On the Utility of Pretraining Language Models on Synthetic Data

Alcides Alcoba Inciarte [λ] **Sang Yun Kwon** [λ]
El Moatez Billah Nagoudi[λ] **Muhammad Abdul-Mageed**[λ,σ,ξ]
[λ]The University of British Columbia [σ]MBZUAI [ξ]Invertible AI
{alcides.alcobainciarte@,skwon01@student.,moatez.nagoudi@,muhammad.mageed@}ubc.ca

Abstract

Development of pre-trained language models has predominantly relied on large amounts of datasets. However, this dependence on abundant data has limited the applicability of these models in low-resource settings. In this work, we investigate the utility of exploiting synthetic datasets acquired from different sources to pre-train language models for Arabic. Namely, we leverage data derived based on four different methods: optical character recognition (OCR), automatic speech recognition (ASR), machine translation (MT), and generative language models. We use these datasets to pre-train models in three different architectures: encoder-only ($\text{BERT}_{\text{Base}}$), encoder-decoder (T5), and decoder-only (GPT-2). We test the capabilities of resulting models on Arabic natural language understanding (NLU) tasks using the ORCA benchmark. Our results show that utilizing synthetic data can achieve performance comparable to, or even surpassing, those trained on gold data. For example, our model based on a GPT-2 architecture trained on a combined synthetic dataset surpasses the baseline model $\text{ARBERT}_{\text{v2}}$. Overall, our models pre-trained on synthetic data demonstrate robust performance across various tasks. This highlights the potential of synthetic datasets in augmenting language model training in low-resource settings.

1 Introduction

The quality and quantity of training data significantly affect model performance in many natural language processing (NLP) tasks. Often, large amounts of training data are essential for neural models to achieve strong performances across a wide range of tasks (Sutskever et al., 2014; Bowman et al., 2015). In areas such as natural language understanding (NLU) (Wang et al., 2022) and neural machine translation (NMT) (Maguerresse et al., 2020) consistent improvements with

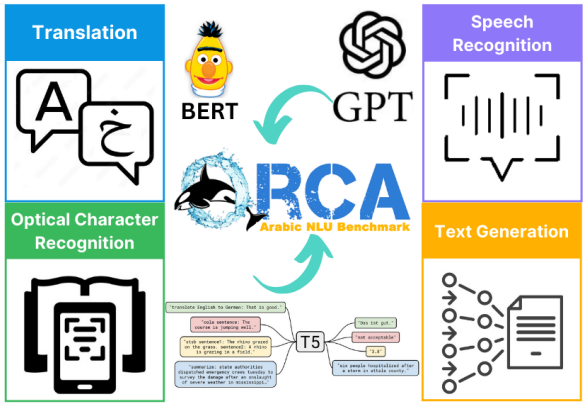

Figure 1: Our four synthetic data sources— optical character recognition (OCR), automatic speech recognition (ASR), machine translation (MT), and text generation (TG) models—used for pre-training Arabic models in different architectures - *encoder-only* (BERT), *decoder-only* (GPT-2), and *encoder-decoder* (T5), tested on the ORCA benchmark.

increased training datasets have been shown, underscoring the significance of abundant training data. However, this becomes particularly challenging in low-resource settings where preparing extensive annotated data is costly, time-consuming, and infeasible to relabel every example whenever new data comes in. This highlights the need for research towards learning with limited labeled data for various NLP tasks (Chen et al., 2023).

There has been growing interest in data augmentation (DA) strategies to generate new data by modifying existing data through transformations (Feng et al., 2021; Wei and Zou, 2019). Moreover, with the recent surge of pre-trained Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs), capabilities in generating data across domains such as images, audio, and text have expanded (Alemohammad et al., 2023). Renowned generative models like ChatGPT (for text) (Brown et al., 2020) and Stable Diffusion

Proceedings of The Second Arabic Natural Language Processing Conference, pages 265–282

(for images) (Rombach et al., 2022) are now readily accessible via APIs, making the generation of high-quality synthetic data across various domains more accessible (Veselovsky et al., 2023).

In this study, we investigate how training data from various sources and methods affect the performance of pre-trained language models (PLMs) on Arabic NLU tasks on the ORCA benchmark (Elmadany et al., 2022). We primarily focus on the Arabic language due to its complex grammar, rich morphology, and limited resources, which present significant challenges. We compare the performance of three different model architectures: *encoder-only* ($\text{BERT}_{\text{Base}}$ architecture), *encoder-decoder* (T5 architecture), and *decoder-only* (GPT-2 architecture), trained on data extracted through automatic speech recognition (ASR), optical character recognition (OCR), machine translation (MT), and text generation (TG). This evaluation aims to determine whether synthetically generated text data can match or potentially replace gold standard data. Our contributions in this paper are as follows:

1. We showcase the importance of leveraging different text extraction and text generation methods to obtain valuable pre-training data.

2. We compare the performance of models pre-trained on gold standard text data against models pre-trained purely on synthetic data.

3. We highlight the limitations of relying on sole and non-diverse data sources, such as solely Wikipedia, in model pre-training.

2 Related Works

Data Augmentation. Data augmentation has become a cornerstone in enhancing training datasets without the need for additional data collection, due to its efficiency, simplicity, and cost-effectiveness. This has led to a surge in its adoption across various domains (Hernández-García and König, 2018; Shorten and Khoshgoftaar, 2019). Earlier methods include *rule-based* approaches, such as token-level random perturbations (Wei and Zou, 2019), and predetermined transformations without model components (Zhang et al., 2015). These methods have demonstrated improved outcomes in various text classification tasks (Xie et al., 2020). With the recent development of large language models (LLMs), the potential for more advanced and adaptable data augmentation techniques has significantly increased (Ding et al., 2020), allowing for greater gains in model performance and robustness across different natural language processing tasks (Perez and Wang, 2017; Yang et al., 2022; Zhuo et al., 2023).

Multimodal Augmentation. Multimodal Augmentation (MA) leverages generative advancements in AI to synthesize data across multiple domains, including image, audio, text, and video (Bewersdorff et al., 2024). These techniques either generate entirely new content or modify existing content to maintain semantic relationships, similar to how traditional Data Augmentation (DA) techniques adapt training examples (Liu et al., 2022). The notable growth and results from publicly released generative models have established MA as a critical component in numerous research sectors, particularly those involving multimodal data (Alemohammad et al., 2023). MA has demonstrated significant effectiveness in various domains, such as developing SoTA image models like Stable Diffusion using the synthetic LAION-5B dataset (Schuhmann et al., 2022). In the vision domain, innovations like MixGen excel in joint data augmentation by combining images and text to create new image-text ensembles (Hao et al., 2023). In the audio domain, advancements such as AudioLM push the boundaries of speech synthesis, capturing the unique voice and nuances of previously unheard speakers (Borsos et al., 2023). These achievements highlight the convergence of self-supervised representation learning and language modeling, pointing to a promising future for the field (Liu et al., 2022).

Arabic NLP. Arabic Natural Language Processing (NLP) faces significant challenges due to its characteristics as a low-resource language, which has historically hindered achieving SoTA results due to data scarcity (Abdul-Mageed et al., 2020b; Belkebir and Habash, 2021). Despite this, there have been multiple efforts to address these challenges. For instance, the KIND dataset collects nuanced dialectal data through social collaboration (Aloui et al., 2024). Similarly, the ArQuAD dataset, a comprehensive, expert-annotated corpus designed for Arabic machine reading comprehension, emphasizes the urgent need for substantial, high-quality datasets to advance Arabic NLP (Obeidat et al., 2024). Furthermore, recent efforts in specific tasks such as Grammatical Error Correction (GEC) (Kwon et al., 2023; Solyman et al., 2021) and task classification (Refai et al., 2023), as well as efforts for evaluation (Khondaker et al., 2023)

Source	Num Sentences	Num Words
Arabic Wikipedia	8.8M	102M
Translated English Wikipedia	6M	103M
OCR	11.6M	101M
ASR	1.9M	25M
TG	8.8M	110M

Table 1: Gold and synthetic data statistics.

and model development (Nagoudi et al., 2022) for Arabic, illustrate ongoing endeavors to enhance the linguistic capabilities and robustness of NLP models for the Arabic language.

3 Datasets & Experimental Setup

3.1 Training Datasets

We consolidate four unique datasets acquired through four distinct methods: *(1) machine translation* (**MT**), *(2) optical character recognition* (**OCR**), *(3) automatic speech recognition* (**ASR**), and *(4) text generation* (**TG**). While having all generated data based on the same corpus or covering the same genre would ensure consistency and fair comparison, this was challenging for certain data extraction techniques. MT and TG datasets align with this criterion, as they both utilize Wikipedia content. For OCR data, we aimed for consistency by selecting academic works from the Arabic public library Hindawi. However, for ASR, achieving this level of consistency was difficult due to the limited availability of domain-specific data. In future work, we aim to improve consistency across data sources where feasible.

We introduce each dataset and outline the extraction methods below. Details on the data statistics can be found in Table 1.

Gold Data. We collect the most recent articles from Arabic Wikipedia dumps and extract raw text at the article level using wikiextractor (Attardi, 2015) as our primary data source. This dataset remained unaltered throughout our study, serving as our baseline dataset for comparison.

MT Data. Following the same approach as the gold data, we gather a subset of English articles from English Wikipedia dumps and translate them into Arabic using Meta's No Language Left Behind model (NLLB) (Costa-jussà et al., 2022). The articles were randomly sampled to ensure a diverse and representative subset of content. After translation, we remove any text that contains repeated duplicates of a word to avoid potential issues during training.

OCR Data. We collect and preprocess 3,000 books from the Arabic public library Hindawi (Ali, 2023). Subsequently, we employ Google Tesseract (Smith, 2007) and perform OCR on all the PDF books, extracting raw text at the page level, then segmenting it into paragraphs.

