern texts as explained in A.1.4. Therefore, the Arabic front-end part of the TTS is responsible for restoring the missing diacritics and text normalization.

Dataset For MSA TTS, we create the first public test dataset, which comprises 30 sentences covering different topics such as psychology, education, health, etc. The average length for each sentence

[19] https://arabicspeech.org/mgb2
[20] https://arabicspeech.org/mgb3
[21] https://arabicspeech.org/mgb5
[22] https://arabicspeech.org/escwa
[23] https://arabicspeech.org/qasr
[24] https://github.com/qcri/Arabic_speech_code_switching
[25] https://catalog.ldc.upenn.edu/LDC97S45

is 8 words. This data is used for objective and subjective evaluation for Arabic TTS.

B Model Parameters

B.1 NLP Models

We used gpt-3.5-turbo-0301 and gpt-4-0314 versions for our tasks. In addition we used Bloomz 176B 8-bit version.

B.2 Speech Models

In Table 7, we provide the details of the speech model parameters.

Model	Layers	Width	Heads	Parameters
W.Small	12	768	12	244M
W.Medium	24	1024	16	769M
W.Large-v2	32	1280	20	1550m
USM	32	1526	16	2B

Table 7: Model parameters and architecture for Large pretrained ASRs. W. stands for Open.AI's Whisper (Radford et al., 2022) and USM is Universal Speech Model from Google (Zhang et al., 2023)

C Prompts

The performance of the model is highly dependent on the prompting strategy. Designing the best prompts for each task is challenging and required several iterations. In many tasks, the output was not consistent for all instances of the datasets. For example, in many cases the model provides the desired labels however, there are cases where the model output different kind of error messages: *(i)* it's trained only on English and cannot handle Arabic texts, *(ii)* the response was filtered due to the prompt triggering Azure OpenAI's content management policy, *(iii)* it often provided extra tokens or swapped the tag (B-PER to PER-B). These resulted in an extra layer of post-processing and filtering of the evaluation dataset. Moreover, from our initial exploration, we noticed that, compared to language-specific (Arabic) prompts, English prompts (task-description) provide superior performance. Our underlying hypothesis is that with English task-description the input representations shift toward the English space that allows the model to process and understand the input better, giving better performance.[26]

[26]Note this observation aligns with other multilingual low-resource language studies.

For the segmentation task, with our initial prompt, we realized that the output was not segmented based on linguistic information but rather more Byte-Pair Encoding (BPE) like encoding. Based on that prompt is further redesigned, which resulted in a better outcome.

For factuality, disinformation, and harmful content detection tasks, the challenges were different from other tasks. One notable example is the propaganda detection task. The task requires determining whether a text snippet contains propagandistic language, and if it does, the model should detect which propaganda technique is used from a pre-defined list of techniques. Even with our best efforts to design the prompt for this task, the model still produced very unexpected responses, sometimes incomplete names of propaganda techniques, or even techniques not among the provided list. Another challenge with designing prompts for these tasks, is the issue of a task's subjectivity where providing a crisp-clear classification task definition to the model is not possible. As an example, one of our tasks is to evaluate whether a tweet is offensive towards a person or an entity. In many instances, the model predicted tweets to be offensive, while in reality they were descriptive of the tweet's author mental or physical state, or they were just repeating common negative statements or Arabic proverbs not directed at anyone indicating the model's understanding of offensiveness is not inline of our definition.

In the following sections, we report the prompts we used for different tasks.

C.1 Word Segmentation, Syntax and Information Extraction

Segmentation

```
A word can be composed of one root and one or
multiple affixes. Segment the following
sentence into its morphological constituents:
{inputSentence}"+". The output format should be
a list of tuples, where each tuple consists of
a word from the input text and its segmented
form joined by a + sign.
```

Named Entity Recognition

```
Task Description: You are working as a named
entity recognition expert and your task is to
label a given arabic text with named entity
labels. Your task is to identify and label any
named entities present in the text without any
explanation. The named entity labels that you
will be using are PER (person), LOC (location),
```