ASR Data: We utilize a wide range of Arabic speech datasets, covering both Modern Standard Arabic (MSA) and various Arabic dialects. For datasets lacking transcriptions, we transcribe the audio using the Arabic whisper model from Talafha et al. (2023). For those datasets that include transcriptions, we consider them only if they were transcribed via ASR systems. The datasets are detailed below:

Multi Genre Broadcast-2 (MGB-2) (Ali et al., 2016) includes 1,200 hours from Aljazeera TV programs. These programs were manually captioned without specific timing details. The QCRI Arabic ASR system (Dalvi et al., 2017) was employed to transcribe all the output, facilitating the alignment of manual captions to generate speech segments for speech recognition training.

Mozilla's Common Voice Project (Ardila et al., 2019) features recordings of sentences in MSA by contributing volunteers. Each recording is validated by at least two users. We use all 11 versions of the dataset, which total 690 hours of audio.

Arabic Speech Corpus (Halabi, 2016) is dedicated to MSA speech synthesis. It features over 3.7 hours of MSA speech, with both phonetic and orthographic transcriptions aligned with the recorded speech at the phoneme level.

Massive Arabic Speech Corpus (Al-Fetyani et al., 2021) contains over 1000 hours of speech from more than 700 YouTube channels. It is multi-genre, multi-regional, and multi-dialect.

TG Data. We take samples from the Gold Data and leverage Jasmine (Nagoudi et al., 2022), an Arabic GPT model, to generate synthetic data by asking it to complete sentences given their beginnings. We chose to use the Jasmine model for generating synthetic data because it is specifically tailored for Arabic, ensuring that the generated text aligns with the linguistic nuances and characteristics of the language. By leveraging Jasmine, we can produce high-quality completions that are coherent and contextually appropriate for Arabic. Additionally, to maintain the quality and diversity of the generated dataset, we implemented measures to prevent repetition. This approach enhances the richness of the synthetic data, which is crucial for robust model

training and evaluation. Examples of the prompts and generated text can be found in Appendix A.1.

Combined Datasets For our final experiment, we compare the performance of the models trained on all the subsets of data used above. We train models with different combinations of the Gold, MT, OCR, ASR, and TG data. Table 2 outlines the different data combinations.

Name	Components
C1	MT + OCR
C2	MT + OCR + ASR
C3	MT + OCR + ASR + TG
C4	MT + OCR + ASR + TG + Gold

Table 2: Configuration of combined datasets.

3.2 ORCA Benchmark

We evaluate our models using the ORCA benchmark, a diverse NLU benchmark for the evaluation of language models in Arabic NLU tasks (Elmadany et al., 2022). The benchmark includes both Modern Standard Arabic (**MSA**) and Dialectal Arabic (**DA**), ensuring broad linguistic spectrum and geographical representation. It is comprised of 60 publicly available datasets and is segmented into seven distinct task clusters: Sentence Classification (**SC**), Text Classification (**TC**), Structured Prediction (**SP**), Semantic Text Similarity (**STS**), Natural Language Inference (**NLI**), Question-Answering (**QA**), and Word Sense Disambiguation (**WSD**). These are further organized into 29 tasks within the mentioned task clusters. For evaluation, a macro-average of scores (F_1) is taken across all tasks and task clusters, with each task given equal weight, referred to as the ORCA$_{score}$. For all individual tasks, the metric is F_1, except for tasks within the STS clsuter and emotion-Reg task within the SC cluster, which use spearman correlation. Samples from distinct tasks are in Table 3. Details regarding the different task clusters, tasks, and data splits can be found in Table 4.

3.3 Models & Pre-training

Common Training Configuration. To pre-train all models outlined below, we run for ten epochs, using a batch size of 16, a learning rate of $5e-3$, and a maximum sequence length of 512.

Baseline Setting. The ARBERT$_{v2}$ (Abdul-Mageed et al., 2020a) model (164 million parameters) serves as baseline for comparative analysis in our experiments. We choose ARBERT$_{v2}$, as it achieved the highest ORCA$_{score}$ on the ORCA benchmark as is shown in the ORCA leaderboard.[1] This ensures a consistent and reliable metric for comparison across our studies.

Encoder-Only. For the encoder-only model, we use the BERT$_{Base}$ model (Devlin et al., 2018) (110 million parameters). We pre-train BERT$_{Base}$ following the methods outlined by Devlin et al. (2018). We employ a BERT Word Piece Tokenizer and adopt the default network architecture of the BERT$_{Base}$ model. We adhere to the hyperparameters and architectural details specified in the original paper, including the tokenizer's vocabulary size, minimum word frequency, and the configuration of special tokens. This approach allows us to maintain alignment with established benchmarks while focusing on the innovations introduced in our work.

Decoder-Only. We use the GPT-2 (Radford et al., 2019) model (1.5 billion parameters) for our decoder-only experiment. To pre-train the GPT-2 model, we use the configured vocabulary size of $50,257$, including special tokens such as <s>, </s>, <pad>, <unk>, and <mask>.

Encoder-Decoder. The T5 (Raffel et al., 2020) model (220 million parameters) serves as the baseline model for our encoder-decoder model experiment. We employ a Sentence Piece Tokenizer with a vocabulary size of $32,128$ and special tokens <unk>, <s>, </s>, and <pad>.

Model Selection. We acknowledge the differences in model sizes across the architectures used in our experiments. The choice of ARBERT$_{v2}$, BERT$_{Base}$, GPT-2, and T5 is driven by their popularity and established benchmarks in the field. Our primary goal is to utilize base "representative" models to ensure the relevance and applicability of our findings. In future work, we aim to focus on similarly sized models for a fairer comparison across different architectures.

3.4 Fine-Tuning on the ORCA Benchmark

Encoder and Decoder Only Models. After pre-training, we fine-tune both the BERT$_{Base}$ and GPT-2 model on the ORCA benchmark, covering all 29 tasks across seven task clusters encompassing 60 datasets. Examples of the specific ORCA tasks are presented in Table 3 while further details on

[1] https://orca.dlnlp.ai/main_leaderboard

Task	Content	Label
Age	والله اني ماسويت شيء يا اسطوره تعالي اضحكك	Under 25
Arabic-NER	وحد صفوف اهل الريف (شمال المغرب) .	O O O O B-LOC O O B-LOC O O
Ans-Stance	***Sentence 1***: استقرار الذهب مع استمرار ضعف الدولار ***Sentence 2***: ضعف الدولار يدعم باتجاه استقرار الذهب	agree
Topic	أعلنت الشرطة في سياتل بالولايات المتحدة الأمريكية، لاحقتا طائرة أقلعت من مطار سياتل F-15 إن مقاتلتا تاكوما،فجر الجمعة، دون الحصول على إذن بالإقلاع،	International News
MQ2Q	***Sentence 1***: ما هو موقع اليمن؟ ***Sentence 2***: أين تقع قارة أوروبا؟	No
QA	***Question***: من هو مكتشف المرو أو الكوارتز ؟ ***Context***: المرو أو الكوارتز هو معدن يعود اكتشافه إلى الفرنسي (بيير كوري) حيث لاحظا ظاهرة1880وأخوه (جاك) اللذان كانا يدرسان عينة من الرمل في سنة غريبة، وهي انه عند تعريض الكوارتز (ثاني أكسيد السيليكون) لجهد آلي فإنه يتولد تيار كهربائي، وبالعكس ففي حال تعرضت بلورة الكوارتز لمجال كهربائي، فإنها تتذبذب وتهتز بتردد معين،كما وجد أن هذا الاهتزاز والتذبذب يتسم بالانتظام والدقة العالية. هذه الظاهرة والتي عرفت بالبيزوكهربائية، مكنت الباحثين من تصنيع الكثير من الأجهزة الحساسة، من أهمها الساعات المصممة لقياس الوقت بدقة عالية، حيث بلغ نصيب الكوارتز في صناعة الساعات أكثر من ٨٥ من سوق الساعات العالمية. ويعود أول نموذج لساعة مصنوعة من الكوارتز إلى سنة ١٩٦٧ حيث تم إنتاج هذه الساعة من قبل الباحثين في مركز الساعات الإلكترونية في نويشتل في سويسرا، وفي سنة ١٩٦٩ تمت صناعة أول ساعة كوارتز في اليابان من قبل سيكو اليابانية تحت اسم أسترون.	(بيير كوري) وأخوه (جاك)

Table 3: Samples from ORCA tasks.

ORCA's diverse task clusters, tasks, and data splits are provided in Table 4. During training, tasks were categorized into token-level tasks, classification tasks, and question-answering tasks, as outlined in the original paper (Elmadany et al., 2022). Each model is fine-tuned according to this categorization to handle the respective tasks. We fine-tune each model for ten epochs using a batch size of 16 and a learning rate $5e-5$. We then select the best-performing model based on our Dev data for blind-testing on the Test sets. We report the mean score of the three runs, along with its standard deviation. Results of the Dev set can be found in Appendix A.4.

Encoder-Decoder Model. Since the T5 model is inherently a generative model, we had to adapt the model for classification tasks on the ORCA benchmark. Given that the ORCA benchmark does not naturally accommodate encoder-decoder architectures, we tailor the tasks to suit a generative format. To achieve this, we use the Dolphin benchmark as a framework (Nagoudi et al., 2023), which is specifically designed for natural language generation (NLG) tasks. To maintain consistency across all model types, including $\text{BERT}_{\text{Base}}$ and GPT-2, we modify the dataset structures while keeping the same sets of inputs and labels. For tasks that involve a direct single input-output relationship, the original structure is preserved. However, for more complex tasks that require multiple inputs, such as the Word Sense Disambiguation (WSD) task, we concatenate the inputs into a unified context. The consolidated context is then used as the input during the fine-tuning phase of the T5 model, ensuring a consistent approach across various task types.