ORG (organization), MISC (miscellaneous). You may encounter multi-word entities, so make sure to label each word of the entity with the appropriate prefix ('B' for first word entity, 'I' for any non-initial word entity). For words which are not part of any named entity, you should return 'O'. Note: Your output format should be a list of tuples, where each tuple consists of a word from the input text and its corresponding named entity label. Input: {inputSentence}

POS

These are the segmentation and POS tags for a sample sentence:
فيلم جاذبية يتصدر ترشيحات جوائز الأكاديمية البريطانية لفنون الفيلم والتلفزيون
فيلم فيلم NOUN
جاذبية جاذبي + ة NOUN+NSUFF
يتصدر يتصدر V
ترشيحات ترشيح + ات NOUN+NSUFF
جوائز جوائز NOUN
الأكاديمية ال + أكاديمي + ة DET+NOUN+NSUFF
البريطانية ال + بريطاني + ة DET+ADJ+NSUFF
لفنون ل + فنون PREP+NOUN
الفيلم ال + فيلم DET+NOUN
والتلفزيون و + ال + تلفزيون CONJ+DET+NOUN

get the segmentation and POS tags for this sentence: {inputSentence}

Assign POS tag to each morphological segment within each word. group the tags for each word with +: {inputSentence}"+". The output should be in the format: [{word: label}, {word: label}]

Label the following sentence with its corresponding PENN Treebank POS Labels.
sentence: {inputSentence}
labels:

Lemmatization

---

for every word in the following sentence, write only the lemmas without diacritics in separate lines without explanation:
{inputSentence}

Diacritization

Diacritize fully the following Arabic sentence: {inputSentence}

Vowelized the following sentence: {inputSentence}. Words that can't be vowelized put them back as they were.

Parsing

Given the following features (in order: ID, Form, Lemma, CPostTag, POSTag, Features), predict the Head of each token in the following sentence, which is either a value of a related ID or 0. A value of zero means the token attaches to the virtual root node:
{inputSentence}

Dialect Identification

Write only the country code of the Arabic country in which this sentence is written in its dialect without any explanation? Write only the country code in ISO 3166-1 alpha-2 format without explanation. Write 'MSA' if the sentence is written in Modern Standard Arabic.
sentence: {inputSentence}
code:

C.2 Sentiment, Stylistic and Emotion Analysis

Sentiment analysis

Choose only one sentiment between: Positive, Negative, Neutral, or Mixed for this sentence:
sentence: {inputSentence}
label:

Emotion detection

Predict all the possible emotions in the following Arabic sentences without explanation and put them in a Python list. List of emotions are: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust
sentence: {inputSentence}
labels:

C.3 Demographic/Protected Attributes

Gender

If the following person name can be considered as male, write 'm' without explnanation, and if it can be considered as female, write 'f' without explnanation.
person name: {inputSentence}
label:

Location

Map the following locations to one of the Arab countries. Write the country code in ISO 3166-1 alpha-2 format without explanation. If the country is outside Arab countries, write 'OTHERS', and if the location cannot be mapped

```
to any country in the world, write 'UNK'
without any explanation.
location: {inputSentence}
label:
```

Name Info

```
Predict the country of citizenship of the
following person name. Write the country code
in ISO 3166-1 alpha-2 format without
explanation.
name: {inputSentence}
code:
```

C.4 Ethics and NLP: Factuality, Disinformation, Harmful content

Offensive Language

```
If the following sentence is offensive, just
write "OFF", otherwise, just write "NOT_OFF"
without explanation:
sentence: {inputSentence}
label:
```

Hate Speech

```
If the following sentence has hate speech, just
write "HS", otherwise, just write "NOT_HS"
without explanation:
sentence: {inputSentence}
label:
```

Adult Content

```
Classify the following Arabic sentence as adult
language (the language used in adult
advertisement and porno advertisement) or not
adult language without illustruation. In case
of adult language, just write "ADULT" without
explaination, and in case of not adult
language, just write "NOT_ADULT" without
explanation.
text: {inputSentence}
label:
```

Spam

```
If the following sentence can be classified as
spam or contains an advertisemnt, write 'ADS'
without explnanation, otherwise write 'NOTADS'
without explanantion.
sentence: {inputSentence}
label:
```

Subjectivity

```
Classify the sentence as subjective or
objective. Provide only label.
sentence: {inputSentence}
label:
```

Checkworthiness

```
Classify the sentence as checkworthy or not
checkworthy. Provide only the label.
sentence: {inputSentence}
label:
```

Claim detection

```
Does this sentence contain a factual claim?
Answer only by yes or no.
sentence: {inputSentence}
label:
```

Harmful content detection

```
Classify the following sentence as harmful or
not harmful. Answer only by yes or no. Provide
only label.
sentence: {inputSentence}
label:
```

Attention-worthy

```
Classify the sentence by whether it should get
the attention of policymakers. Answer by yes or
no. If the predicted label is yes then classify
the sentence into one of the following
categories: asks question, blame authorities,
calls for action, Harmful, contains advice,
discusses action taken, discusses cure, or
other.
text: {input_sample}
label:
```

C.5 Semantics

Semantic Textual Similarity

```
Given two sentences, produce a continuous
valued similarity score on a scale from 0 to 5,
with 0 indicating that the semantics of the
sentences are completely independent and 5
indicating semantic equivalence. The output
should be exactly in the form of a similarity
score.
sentence 1: {inputSentence1}
sentence 2: {inputSentence2}
score:
```

Natural Language Inference

```
You are provided with a premise and a
hypothesis. Your task is to classify the
hypothesis as true (entailment), false
(contradiction), or unknown (neutral) based on
the given premise. The output should be true,
false or unknown.
premise: {inputSentence1}
hypothesis: {inputSentence2}
output:
```

Classification (Question Similarity)

```
Are the following two questions semantically
similar? The output should be exactly either
yes or no.
question 1: {inputQuestion1}
question 2: {inputQuestion2}
label:
```

C.6 Question answering (QA)

```
Your task is to answer questions in Arabic
based on a given context.
Note: Your answers should be spans extracted
from the given context without any
illustrations.
You don't need to provide a complete answer.
context:{context}
question:{question}
answer:
```

D Post-processing

Post-processing was needed for almost all tasks in order to match gold labels, which include reformatting the output handling exceptions, missing values, and unexpected values. Much like NLP tasks, post-processing the transcription output from the speech models is an important step. We noticed that the performance of the Whisper models is highly dependent on the post-processing. As the models (Whisper family) are trained with massive dataset created by weak supervision, the output is quite noisy and needs extra care for post-processing. In this study, we opt for a simple post-processing pipeline so that the process is not overfitted to task-based data styles.

E Benchmarks on Arabic: Details

Noteworthy among them is ORCA (Elmadany et al., 2022), a large-scale benchmark that incorporates 60 diverse datasets organized into seven comprehensive task clusters. This large-scale organization allows for a more in-depth and diverse analysis of model performance across a multitude of language tasks including but not limited to sentence classification, text classification, structured prediction, semantic similarity, natural language inference, question-answering, and word sense disambiguation.

AraBench (Sajjad et al., 2020) is an evaluation suite for dialectal Arabic-to-English machine translation. It offers a wide range of dialect categories including 4 coarse, 15 fine-grained, and 25 city-level dialects from various genres like media, chat, and travel. It also provides robust baselines that utilize different training methods like fine-tuning, back-translation, and data augmentation.

The ***ALUE*** (Seelawi et al., 2021) benchmark offers 8 curated tasks and private evaluation datasets, covering areas like emotion classification, hate speech, and fine-grained dialect identification. AraBicBERT tops the performance in 7 of these 8 tasks, with evaluations also including BERT variants with AraVec and FastText models.

ARLUE (Abdul-Mageed et al., 2021) benchmark employs 42 datasets for six task clusters to evaluate multi-dialectal Arabic language understanding, featuring BERT and XLM model variants. Fine-tuned models utilizing ARLUE lead the performance in all six clusters.