4 Results

4.1 Encoder Only Model

Table 5 presents the performance of the $\text{BERT}_{\text{Base}}$ model. The model trained on ASR data achieves the highest ORCA_{score}, surpassing the scores of the gold dataset and significantly enhancing performance, particularly in the adult, dialect-binary, and emotion-reg tasks, with improvements of 15.20,

Cluster	Task	Level	#Data	Train	Dev	Test
SC	SA	Sent	19	50K	5K	5K
	SM	Sent	11	50K	5K	5K
	Dia-b	Sent	2	50K	5K	5K
	Dia-r	Sent	3	38.5K	4.5K	5K
	Dia-c	Sent	4	50K	5K	5K
	CL	Sent	1	3.2K	0.9K	0.4K
	MG	Sent	1	50K	5K	5K
SP	NER	Word	2	5.2K	1.1K	1.2K
	POS	Word	2	5.2K	1.1K	1.2K
TC	Topic	Doc	5	47.5K	5K	5K
QA	QA	Parag	4	101.6K	517	7.4K
STS	STS-reg	Sent	1	0.8K	0.2K	0.2K
	STS-cls	Sent	1	9.6K	1.2K	1.2K
NLI	XNLI	Sent	1	4.5K	0.5K	2.5K
	FC	Doc	2	5K	1K	1K
WSD	WSD	Word	1	21K	5K	5K
Total			**60**	487.1K	46.0K	55.1K

Table 4: The different task clusters, tasks, and data splits in the ORCA benchmark. **SC**: Sentence Classification. **SP**: Structured Prediction. **TC**: Topic Classification. **STS**: Textual Semantic Similarity. **NLI**: Natural Language Inference. **QA**: Question Answering. **SM**: Social Meaning. This table is adapted from the ORCA Elmadany et al. (2022).

15.02, and 17.36 points respectively. The model trained on TG data demonstrates strong performance within the SC and SP task cluster, achieving the highest scores in adult (87.19), age (41.92), dialect-country (23.99), sarcasm (65.74), arabic-ner (66.35), and aqmar-ner (57.68). Notably, this model also outperforms the gold dataset in ORCA$_{score}$. In contrast, the model trained on both OCR and MT data scores lower than the gold data. However, the model trained on OCR data shows notable improvements in the emotion-reg task, with an increase of 17.51.

4.2 Decoder Only Model

Table 6 outlines the performance of the GPT-2 model. The model scores the highest ORCA$_{score}$ in all datasets compared to all other model architectures, with the model trained on gold data scoring the best. However, the model struggles in tasks within the STS cluster. Consistent trends are observed across all tasks with the BERT$_{\text{Base}}$ model, with the model trained on OCR and ASR data showing significant performance improvements. The model trained on OCR data excels in hate-speech, scoring 72.47, and emotion-reg, with a notable increase to 61.59. Meanwhile, the model trained on ASR data achieves strong results in sarcasm (63.22) and dialect-binary (79.99). The TG dataset also show competitive performance, particularly in adult (86.63) and emotion (46.53) tasks. In contrast, the model trained on the MT dataset generally underperforms compared to the gold dataset but shows comparable results in irony (78.61) and offensive (78.60) tasks.

The results from the GPT-2 model suggest that this architecture is particularly adept at learning from limited data quantities. Overall, the results from GPT-2 are better than those from BERT$_{\text{Base}}$ and T5 model. This may be attributed to the fact that GPT-2 has the largest model size among those tested. Additionally, while most other model architectures benefit from training on ASR data, the GPT-2 model does not show the same improvements. This is likely because ASR data is the smallest among those used.

4.3 Encoder-Decoder Model

Table 7 presents the performance of the T5 model. The model trained on the ASR data achieves the highest ORCA$_{score}$, surpassing the gold dataset. The model trained on ASR data excels particularly in tasks such as abusive (64.77), emotion (45.97), and emotion-reg (11.12), demonstrating significant improvements. The model trained on the TG dataset also shows strong performance, especially in dialect-binary (84.22), offensive (77.98), and sarcasm (57.10) tasks, while maintaining competitive scores in most other tasks. The OCR-trained model has notable success in offensive (81.54) and sarcasm (67.79), although it underperforms in tasks like emotion-reg (2.52). The MT dataset generally enable lower performance than the gold data but is effective in enabling irony (73.07) and adult (88.56) tasks. Overall, the ASR and TG datasets support superior performance in the majority of tasks compared to the gold dataset.

5 Discussion

5.1 MT, OCR, ASR, and TG Data

The ORCA$_{score}$ from each model suggests that models trained on synthetic data from various sources can be as effective as those trained on gold data. Specifically, the BERT$_{\text{Base}}$ model trained on the ASR dataset outperformed all other datasets by an average of approximately 5 F_1 points. This can be attributed to the diverse domain sources included in the ASR data, unlike other datasets confined to a single domain. Furthermore, the T5 model trained on ASR data exhibits comparable performance to

Cluster	Task	Gold	MT	OCR	ASR	TG
SC	abusive	**67.86**$^{\pm 2.30}$	65.68$^{\pm 0.75}$	61.73$^{\pm 2.73}$	65.20$^{\pm 0.43}$	65.85$^{\pm 1.45}$
	adult	72.00$^{\pm 17.94}$	74.92$^{\pm 16.64}$	73.52$^{\pm 19.02}$	**87.20**$^{\pm 0.30}$	87.19$^{\pm 0.19}$
	age	33.49$^{\pm 11.47}$	41.73$^{\pm 0.40}$	33.50$^{\pm 1.41}$	41.40$^{\pm 1.11}$	**41.92**$^{\pm 0.54}$
	ans-claim	55.06$^{\pm 10.54}$	56.41$^{\pm 11.50}$	60.95$^{\pm 2.07}$	**62.49**$^{\pm 1.93}$	61.20$^{\pm 1.57}$
	dangerous	59.74$^{\pm 1.60}$	57.03$^{\pm 2.29}$	54.61$^{\pm 5.88}$	**62.84**$^{\pm 1.97}$	61.13$^{\pm 2.75}$
	dialect-binary	65.74$^{\pm 20.07}$	65.89$^{\pm 20.17}$	66.58$^{\pm 0.66}$	**80.76**$^{\pm 0.44}$	79.95$^{\pm 0.78}$
	dialect-country	16.93$^{\pm 10.40}$	17.03$^{\pm 0.98}$	22.61$^{\pm 0.87}$	23.72$^{\pm 0.27}$	**23.99**$^{\pm 0.25}$
	dialect-region	59.36$^{\pm 7.76}$	**60.89**$^{\pm 0.97}$	60.09$^{\pm 1.80}$	60.59$^{\pm 1.30}$	60.71$^{\pm 0.92}$
	emotion	47.10$^{\pm 0.28}$	**48.53**$^{\pm 0.46}$	46.34$^{\pm 0.35}$	47.00$^{\pm 0.09}$	46.29$^{\pm 0.94}$
	emotion-reg	14.49$^{\pm 13.91}$	14.98$^{\pm 11.78}$	**32.03**$^{\pm 0.66}$	31.85$^{\pm 0.54}$	23.71$^{\pm 7.06}$
	gender	**50.07**$^{\pm 10.91}$	41.94$^{\pm 10.31}$	35.72$^{\pm 1.48}$	34.86$^{\pm 0.29}$	42.89$^{\pm 11.65}$
	hate-speech	63.87$^{\pm 10.73}$	65.34$^{\pm 9.64}$	55.40$^{\pm 1.81}$	**72.19**$^{\pm 1.12}$	62.02$^{\pm 8.13}$
	irony	79.14$^{\pm 1.16}$	**79.71**$^{\pm 0.88}$	79.48$^{\pm 0.88}$	79.58$^{\pm 1.36}$	78.89$^{\pm 1.02}$
	offensive	78.56$^{\pm 0.63}$	79.08$^{\pm 0.63}$	**80.07**$^{\pm 0.45}$	79.61$^{\pm 0.60}$	78.84$^{\pm 0.62}$
	machine-generation	73.97$^{\pm 2.85}$	72.12$^{\pm 3.71}$	74.87$^{\pm 0.94}$	**75.89**$^{\pm 0.61}$	74.87$^{\pm 0.97}$
	sarcasm	59.58$^{\pm 9.96}$	59.57$^{\pm 6.81}$	61.78$^{\pm 1.96}$	65.34$^{\pm 1.70}$	**65.74**$^{\pm 2.01}$
	sentiment	**70.56**$^{\pm 0.43}$	69.64$^{\pm 0.46}$	69.36$^{\pm 0.46}$	69.76$^{\pm 0.46}$	69.64$^{\pm 0.46}$
SP	arabic-ner	63.81$^{\pm 0.54}$	57.33$^{\pm 0.72}$	57.94$^{\pm 0.79}$	64.40$^{\pm 0.30}$	**66.35**$^{\pm 0.49}$
	aqmar-ner	53.92$^{\pm 1.06}$	47.29$^{\pm 1.32}$	45.82$^{\pm 0.93}$	53.76$^{\pm 0.72}$	**57.68**$^{\pm 0.72}$
	msa-pos	14.71$^{\pm 1.23}$	12.40$^{\pm 0.71}$	14.38$^{\pm 1.74}$	**16.10**$^{\pm 0.59}$	13.18$^{\pm 0.59}$
	dialect-pos	81.67$^{\pm 0.32}$	81.38$^{\pm 0.18}$	81.18$^{\pm 0.37}$	**86.03**$^{\pm 0.71}$	84.20$^{\pm 0.13}$
NLI	ans-stance	37.87$^{\pm 8.62}$	42.25$^{\pm 4.29}$	43.07$^{\pm 2.89}$	**51.25**$^{\pm 2.97}$	47.91$^{\pm 2.92}$
	baly-stance	25.65$^{\pm 4.11}$	25.70$^{\pm 4.29}$	19.98$^{\pm 0.04}$	**29.29**$^{\pm 0.97}$	26.79$^{\pm 0.00}$
	xlni	36.33$^{\pm 13.92}$	16.70$^{\pm 0.04}$	35.93$^{\pm 1.69}$	**47.89**$^{\pm 1.10}$	36.88$^{\pm 14.29}$
STS	sts	11.08$^{\pm 6.61}$	15.43$^{\pm 2.06}$	17.54$^{\pm 4.23}$	**20.99**$^{\pm 4.48}$	14.28$^{\pm 4.79}$
	mq2q	55.09$^{\pm 0.50}$	**63.89**$^{\pm 15.97}$	52.80$^{\pm 0.29}$	52.92$^{\pm 0.29}$	53.26$^{\pm 0.43}$
TC	topic	**91.63**$^{\pm 6.35}$	90.86$^{\pm 6.44}$	89.82$^{\pm 0.99}$	91.36$^{\pm 0.50}$	90.95$^{\pm 0.38}$
QA	qa	23.19$^{\pm 16.36}$	13.57$^{\pm 4.29}$	14.42$^{\pm 0.99}$	23.41$^{\pm 0.50}$	**33.75**$^{\pm 5.37}$
WSD	wsd	42.04$^{\pm 6.27}$	36.24$^{\pm 4.05}$	48.01$^{\pm 6.68}$	**55.76**$^{\pm 12.58}$	33.30$^{\pm 0.19}$
***ORCA*$_{score}$**		51.88$^{\pm 0.43}$	50.81$^{\pm 1.35}$	51.36$^{\pm 0.50}$	**56.33**$^{\pm 0.97}$	54.60$^{\pm 0.38}$