As shown in Table 8, Our study provides a comprehensive evaluation platform that advances the current benchmarks by presenting 33 distinct tasks over 61 datasets, which is the most extensive task coverage among current benchmarks. Unlike the AraBench, which focuses exclusively on Arabic-to-English translation tasks, and ALUE and ARLUE, which have a narrower task focus or a lesser number of tasks, LAraBench provides a broader scope of evaluation tasks. This benchmark encompasses a multitude of language tasks that are paramount to understanding the robustness and generalizability of language models. Furthermore, LAraBench distinguishes itself by not only including text modality but also speech modality, thereby increasing the robustness and utility of our benchmark. Additionally, we successfully implemented GPT-3.5 and GPT-4, demonstrating its compatibility with cutting-edge language models.

Notably, the models employed in LAraBench have displayed comparable performance with the SOTA models, attesting to its robustness and high standard of evaluation. While SOTA models generally outperform LLMs, our benchmark reveals that these LLMs can close the performance gap in certain tasks, particularly when increasing prompt complexity and transitioning from zero-shot to few-shot learning. This highlights LAraBench's utility not only as a tool for model evaluation but also as an instrumental platform for identifying tasks under which LLMs might be able to match or even surpass SOTA performance. This benchmark serves as a challenging testbed for future language models and contributes to the advancement of Arabic language understanding models.

Reference	# tasks	# datasets	Fine-tuned Models	Zero-shot GPT-3.5	Few-shot GPT-3.5	Zero-shot GPT-4	Few-shot GPT-4	Zero-shot Bloomz	SOTA Comp.	Modality
AraBench (Sajjad et al., 2020)	1	6	Seq2Seq (transformer)	✗	✗	✗	✗	✗	✓	T, S
ARLUE (Abdul-Mageed et al., 2021)	13	42	ARBERT, MARBERT	✗	✗	✗	✗	✗	✓	T
ALUE (Seelawi et al., 2021)	8	8	AraBERT, mBERT	✗	✗	✗	✗	✗	✓	T
ORCA (Elmadany et al., 2022)	29	60	mBERT, ARBERT, CamelBERT, MARBERT	✗	✗	✗	✗	✗	✓	T
GPTAraEval (Tawkat et al., 2023)	32	60	✗	✓	✓	✗	✓	✗	✗	T
LAraBench (Ours)	33	61	✗	✓	✗	✓	✓	✓	✓	T, S

Table 8: A comparison with prior studies. T: Text, S: Speech.

F Extended Experiments and Results

In this section, we provide extended versions of the results reported earlier in the paper.

F.1 Random Baseline

For different tasks, we used different approaches to compute random baseline, as discussed below.

- **Segmentation:** We first randomly decide how many segments a token should have (between 0, 1 and 2), and then randomly split the characters of that token into the chosen number of segments.
- **Lemmatization:** We first randomly decide the length of the lemma, and then randomly divide the remaining length between a prefix and suffix.
- **Diacritization:** we randomly choose between 9 choices for every character (8 diacritics and 1 choice for no diacritic).
- **QA:** Randomly select a span of tokens from the given context of each question.
- **Others (Multiclass and multilabel classification tasks):** For multiclass classification, we randomly assign a label to the test instance, with label selection based on the labels from the training set. For multilabel classification, which requires assigning multiple labels from a predefined set, both the number of labels and their selection were random, and these were assigned to the test instance.

F.2 Extended Few-shot Results

We conducted experiments using GPT-4 by incrementally increasing the number of shots. For this purpose, we chose one task from each of the seven groups listed in Table 1 in the paper. We tested the models using 3, 5, and 10 shots. For each task, we observed a general trend of increasing performance, with the exception of the gender task. On average, performance improved from 0.656 in the 0-shot setting to 0.721 in the 10-shot setting. The results are presented in Table 9. To provide a clear overview of the comparison across different few-shot scenarios, we present the average performance in Figure 2.

Task Name	Metric	0-shot	3-shot	5-shot	10-shot
NER	Macro-F1	0.355	0.420	0.426	0.451
Sentiment	Macro-F1	0.569	0.598	0.619	0.639
News Cat.	Macro-F1	0.667	0.594	0.674	0.723
Gender	Macro-F1	0.868	0.980	0.931	0.937
Subjectivity	Macro-F1	0.677	0.745	0.740	0.771
XNLI (Arabic)	Acc	0.753	0.774	0.789	0.809
QA	F1 (exact match)	0.705	0.704	0.718	0.716
Average		**0.656**	**0.688**	**0.700**	**0.721**

Table 9: Results from few-shot experiments over seven tasks with GPT-4.

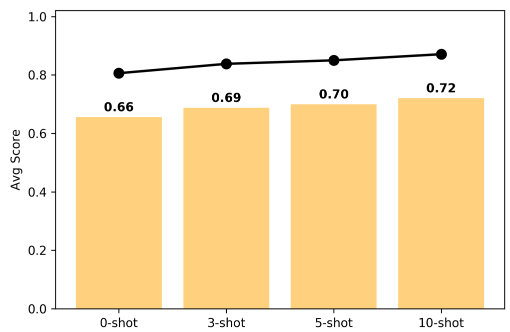

Figure 2: An average performance comparison (over seven tasks) of different few-shot experiments with GPT-4.

F.3 Native Language Prompts

We have conducted experiments using Arabic prompts for the *seven selected tasks*. The Arabic prompts were created by native Arabic speakers. The results are reported in Table 10. Using the Arabic prompts, three out of the seven tasks outperformed their counterparts that used English prompts, two underperformed, and one showed equivalent performance. This finding partially supports the findings reported by Ahuja et al. (2023), which states that "the monolingual prompting setup outperforms the cross-lingual prompting strategy". However, they also report that using Davinci-003, the English prompts yield better results than their translated version in the native language.

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**

Table 10: Results from GPT-4 using zero-shot prompts in both English and native languages.

F.4 Semantic vs. Syntactic Task Differences

We computed the performance difference between POS and MT, as shown in Table 11. The gap between SOTA and the three LLMs for POS (a syntactic task) is considerably larger than for MT (a semantic task). Moreover, the performance gap is much lower for semantic tasks compared to syntactic tasks, on average, across the three LLMs, as depicted in Table 11. This implies that these models might be better equipped to encode and express semantic information than to handle specific syntactic phenomena in their inputs.

	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

Table 11: Average performance difference between semantic and syntactic tasks.

F.5 Data Contamination Assessment

The presence of test data from standard downstream NLP tasks in the training dataset of pretrained LLMs' may effect the evaluations. It is important to have blind test-sets to reliably assert that the models are not merely memorizing data patterns but have truly acquired the ability to generalize. Identifying whether the data has been contaminated or not is a challenging problem. In our study, we have used the dataset that has been released after September 2021, which is a cut-off date for OpenAI's GPT models.[27]. The tasks include CT–CWT–22 tasks (Checkworthy, Claim, Harmful content, and Attention-worthy) introduced in 2022. Consequently, for these specific tasks, the potential for data contamination is none. Both GPT-3.5 and GPT-4 (in zero-shot and 3-shot scenarios) demonstrate results closely aligned with the state-of-the-art, mirroring trends seen in other 2021 test sets. In addition, the dataset for the subjectivity task is our in-house developed dataset, created at the end of 2022.

To further validate whether evaluation datasets have been exposed to the LLMs, we assessed various datasets using the methodology outlined in (Golchin and Surdeanu, 2023). This approach employs tailored instructions to ascertain if a model has encountered particular evaluation data. When applying this methodology to GPT-4, across a representative array of datasets, namely *(1)* Sentiment (ArSAS 2018), *(2)* Emotion (SemEval-2018 Task 1, Arabic), *(3)* Sarcasm (ArSarcasm-OSACT2020, ArSarcasm-v2-WANLP2021), *(4)* News Category (ASND 2020), *(5)* Gender (Arap-Tweet 2022), *(6)* Subjectivity (In-house 2022), *(7)* XNLI 2020 (Arabic), *(8)* Question Answering (XQuAD 2019), no instances were generated by GPT-4 from these datasets. For none of the 9 datasets and 8 tasks, GPT-4 could produce any example from it. Thus, based on these experiments, we can conclude that the Arabic datasets for different tasks are not included in the training data of GPT models.