Table 5: Performance of BERT$_{\text{Base}}$ model on Test splits (F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold** : Data obtained from Arabic Wikipedia, **MT** : Data obtained through Machine Translation, **OCR** : Data obtained through Optical Character Recognition, **ASR** : Data obtained through Automatic Speech Recognition, **TG** : Data obtained through Text Generation.

the gold data in tasks such as adult, dialect-binary, hate-speech, and sarcasm within the SC task cluster. This improvement is likely because, unlike the more formal and academically oriented texts from datasets like Wikipedia articles, ASR data are varied and include informal content, thus aiding in tasks such as predicting hate-speech and adult content. Additionally, other datasets contain predominantly Modern Standard Arabic (MSA) and lack sufficient exposure to different Dialectal Arabic (DA), impacting their effectiveness in DA classification tasks. Notably, the ASR dataset includes DA data, enhancing its performance in these tasks.

Models trained on MT data consistently score a lower ORCA$_{score}$ than any other dataset. This is likely due to the fact that, unlike datasets sourced from completely new domains, MT is generated by translating different subsets of English Wikipedia. Therefore, it is expected to perform worse than the original source. However, for the TG dataset, although it is also generated with Wikipedia as a base, it is more novel as it is produced from a language model (JASMINE) trained to perform well on both MSA and DA, giving it a slight edge over models solely trained on Wikipedia data or translated text.

We see a similar pattern to MT in the OCR dataset with larger declines in performance on tasks involving informal language. This can be attributed to the fact that OCR data is acquired from a repository of academic texts, making it very unlikely to have encountered foul language during training.

Ablation Study We conduct an ablation study to determine if aggregating synthetic data from different sources improves results. We choose specific data combinations based on these principles to assess their individual and collective impact on model performance. We only compare ARBERT$_{\text{v2}}$ against the GPT-2 model, as GPT-2 achieved the best results among the three model architectures. As shown in Table 8, GPT-2 exhibit continuous improvements with additional data, achieving the

Task	Gold	MT	OCR	ASR	TG
abusive	66.88	66.03	**69.14**	63.62	66.53
adult	86.68	86.69	86.68	**86.76**	86.63
age	**43.08**	41.94	41.91	41.99	41.83
ans-claim	**64.68**	62.83	64.09	63.55	62.73
dangerous	60.10	59.54	59.88	57.93	**61.12**
dialect-binary	79.94	79.71	79.82	79.99	79.74
dialect-country	**24.87**	23.13	24.33	23.09	22.92
dialect-region	**60.05**	59.78	59.93	59.76	59.96
emotion	**49.73**	44.06	46.09	43.80	46.53
emotion-reg	**44.27**	32.94	41.59	33.92	34.16
gender	**62.03**	61.54	61.77	61.59	60.57
hate-speech	69.60	69.07	**72.47**	68.75	69.42
irony	80.22	78.68	**81.47**	78.52	79.31
offensive	80.09	78.97	**80.60**	78.82	79.33
machine-generation	**76.04**	75.05	75.32	76.00	75.51
sarcasm	**65.59**	64.65	65.52	65.41	63.22
sentiment	69.75	69.67	69.84	69.70	**69.98**
arabic-ner	**75.35**	69.88	71.53	66.29	71.20
aqmar-ner	**66.91**	62.57	63.08	55.87	64.78
msa-pos	**27.38**	17.95	12.40	19.52	22.27
dialect-pos	82.33	82.03	**82.41**	82.04	82.27
ans-stance	44.82	41.49	**45.15**	38.81	38.70
baly-stance	**29.12**	25.75	26.00	25.70	26.69
xlni	**52.74**	47.93	48.45	45.58	50.69
sts	5.20	6.61	13.29	**18.37**	14.07
mq2q	**58.15**	52.88	55.51	52.04	52.02
topic	91.68	91.37	**91.70**	90.89	91.34
qa	**37.20**	33.24	34.71	21.39	36.20
wsd	68.01	66.84	66.71	67.44	**68.09**
$ORCA_{score}$	**59.40**	57.40	58.32	56.45	57.86

Table 6: Performance of GPT2 model on Test splits (F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold** : Data obtained from Arabic Wikipedia, **MT** : Data obtained through Machine Translation, **OCR** : Data obtained through Optical Character Recognition, **ASR** : Data obtained through Automatic Speech Recognition, **TG** : Data obtained through Text Generation.

highest $ORCA_{score}$ of 62.58 on the C4 dataset. This improvement can be attributed to GPT-2's larger model size, which benefits more from increased training data. Notably, the GPT-2 model consistently outperform $ARBERT_{v2}$ across most tasks, particularly in the SC and SP clusters, demonstrating its superior ability to handle a diverse range of tasks when trained on combined datasets.

5.2 Comparing Against Baseline

We compare the results from our trained models against our baseline model, $ARBERT_{v2}$. Despite

Task	Gold	MT	OCR	ASR	TG
abusive	48.85	48.76	25.48	**64.77**	49.60
adult	88.52	**88.56**	88.06	83.19	87.84
age	43.22	44.59	29.10	**45.97**	45.01
ans-claim	**61.92**	61.53	40.15	61.01	61.62
dangerous	**65.91**	65.60	46.75	63.50	60.61
dialect-binary	82.41	82.47	**84.61**	78.69	84.22
dialect-country	12.01	6.60	3.58	**22.63**	14.53
dialect-region	58.99	58.41	61.77	**61.84**	59.98
emotion	27.08	29.49	24.36	**45.97**	29.70
emotion-reg	3.18	5.17	2.52	**11.12**	7.02
gender	**62.87**	62.10	60.16	51.91	62.84
hate-speech	64.14	56.05	48.70	**71.13**	48.70
irony	76.05	73.07	78.07	**79.29**	77.00
offensive	78.15	75.59	81.54	**84.08**	77.98
machine-generation	76.41	75.77	78.84	**79.43**	77.25
sarcasm	54.05	55.45	45.55	**67.79**	57.10
sentiment	69.00	70.63	71.84	**73.07**	70.91
ans-stance	27.23	33.77	25.98	**35.97**	27.13
baly-stance	24.38	**29.06**	19.9	26.24	26.12
xlni	32.67	30.17	**57.32**	48.99	46.93
sts	5.5	5.10	**16.34**	7.11	12.21
mq2q	85.41	84.10	**89.77**	78.46	88.08
topic	87.53	88.68	40.21	**91.67**	90.47
qa	**40.74**	36.46	36.83	21.36	36.55
wsd	64.63	64.99	65.67	63.69	**65.85**
$ORCA_{score}$	45.84	45.57	42.55	**49.98**	47.27

Table 7: Performance of T5 model on Test splits (F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold** : Data obtained from Arabic Wikipedia, **MT** : Data obtained through Machine Translation, **OCR** : Data obtained through Optical Character Recognition, **ASR** : Data obtained through Automatic Speech Recognition, **TG** : Data obtained through Text Generation. Token Level Tasks were excluded as they yielded scores of 0 F_1 score.

$ARBERT_{v2}$ being trained on a vast and diverse dataset specifically designed for Arabic, our models achieve comparable results in tasks such as dialect, age, adult, dangerous, and topic classification. Additionally, the Spearman Correlation between the highest $ORCA_{score}$ from $ARBERT_{v2}$ and each model architecture shows significant positive correlations, indicating a strong relationship between the two sets of scores. Specifically, the decoder-only architecture exhibit a high correlation of 0.85, suggesting a close alignment with the $ARBERT_{v2}$ model. The encoder-only and encoder-decoder models also show substantial correlations, with coefficients of 0.728 and 0.725, respectively. These results demonstrate not only the competitive

Task	c1	c2	c3	c4	BL
abusive	69.06	68.98	69.14	67.27	75.99
adult	86.48	87.00	86.76	87.20	89.67
age	42.68	42.60	43.08	42.79	45.57
ans-claim	63.23	61.56	64.68	63.08	67.38
dangerous	58.43	60.94	61.12	59.44	64.96
dialect-binary	80.00	79.72	79.99	79.91	86.92
dialect-country	24.76	24.19	24.87	23.06	35.69
dialect-region	60.61	59.89	60.05	59.37	65.21
emotion	46.15	49.46	49.73	45.43	64.81
emotion-reg	43.39	40.28	44.27	42.10	67.73
gender	61.94	61.39	62.03	62.12	63.18
hate-speech	70.99	73.43	72.47	70.59	82.26
irony	80.40	80.29	81.47	81.46	83.83
offensive	79.54	81.08	80.60	80.47	89.55
machine-generation	76.52	75.80	76.04	76.79	87.94
sarcasm	64.15	63.53	65.59	65.82	74.16
sentiment	70.18	69.88	69.98	70.15	78.60
arabic-ner	73.58	74.30	75.35	72.64	90.83
aqmar-ner	66.36	66.11	66.91	65.88	81.70
msa-pos	26.19	29.63	27.38	21.23	52.55
dialect-pos	82.49	82.98	83.54	83.09	93.92
ans-stance	43.14	45.58	45.15	42.14	91.02
baly-stance	28.39	32.24	29.12	27.16	49.34
xlni	50.99	52.82	52.74	51.34	68.17
sts	7.97	22.94	36.23	11.99	71.90
mq2q	55.77	58.37	91.80	54.47	96.73
topic	91.70	91.63	91.70	91.64	93.96
qa	37.20	39.42	41.30	35.08	61.56
wsd	67.97	67.96	68.32	67.63	71.01
$ORCA_{score}$	58.65	58.99	60.16	62.58	74.02

Table 8: Performance of GPT2 model on Test splits (F_1) and ARBERTv2. Metric for the sts and emotion-reg tasks is spearman correlation. **C1** : Combined dataset 1, **C2** : Combined dataset 2, **C3** : Combined dataset 3, **C4** : Combined dataset 4, **BL** : Baseline Model $ARBERT_{v2}$.

performance of our models against a fully trained PLMs but also highlight the robustness of training on various sources across different architectures.

6 Conclusion

In this work, we highlight the importance of leveraging different data extraction techniques for the augmentation and development of LMs. Our results show noticeable benefits from leveraging different sources of text to augment the performance of LMs. Our findings indicate that models trained on diverse synthetic data sources can achieve performance comparable to those trained on gold data. Specifically, the diverse domain sources in ASR data and the informal content included in such datasets significantly enhance model capabilities in various tasks. While MT data enable lower performance due to its translation-based generation, TG data demonstrate improved results due to its novelty and the robust language model used in its creation. Our work, emphasizes the value of incorporating a variety of data sources to improve the effectiveness of language models in handling different linguistic tasks.

7 Limitations

We identify the following limitations in this work:

1. Only the ASR dataset contains dialectal Arabic (DA) data, whereas all other datasets are limited to Modern Standard Arabic (MSA). This inclusion of DA data in the ASR dataset boosts the performance of models trained on ASR data, particularly for tasks related to DA. Furthermore, the varying number of datasets from each source makes it challenging to ensure a fair comparison of the efficacy of these datasets.

2. The model sizes differ significantly, which may affect performance and does not ensure a fair comparison. Larger models, such as GPT-2, benefit more from increased training data, potentially biasing the results. This variability in model size complicates direct performance comparisons across different architectures.

3. The scope of this study is primarily centered around NLU tasks. While our findings provide valuable insights into the performance of various datasets and model architectures for NLU, they may not generalize to other types of natural language processing tasks, such as NLG tasks.

4. The datasets used in this study may carry inherent biases specific to their domains of origin. For example, the OCR dataset, sourced from academic texts, may not adequately represent informal language use, affecting model performance on tasks involving colloquial or slang expressions. This domain-specific bias can limit the applicability of our findings to more diverse linguistic contexts.

8 Ethics Statement and Broad Impact

Promoting inclusive NLP research and resource development. NLP for Arabic languages has been under-resourced compared to other languages. This scarcity of resources has hindered the development of robust NLP applications for Arabic. Our work aims to bridge this gap by leveraging diverse datasets to improve model performance across various tasks. We hope to stimulate further research and development in Arabic NLP, fostering innovation and enabling the creation of more effective and culturally aware language technologies.

Encouraging future research and innovation. Our research underscores the potential of using diverse data sources to enhance the performance of language models. We hope that our findings will inspire further exploration into the integration of underrepresented language varieties in NLP research. This pursuit can lead to the creation of more robust, adaptable, and inclusive language models. By advancing the state of NLP for Arabic and other less-represented languages, we contribute to the broader goal of developing AI technologies that serve a global audience, promoting cross-cultural understanding and communication.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[2] and UBC ARC-Sockeye.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020a. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020b. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2021. Masc: Massive arabic speech corpus.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. 2023. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*.

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.

Elfilali Ali. 2023. Hindawi books dataset. Dataset.

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. 101 billion arabic words dataset. *arXiv preprint arXiv:2405.01590*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for arabic. *arXiv preprint arXiv:2109.08068*.

Arne Bewersdorff, Christian Hartmann, Marie Hornberger, Kathrin Seßler, Maria Bannert, Enkelejda Kasneci, Gjergji Kasneci, Xiaoming Zhai, and Claudia Nerdel. 2024. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *arXiv preprint arXiv:2401.00832*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot

[2] https://alliancecan.ca

learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Fahim Dalvi, Yifan Zhang, Sameer Khurana, Nadir Durrani, Hassan Sajjad, Ahmed Abdelali, Hamdy Mubarak, Ahmed Ali, and Stephan Vogel. 2017. QCRI live speech translation system. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 61–64, Valencia, Spain. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks. *arXiv preprint arXiv:2011.01549*.

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

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Nawar Halabi. 2016. *Modern standard Arabic phonetics for speech synthesis*. Ph.D. thesis, University of Southampton.

Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. 2023. Mixgen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 379–389.

Alex Hernández-García and Peter König. 2018. Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*.

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*.

Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond English: Evaluating LLMs for Arabic grammatical error correction. In *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid). Association for Computational Linguistics.

Zichang Liu, Zhiqiang Tang, Xingjian Shi, Aston Zhang, Mu Li, Anshumali Shrivastava, and Andrew Gordon Wilson. 2022. Learning multimodal data augmentation in feature space. *arXiv preprint arXiv:2212.14453*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Alcoba Inciarte, and Md Tawkat Islam Khondaker. 2022. Jasmine: Arabic gpt models for few-shot learning. *arXiv preprint arXiv:2212.10755*.

El Moatez Billah Nagoudi, Ahmed El-Shangiti, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for arabic nlg. *arXiv preprint arXiv:2305.14989*.

Rasha Obeidat, Marwa Al-Harbi, Mahmoud Al-Ayyoub, and Luay Alawneh. 2024. Arquad: An expert-annotated arabic machine reading comprehension dataset. *Cognitive Computation*, pages 1–20.

Luis Perez and Jason Wang. 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Dania Refai, Saleh Abu-Soud, and Mohammad J Abdel-Rahman. 2023. Data augmentation using transformers and similarity measures for improving arabic text classification. *IEEE Access*, 11:132516–132531.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open

large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60.

R. Smith. 2007. An overview of the tesseract ocr engine. volume 2, pages 629 – 633.

Aiman Solyman, Wang Zhenyu, Tao Qian, Arafat Abdulgader Mohammed Elhag, Muhammad Toseef, and Zeinab Aleibeid. 2021. Synthetic data with neural machine translation for automatic correction in arabic grammar. *Egyptian Informatics Journal*, 22(3):303–315.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-shot benchmarking of whisper on diverse arabic speech recognition. *arXiv preprint arXiv:2306.02902*.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. Promda: Prompt-based data augmentation for low-resource nlu tasks. *arXiv preprint arXiv:2202.12499*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Suorong Yang, Weikang Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Furao Shen. 2022. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Terry Yue Zhuo, Zhou Yang, Zhensu Sun, Yufei Wang, Li Li, Xiaoning Du, Zhenchang Xing, and David Lo. 2023. Data augmentation approaches for source code models: A survey. *arXiv preprint arXiv:2305.19915*.

A Appendices

A.1 Jasmine Sample

The prompt and generated examples from Jasmine can be found in Table 9.

A.2 Test Results with Standard Deviation

Results of the Test sets differences in respect to the gold data for each model architecture and standard deviation can be found in Table 10, 11.

A.3 Test Results as Compared to Gold

Results of the Test sets with positive and negative differences in respect to the gold data for each model architecture can be found in Table 12, 13, 14.

A.4 Dev Results

Results of the Dev sets for each model architecture can be found in Table 15, 16, 17.