F.6 Machine Translation (MT)

In Table 12, we report detailed results for MT, considering both dialect and city levels.

[27] https://platform.openai.com/docs/models/overview

Dataset	Dialect	SC	City	#Sent	BloomZ	Zero-shot GPT-3.5	Zero-shot GPT-4	SOTA
APT	LEV	lv	-	1000	11.38	18.55	17.77	21.9
APT	Nile	eg	-	1000	12.95	21.58	18.99	22.6
MADAR	Gulf	iq	Baghdad	2000	30.99	32.47	34.83	29.1
MADAR	Gulf	iq	Basra	2000	29.63	32.92	34.72	29
MADAR	Gulf	iq	Mosul	2000	29.17	30.82	35.32	31.3
MADAR	Gulf	om	Muscat	2000	39.91	39.37	39.9	39.5
MADAR	Gulf	qa	Doha	2000	31.1	33.6	33.62	29.3
MADAR	Gulf	sa	Jeddah	2000	40.37	42.62	42.69	29.4
MADAR	Gulf	sa	Riyadh	2000	27.73	32.51	33.71	40.7
MADAR	Gulf	ye	Sana'a	2000	29.79	32.48	34.63	31.4
MADAR	LEV	jo	Amman	2000	35.56	35.09	36.24	35.1
MADAR	LEV	jo	Salt	2000	34.54	35.78	37.54	34.9
MADAR	LEV	lb	Beirut	2000	24.01	26.14	28.95	23.7
MADAR	LEV	ps	Jerusalem	2000	34.02	35.22	35.5	33.6
MADAR	LEV	sy	Aleppo	2000	30.92	34.09	35.47	34.3
MADAR	LEV	sy	Damascus	2000	29.1	34.19	37.74	33.1
MADAR	MGR	dz	Algiers	2000	23.13	22.43	25.95	21.3
MADAR	MGR	ly	Benghazi	2000	25.41	26.99	30.12	32
MADAR	MGR	ly	Tripoli	2000	30.05	32.82	38.63	25.9
MADAR	MGR	ma	Fes	2000	23.73	22.53	26.15	29.9
MADAR	MGR	ma	Rabat	2000	31.02	31.95	34.71	23.1
MADAR	MGR	tn	Sfax	2000	15	15.93	20.74	13.8
MADAR	MGR	tn	Tunis	2000	16.79	14.69	18.51	16
MADAR	MSA	ms	-	2000	42.33	37.55	37.67	43.4
MADAR	Nile	eg	Alexandria	2000	29.24	32.05	32.46	38.3
MADAR	Nile	eg	Aswan	2000	39.97	41.77	42.42	30.4
MADAR	Nile	eg	Cairo	2000	32.79	32.77	32.69	32.9
MADAR	Nile	sd	Khartoum	2000	37.48	41.27	44.13	39
MDC	LEV	jo	-	1000	10.43	17.75	16.96	17.7
MDC	LEV	ps	-	1000	9.32	15.72	14.22	15.3
MDC	LEV	sy	-	1000	10.24	18.66	16.96	19.9
MDC	MGR	tn	-	1000	8.28	14.46	14.2	13.9
MDC	MSA	ms	-	1000	15.75	21.05	19.34	20.4
Media	Gulf	om	-	467	14.22	22.68	22.76	19.6
Media	LEV	lb	-	250	7.54	17.65	16.65	16.8
Media	MGR	ma	-	526	4.87	11.58	10.2	9.6
Media	MSA	ms	-	637	22.14	37.87	34.41	29.7
Media	MSA	ms	-	621	19.17	32.8	32.73	35.6
QAraC	Gulf	qa	-	6713				16
Bible	MGR	ma	-	600	16.34	16.16	15.14	28.8
Bible	MGR	tn	-	600	17.83	17.27	15.43	29.2
Bible	MSA	ms	-	600	24.37	23.96	18.38	33.2
Bible	MSA	ms	-	600	21.44	20.2	16.68	29.2

Table 12: Results (BLEU score) on machine translation for different datasets using zero-shot prompts. #Sent. indicates number of sentences in test set. SOTA results are reported in (Sajjad et al., 2020).