Jasmine Generated Text Sample 1
لماء مادةٌ شفافةٌ عديمة اللون والرائحة، وهو المكوّن الأساسي للجداول والبحيرات والبحار والمحيطات وكذلك للسوائل في جميع الكائنات الحيّة، وهو أكثر المركّبات الكيميائيّة انتشاراً على سطح الأرض. يتألّف جزيء الماء من ذرّة أكسجين مركزية ترتبط بها ذرّتا هيدروجين على طرفيها . عند الظروف القياسية من الضغط ودرجة الحرارة يكون الماء سائلاً؛ أمّا H2O برابطة تساهميّة بحيث تكون صيغته الكيميائية الحالة الصلبة فتتشكّل عند نقطة التجمّد، وتدعى بالجليد؛ أمّا الحالة الغازية فتتشكّل عند نقطة الغليان، وتسمّى بخار الماء. ل الديناميكي: هو عبارة عن مجموعة الذرات المرتبطة بروابط هيدروجينية أو ربانية (الرباط الهيدروجيني) والتي تشكل معا جزيئات السائل الذي يسمى بالمائع... أما إذا كانت الرابطة بين ذرة الهيدروجين وذرة الأكسجين رابطة تشاركية فيكون هذا النوع سائل متبلور مثل المياه السائلة والمخلوطة مع بعض المواد الأخرى ، وأما إن لم يكن هناك أي رابط مشترك فتكون هذه السوائل غير المتبلورة هي ما تسمى بالسائل اللزج او المائي وتكون سائلة فقط عندما تصل درجة حرارتها إلى ١٠٠ م حيث تتصلب وتصبح لينة ومطاطة وتسمى حينها بالمياه الباردة.. ومن الأمثلة عليها مياه البحر ومياه الينابيع العذبة التي تحتوي علي نسبة عالية جدا»من الأملاح المعدنية المذابة فيها نتيجة التبخر والتعرق الشديدين،كما أن معظم سوائل البحار لا طعم لها ولا لون وهي ليست ماء بل أملاح معدنية مذابه فيه بسبب تبخره عبر مسام الصخور الموجودة تحت السطح وبسبب وجود الملح بشكل كبير أيضا» . . .
Sample 2
يمكن لشدة اللون الطيفي أن تغير الإحساس به إلى حد بعيد، فمثلا، اللون البرتقالي ـ الأصفر ذو الشدة المنخفضة يبدو بنيا،كما يبدو اللون الأصفر ـ الأخضر ذو الشدة المنخفضة أخضرا زيتونيا. الديناميكي هو لون آخر من الألوان الأساسية التي يمكن رؤيتها في الضوء المرئي فقط؛ إذ إن شدة هذا الطيف تكون أقل بكثير عند النظر إليها تحت ضوء الشمس المباشر أو الأشعة فوق البنفسجية. على سبيل المثال: يكون طيف الأزرق أكثر حدة عندما ينظر إليه مع خلفية زرقاء داكنة جدا ذات إضاءة منخفضة للغاية مقارنة بإضاءة أشعة شمس الظهيرة العادية. أما بالنسبة للون الأحمر فهو أيضا غير مرئي للعين البشرية ولكن له تأثير كبير عليها حيث أنه يجعل الناس يشعرون بالدفء والحرارة ، وهو ما يعرف بظاهرة»تأثير البيت المحمي». وقد تم استخدام مصطلح »الأحمر العميق»، الذي يعني العمق واللون الأكثر عمقا للأحمر والذي لا يصله البصر عادة إلا بعد فترة طويلة نسبيا قد تصل لعدة أيام أحيانا لوصف هذه الظاهرة بشكل دقيق وواضح وذلك بسبب طول الفترة الزمنية بين التعرض للضوء وظهور التأثير الفعلي الناتج . . .

Table 9: Examples of generated wikipedia text. We color the initial prompt with blue.

Task	Gold	MT	OCR	ASR	TG
abusive	66.88$^{\pm 1.52}$	66.03$^{\pm 1.51}$	**69.14**$^{\pm 1.95}$	63.62$^{\pm 2.22}$	66.53$^{\pm 2.92}$
adult	86.68$^{\pm 0.91}$	86.69$^{\pm 0.23}$	86.68$^{\pm 0.23}$	**86.76**$^{\pm 0.19}$	86.63$^{\pm 0.15}$
age	**43.08**$^{\pm 0.84}$	41.94$^{\pm 0.45}$	41.91$^{\pm 0.56}$	41.99$^{\pm 0.71}$	41.83$^{\pm 0.71}$
ans-claim	**64.68**$^{\pm 0.60}$	62.83$^{\pm 1.71}$	64.09$^{\pm 0.59}$	63.55$^{\pm 0.76}$	62.73$^{\pm 0.44}$
dangerous	60.10$^{\pm 1.56}$	59.54$^{\pm 1.45}$	59.88$^{\pm 3.35}$	57.93$^{\pm 2.65}$	**61.12**$^{\pm 3.42}$
dialect-binary	79.94$^{\pm 0.69}$	79.71$^{\pm 0.29}$	79.82$^{\pm 0.78}$	**79.99**$^{\pm 0.27}$	79.74$^{\pm 0.47}$
dialect-country	**24.87**$^{\pm 1.40}$	23.13$^{\pm 0.97}$	24.33$^{\pm 0.73}$	23.09$^{\pm 0.56}$	22.92$^{\pm 0.30}$
dialect-region	**60.05**$^{\pm 0.56}$	59.78$^{\pm 0.18}$	59.93$^{\pm 0.29}$	59.76$^{\pm 0.29}$	59.96$^{\pm 0.07}$
emotion	**49.73**$^{\pm 1.93}$	44.06$^{\pm 0.26}$	46.09$^{\pm 2.37}$	43.80$^{\pm 0.59}$	46.53$^{\pm 1.85}$
emotion-reg	**44.27**$^{\pm 0.47}$	32.94$^{\pm 1.99}$	41.59$^{\pm 1.28}$	33.92$^{\pm 3.24}$	34.16$^{\pm 1.33}$
gender	**62.03**$^{\pm 4.04}$	61.54$^{\pm 0.36}$	61.77$^{\pm 0.58}$	61.59$^{\pm 0.35}$	60.57$^{\pm 0.86}$
hate-speech	69.60$^{\pm 1.64}$	69.07$^{\pm 0.87}$	**72.47**$^{\pm 1.94}$	68.75$^{\pm 5.32}$	69.42$^{\pm 1.59}$
irony	80.22$^{\pm 0.73}$	78.68$^{\pm 0.56}$	**81.47**$^{\pm 0.48}$	78.52$^{\pm 0.97}$	79.31$^{\pm 0.98}$
offensive	80.09$^{\pm 0.95}$	78.97$^{\pm 0.37}$	**80.60**$^{\pm 0.46}$	78.82$^{\pm 0.46}$	79.33$^{\pm 0.49}$
machine-generation	**76.04**$^{\pm 0.35}$	75.05$^{\pm 0.38}$	75.32$^{\pm 0.39}$	76.00$^{\pm 0.55}$	75.51$^{\pm 1.06}$
sarcasm	**65.59**$^{\pm 2.07}$	64.65$^{\pm 1.99}$	65.52$^{\pm 2.08}$	65.41$^{\pm 1.91}$	63.22$^{\pm 1.05}$
sentiment	69.75$^{\pm 0.58}$	69.67$^{\pm 0.23}$	69.84$^{\pm 0.83}$	69.70$^{\pm 0.34}$	**69.98**$^{\pm 0.26}$
arabic-ner	**75.35**$^{\pm 0.64}$	69.88$^{\pm 0.31}$	71.53$^{\pm 0.56}$	66.29$^{\pm 0.44}$	71.20$^{\pm 0.44}$
aqmar-ner	**66.91**$^{\pm 1.35}$	62.57$^{\pm 0.56}$	63.08$^{\pm 2.01}$	55.87$^{\pm 0.72}$	64.78$^{\pm 0.72}$
msa-pos	**27.38**$^{\pm 2.45}$	17.95$^{\pm 1.27}$	12.40$^{\pm 0.77}$	19.52$^{\pm 2.13}$	22.27$^{\pm 2.42}$
dialect-pos	82.33$^{\pm 0.32}$	82.03$^{\pm 0.20}$	**82.41**$^{\pm 0.38}$	82.04$^{\pm 0.41}$	82.27$^{\pm 0.52}$
ans-stance	44.82$^{\pm 8.62}$	41.49$^{\pm 1.83}$	**45.15**$^{\pm 2.39}$	38.81$^{\pm 0.40}$	38.70$^{\pm 2.32}$
baly-stance	**29.12**$^{\pm 0.23}$	25.75$^{\pm 2.21}$	26.00$^{\pm 2.30}$	25.70$^{\pm 0.91}$	26.69$^{\pm 1.19}$
xlni	**52.74**$^{\pm 2.69}$	47.93$^{\pm 1.72}$	48.45$^{\pm 5.36}$	45.58$^{\pm 1.39}$	50.69$^{\pm 1.39}$
sts	5.20$^{\pm 2.26}$	6.61$^{\pm 3.64}$	13.29$^{\pm 4.29}$	**18.37**$^{\pm 6.78}$	14.07$^{\pm 8.10}$
mq2q	**58.15**$^{\pm 2.95}$	52.88$^{\pm 2.37}$	55.51$^{\pm 2.16}$	52.04$^{\pm 2.40}$	52.02$^{\pm 2.52}$
topic	91.68$^{\pm 0.31}$	91.37$^{\pm 0.21}$	**91.70**$^{\pm 0.24}$	90.89$^{\pm 0.69}$	91.34$^{\pm 0.26}$
qa	**37.20**$^{\pm 0.99}$	33.24$^{\pm 0.38}$	34.71$^{\pm 0.34}$	21.39$^{\pm 0.62}$	36.20$^{\pm 0.62}$
wsd	68.01$^{\pm 0.18}$	66.84$^{\pm 0.59}$	66.71$^{\pm 1.97}$	67.44$^{\pm 0.19}$	**68.09**$^{\pm 0.25}$
ORCA$_{score}$	**59.40**$^{\pm 0.32}$	57.40$^{\pm 0.40}$	58.32$^{\pm 0.25}$	56.45$^{\pm 0.25}$	57.86$^{\pm 0.25}$

Table 10: Performance of GPT2 model on Test splits (F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold** : Data obtained from Arabic Wikipedia, **MT** : Data obtained through Machine Translation, **OCR** : Data obtained through Optical Character Recognition, **ASR** : Data obtained through Automatic Speech Recognition, **TG** : Data obtained through Text Generation.

Task	Gold	MT	OCR	ASR	TG
abusive	48.85$^{\pm 0.20}$	48.76$^{\pm 0.53}$	25.48$^{\pm 1.95}$	**64.77**$^{\pm 0.39}$	49.60$^{\pm 0.40}$
adult	88.52$^{\pm 0.36}$	**88.56**$^{\pm 0.16}$	88.06$^{\pm 0.76}$	83.19$^{\pm 0.19}$	87.84$^{\pm 0.26}$
age	43.22$^{\pm 0.10}$	44.59$^{\pm 0.44}$	29.10$^{\pm 0.76}$	**45.97**$^{\pm 0.50}$	45.01$^{\pm 0.35}$
ans-claim	**61.92**$^{\pm 0.87}$	61.53$^{\pm 0.16}$	40.15$^{\pm 2.66}$	61.01$^{\pm 1.47}$	61.62$^{\pm 2.66}$
dangerous	**65.91**$^{\pm 0.85}$	65.60$^{\pm 0.88}$	46.75$^{\pm 1.87}$	63.50$^{\pm 1.87}$	60.61$^{\pm 1.87}$
dialect-binary	82.41$^{\pm 0.17}$	82.47$^{\pm 0.29}$	**84.61**$^{\pm 0.11}$	78.69$^{\pm 0.40}$	84.22$^{\pm 0.39}$
dialect-country	12.01$^{\pm 0.21}$	6.60$^{\pm 0.12}$	3.58$^{\pm 0.45}$	**22.63**$^{\pm 0.19}$	14.53$^{\pm 0.33}$
dialect-region	58.99$^{\pm 0.15}$	58.41$^{\pm 0.13}$	61.77$^{\pm 0.09}$	**61.84**$^{\pm 0.04}$	59.98$^{\pm 0.17}$
emotion	27.08$^{\pm 0.58}$	29.49$^{\pm 0.25}$	24.36$^{\pm 0.36}$	**45.97**$^{\pm 0.23}$	29.70$^{\pm 0.36}$
emotion-reg	3.18$^{\pm 1.18}$	5.17$^{\pm 2.30}$	2.52$^{\pm 1.64}$	**11.12**$^{\pm 1.78}$	7.02$^{\pm 1.78}$
gender	**62.87**$^{\pm 0.39}$	62.10$^{\pm 0.23}$	60.16$^{\pm 0.31}$	51.91$^{\pm 0.12}$	62.84$^{\pm 0.86}$
hate-speech	64.14$^{\pm 0.89}$	56.05$^{\pm 1.34}$	48.70$^{\pm 1.31}$	**71.13**$^{\pm 1.15}$	48.70$^{\pm 1.26}$
irony	76.05$^{\pm 0.81}$	73.07$^{\pm 0.50}$	78.07$^{\pm 0.62}$	**79.29**$^{\pm 0.27}$	77.00$^{\pm 0.27}$
offensive	78.15$^{\pm 0.94}$	75.59$^{\pm 1.57}$	81.54$^{\pm 0.63}$	**84.08**$^{\pm 0.47}$	77.98$^{\pm 0.47}$
machine-generation	76.41$^{\pm 0.18}$	75.77$^{\pm 0.26}$	78.84$^{\pm 0.77}$	**79.43**$^{\pm 0.21}$	77.25$^{\pm 0.21}$
sarcasm	54.05$^{\pm 2.00}$	55.45$^{\pm 1.67}$	45.55$^{\pm 1.66}$	**67.79**$^{\pm 1.27}$	57.10$^{\pm 1.28}$
sentiment	69.00$^{\pm 0.19}$	70.63$^{\pm 0.25}$	71.84$^{\pm 0.34}$	**73.07**$^{\pm 0.03}$	70.91$^{\pm 0.23}$
ans-stance	27.23$^{\pm 1.40}$	33.77$^{\pm 3.96}$	25.98$^{\pm 1.83}$	**35.97**$^{\pm 0.85}$	27.13$^{\pm 2.02}$
baly-stance	24.38$^{\pm 0.27}$	**29.06**$^{\pm 0.98}$	19.98$^{\pm 0.31}$	26.24$^{\pm 0.71}$	26.12$^{\pm 1.49}$
xlni	32.67$^{\pm 0.73}$	30.17$^{\pm 0.66}$	**57.32**$^{\pm 0.31}$	48.99$^{\pm 0.76}$	46.93$^{\pm 0.69}$
sts	5.5$^{\pm 2.89}$	5.10$^{\pm 3.64}$	**16.34**$^{\pm 4.27}$	7.11$^{\pm 0.58}$	12.21$^{\pm 5.10}$
mq2q	85.41$^{\pm 0.31}$	84.10$^{\pm 0.35}$	**89.77**$^{\pm 0.19}$	78.46$^{\pm 0.24}$	88.08$^{\pm 0.45}$
topic	87.53$^{\pm 2.11}$	88.68$^{\pm 1.17}$	40.21$^{\pm 0.33}$	**91.67**$^{\pm 1.58}$	90.47$^{\pm 2.71}$
qa	**40.74**$^{\pm 0.84}$	36.46$^{\pm 0.18}$	36.83$^{\pm 0.49}$	21.36$^{\pm 0.17}$	36.55$^{\pm 0.72}$
wsd	64.63$^{\pm 0.48}$	64.99$^{\pm 0.15}$	65.67$^{\pm 0.23}$	63.69$^{\pm 0.57}$	**65.85**$^{\pm 0.06}$
ORCA$_{score}$	45.84$^{\pm 2.11}$	45.57$^{\pm 1.17}$	42.55$^{\pm 0.33}$	**49.98**$^{\pm 1.58}$	47.27$^{\pm 2.71}$

Table 11: Performance of T5 model on Test splits (F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold** : Data obtained from Arabic Wikipedia, **MT** : Data obtained through Machine Translation, **OCR** : Data obtained through Optical Character Recognition, **ASR** : Data obtained through Automatic Speech Recognition, **TG** : Data obtained through Text Generation. Token Level Tasks were excluded as they yielded scores of 0 F_1 score.

Cluster	Task	Gold	MT	OCR	ASR	TG
SC	abusive	67.86	−2.18	−6.13	−2.66	−2.01
	adult	72.00	+2.92	+1.52	+15.20	+15.19
	age	33.49	+8.24	+0.01	+7.91	+8.43
	ans-claim	55.06	+1.35	+5.89	+7.43	+6.14
	dangerous	59.74	−2.71	−5.13	+3.10	+1.39
	dialect-binary	65.74	+0.15	+0.84	+15.02	+14.21
	dialect-country	16.93	+0.10	+5.68	+6.79	+7.06
	dialect-region	59.36	+1.53	+0.73	+1.23	+1.35
	emotion	47.10	+1.43	−0.76	−0.10	−0.81
	emotion-reg	14.49	+0.49	+17.54	+17.36	+9.22
	gender	50.07	−8.13	−14.35	−15.21	−7.18
	hate-speech	63.87	+1.47	−8.47	+8.32	−1.85
	irony	79.14	+0.57	+0.34	+0.44	−0.25
	offensive	78.56	+0.52	+1.51	+1.05	+0.28
	machine-generation	73.97	−1.85	+0.90	+1.92	+0.90
	sarcasm	59.58	−0.01	+2.20	+5.76	+6.16
	sentiment	70.56	−0.92	−1.20	−0.80	−0.92
SP	arabic-ner	63.81	−6.48	−5.87	+0.59	+2.54
	aqmar-ner	53.92	−6.63	−8.10	−0.16	+3.76
	msa-pos	14.71	−2.31	−0.33	+1.39	−1.53
	dialect-pos	81.67	−0.29	−0.49	+4.36	+2.53
NLI	ans-stance	37.87	+4.38	+5.20	+13.38	+10.04
	baly-stance	25.65	+0.05	−5.67	+3.64	+1.14
	xlni	36.33	−19.63	−0.40	+11.56	+0.55
STS	sts	11.08	+4.35	+6.46	+9.91	+3.20
	mq2q	55.09	+8.80	−2.29	−2.17	−1.83
TC	topic	91.63	−0.77	−1.81	−0.27	−0.68
QA	qa	23.19	−9.62	−8.77	+0.22	+10.56
WSD	wsd	42.04	−5.80	+5.97	+13.72	−8.74
ORCA$_{score}$		51.88	−1.07	−0.52	+4.45	+2.72

Table 12: Difference in performance of $\text{BERT}_{\text{Base}}$ model on Test set compared to Gold(F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold** : Data obtained from Arabic Wikipedia, **MT** : Data obtained through Machine Translation, **OCR** : Data obtained through Optical Character Recognition, **ASR** : Data obtained through Automatic Speech Recognition, **TG** : Data obtained through Text Generation.

Task	Gold	MT	OCR	ASR	TG
abusive	66.88	−0.85	+2.26	−3.26	−0.35
adult	86.68	+0.01	+0.00	+0.08	−0.05
age	43.08	−1.14	−1.17	−1.09	−1.25
ans-claim	64.68	−1.85	−0.59	−1.13	−1.95
dangerous	60.10	−0.56	−0.22	−2.17	+1.02
dialect-binary	79.94	−0.23	−0.12	+0.05	−0.20
dialect-country	24.87	−1.74	−0.54	−1.78	−1.95
dialect-region	60.05	−0.27	−0.12	−0.29	−0.09
emotion	49.73	−5.67	−3.64	−5.93	−3.20
emotion-reg	44.27	−11.33	−2.68	−10.35	−10.11
gender	62.03	−0.49	−0.26	−0.44	−1.46
hate-speech	69.60	−0.53	+2.87	−0.85	−0.18
irony	80.22	−1.54	+1.25	−1.70	−0.91
offensive	80.09	−1.12	+0.51	−1.27	−0.76
machine-generation	76.04	−0.99	−0.72	−0.04	−0.53
sarcasm	65.59	−0.94	−0.07	−0.18	−2.37
sentiment	69.75	−0.08	+0.09	−0.05	+0.23
arabic-ner	75.35	−5.47	−3.82	−9.06	−4.15
aqmar-ner	66.91	−4.34	−3.83	−11.04	−2.13
msa-pos	27.38	−9.43	−14.98	−7.86	−5.11
dialect-pos	82.33	−0.30	+0.08	−0.29	−0.06
ans-stance	44.82	−3.33	+0.33	−6.01	−6.12
baly-stance	29.12	−3.37	−3.12	−3.42	−2.43
xlni	52.74	−4.81	−4.29	−7.16	−2.05
sts	5.20	+1.41	+8.09	+13.17	+8.87
mq2q	58.15	−5.27	−2.64	−6.11	−6.13
topic	91.68	−0.31	+0.02	−0.79	−0.34
qa	37.20	−3.96	−2.49	−15.81	−1.00
wsd	68.01	−1.17	−1.30	−0.57	+0.08
ORCA$_{score}$	59.40	−2.00	−1.08	−2.95	−1.54

Table 13: Difference in performance of GPT2 model on Test set compared to Gold(F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold**: Data obtained from Arabic Wikipedia, **MT**: Data obtained through Machine Translation, **OCR**: Data obtained through Optical Character Recognition, **ASR**: Data obtained through Automatic Speech Recognition, **TG**: Data obtained through Text Generation.

Task	Gold	MT	OCR	ASR	TG
abusive	48.85	−0.09	−23.37	+15.92	+0.75
adult	88.52	+0.04	−0.46	−5.33	−0.68
age	43.22	+1.37	−14.12	+2.75	+1.79
ans-claim	61.92	−0.39	−21.77	−0.91	−0.30
dangerous	65.91	−0.31	−19.16	−2.41	−5.30
dialect-binary	82.41	+0.06	+2.20	−3.72	+1.81
dialect-country	12.01	−5.41	−8.43	+10.62	+2.52
dialect-region	58.99	−0.58	+2.78	+2.85	+0.99
emotion	27.08	+2.41	−2.72	+18.89	+2.62
emotion-reg	3.18	+1.99	−0.66	+7.94	+3.84
gender	62.87	−0.77	−2.71	−10.96	−0.03
hate-speech	64.14	−8.09	−15.44	+7.00	−15.44
irony	76.05	−2.98	+2.02	+3.24	+0.95
offensive	78.15	−2.56	+3.39	+5.93	−0.17
machine-generation	76.41	−0.64	+2.43	+3.02	+0.84
sarcasm	54.05	+1.40	−8.50	+13.74	+3.05
sentiment	69.00	+1.63	+2.84	+4.07	+1.91
ans-stance	27.23	+6.54	−1.25	+8.74	−0.10
baly-stance	24.38	+4.68	−4.40	+1.86	+1.74
xlni	32.67	−2.50	+24.65	+16.32	+14.26
sts	5.50	−0.40	+10.84	+1.61	+6.71
mq2q	85.41	−1.31	+4.36	−6.95	+2.67
topic	87.53	+1.15	−47.32	+4.14	+2.94
qa	40.74	−4.28	−3.91	−19.38	−4.19
wsd	64.63	+0.36	+1.04	−0.94	+1.22
ORCA$_{score}$	45.84	−0.27	−3.29	+4.14	+1.43

Table 14: Difference in performance of T5 model on Test set compared to Gold(F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold**: Data obtained from Arabic Wikipedia, **MT**: Data obtained through Machine Translation, **OCR**: Data obtained through Optical Character Recognition, **ASR**: Data obtained through Automatic Speech Recognition, **TG**: Data obtained through Text Generation. Token Level Tasks were excluded as they yielded scores of 0 F_1 score.

Cluster	Task	Gold	MT	OCR	ASR	TG
SC	abusive	67.66	67.43	64.84	67.97	66.07
	adult	73.06	76.76	74.60	88.37	89.35
	age	68.37	87.47	69.94	96.09	96.41
	ans-claim	56.62	55.89	62.07	64.81	65.15
	dangerous	65.44	64.59	57.53	64.42	66.19
	dialect-binary	67.12	67.01	67.59	83.18	82.54
	dialect-country	16.11	15.99	21.73	22.68	22.91
	dialect-region	82.35	84.57	83.11	84.29	84.38
	emotion	48.86	47.69	47.62	48.24	47.38
	emotion-reg	11.61	10.33	26.51	24.39	20.03
	gender	49.74	41.90	35.27	35.27	43.09
	hate-speech	64.10	65.12	54.71	72.11	66.27
	irony	79.36	77.72	78.84	77.89	78.45
	offensive	80.44	78.19	79.88	78.98	78.97
	machine-generation	75.41	73.02	75.89	75.89	75.70
	sarcasm	60.08	61.67	60.01	70.04	66.89
	sentiment	80.73	80.01	79.62	80.22	79.72
SP	arabic-ner	61.35	54.95	55.14	62.11	65.59
	aqmar-ner	55.58	49.77	49.22	53.52	56.87
	msa-pos	16.85	15.22	17.02	19.60	15.39
	dialect-pos	81.78	81.91	81.64	86.16	84.79
NLI	ans-stance	38.93	43.45	37.44	50.09	45.35
	baly-stance	30.99	33.73	20.44	39.14	38.51
	xlni	34.32	17.62	30.16	46.27	36.43
STS	sts	71.06	70.51	70.17	69.85	70.20
	mq2q	55.66	53.84	54.02	53.64	53.36
TC	topic	91.99	90.71	90.14	91.86	91.94
QA	qa	31.19	14.60	15.64	21.06	29.43
WSD	wsd	42.39	36.22	47.64	56.14	38.87
ORCA$_{score}$		51.88	50.81	51.36	56.33	54.60

Table 15: Performance of $\text{BERT}_{\text{Base}}$ model on Dev set (F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold** : Data obtained from Arabic Wikipedia, **MT** : Data obtained through Machine Translation, **OCR** : Data obtained through Optical Character Recognition, **ASR** : Data obtained through Automatic Speech Recognition, **TG** : Data obtained through Text Generation.

Task	Gold	MT	OCR	ASR	TG
abusive	66.37	64.91	68.32	63.53	64.15
adult	88.93	88.75	89.24	89.03	88.30
age	99.61	98.86	98.61	99.14	99.21
ans-claim	65.81	65.72	65.45	65.39	65.99
dangerous	67.38	65.78	64.90	64.28	65.14
dialect-binary	82.23	81.66	81.91	81.76	81.88
dialect-country	23.42	22.35	22.93	21.78	22.43
dialect-region	83.32	83.32	83.20	84.29	83.49
emotion	47.72	45.12	47.57	44.21	45.79
emotion-reg	36.76	30.45	32.50	31.28	31.82
gender	60.65	60.32	60.77	60.20	59.90
hate-speech	73.15	71.26	72.54	70.04	70.66
irony	80.16	78.52	81.88	80.18	78.42
offensive	80.87	79.81	83.21	78.93	80.01
machine-generation	76.10	75.27	75.29	75.45	75.15
sarcasm	66.00	65.88	67.25	67.66	66.24
sentiment	81.04	80.75	80.69	79.94	79.98
arabic-ner	75.19	68.80	70.82	66.01	70.71
aqmar-ner	67.85	62.78	63.86	57.91	64.53
msa-pos	28.79	20.42	15.51	22.21	24.30
dialect-pos	83.12	82.41	83.36	83.06	82.52
ans-stance	45.26	41.65	45.14	44.14	45.68
baly-stance	43.43	41.65	40.44	39.99	42.03
xlni	45.66	42.52	44.84	41.10	45.29
sts	75.31	71.96	72.19	71.22	71.38
mq2q	58.99	54.14	55.76	52.32	54.11
topic	92.56	92.21	92.49	91.42	92.10
qa	67.84	67.23	66.80	67.32	68.15
wsd	65.41	63.44	64.15	62.48	63.98
$ORCA_{score}$	65.00	61.80	63.32	62.48	63.22

Table 16: Performance of GPT-2 model on Dev set (F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold** : Data obtained from Arabic Wikipedia, **MT** : Data obtained through Machine Translation, **OCR** : Data obtained through Optical Character Recognition, **ASR** : Data obtained through Automatic Speech Recognition, **TG** : Data obtained through Text Generation.

Task	Gold	MT	OCR	ASR	TG
abusive	47.25	48.47	25.03	63.88	47.90
adult	88.26	88.38	89.79	90.24	89.70
age	44.92	43.88	32.15	65.99	47.94
ans-claim	61.17	61.01	40.47	64.72	61.65
dangerous	53.86	53.37	43.37	65.48	55.77
dialect-binary	84.16	85.25	86.45	86.45	85.62
dialect-country	11.74	11.27	3.79	22.10	12.92
dialect-region	80.42	80.40	84.75	85.44	81.35
emotion	28.81	27.88	24.27	44.15	30.91
emotion-reg	1.34	1.90	−1.50	12.06	6.61
gender	61.19	61.01	59.54	61.62	62.45
hate-speech	61.65	69.67	48.88	71.49	48.88
irony	73.98	73.61	78.46	81.61	77.95
offensive	80.70	79.50	84.20	84.20	80.93
machine-generation	76.82	77.33	80.28	79.30	78.09
sarcasm	54.80	54.86	45.86	69.40	60.60
sentiment	79.04	58.51	82.18	83.30	80.81
ans-stance	26.60	25.61	25.61	36.18	28.47
baly-stance	32.08	20.44	20.44	36.82	30.87
xlni	30.21	55.25	55.25	47.69	47.13
sts	58.51	55.12	55.12	68.08	68.47
mq2q	86.62	90.14	90.14	81.14	90.51
topic	90.58	92.10	92.10	92.01	90.51
qa	36.77	36.77	36.77	21.05	31.62
wsd	64.53	64.53	64.53	63.11	66.20
total	56.64	52.30	52.30	63.08	58.34

Table 17: Performance of T5 model on Dev set (F_1). Metric for the sts and emotion-reg tasks is spearman correlation. **Gold** : Data obtained from Arabic Wikipedia, **MT** : Data obtained through Machine Translation, **OCR** : Data obtained through Optical Character Recognition, **ASR** : Data obtained through Automatic Speech Recognition, **TG** : Data obtained through Text Generation. Token Level Tasks were excluded as they yielded scores of 0 F_1 - more details can be found on section 4.3