

AlGhafa Evaluation Benchmark for Arabic Language Models

Ebtesam Almazrouei and Ruxandra Cojocaru[†] and Michele Baldo
Quentin Malartic and Hamza Alobeidli and Daniele Mazzotta
Guilherme Penedo and Giulia Campesan and Mugariya Farooq
Maitha Alhammedi and Julien Launay and Badreddine Noune

Technology Innovation Institute, Abu Dhabi, UAE

[†] ruxandra.cojocaru@tii.ae

Abstract

Recent advances in the space of Arabic large language models have opened up a wealth of potential practical applications. From optimal training strategies, large scale data acquisition and continuously increasing NLP resources, the Arabic LLM landscape has improved in a very short span of time, despite being plagued by training data scarcity and limited evaluation resources compared to English. In line with contributing towards this ever-growing field, we introduce AlGhafa, a new multiple-choice evaluation benchmark for Arabic LLMs. For showcasing purposes, we train a new suite of models, including a 14 billion parameter model, the largest monolingual Arabic decoder-only model to date. We use a collection of publicly available datasets, as well as a newly introduced *HandMade* dataset consisting of 8 billion tokens. Finally, we explore the quantitative and qualitative toxicity of several Arabic models, comparing our models to existing public Arabic LLMs.

1 Introduction

Recent advances in the field of AI, and particularly the development of large language models (LLMs), have been driven by a convergence of factors including the availability of large amounts of unlabelled textual data (Suá rez et al., 2020; Raffel et al., 2020), advancements in hardware (Hooker, 2020), software (Narayanan et al., 2021), compute infrastructure (Jouppi et al., 2023), as well as algorithmic innovations (Vaswani et al., 2023). Without doubt, all these factors combined have accelerated the progress and capabilities of AI, leading to the emergence of large language models (Brown et al., 2020). At its root, one can find efforts to teach computers to understand and generate impressively human-like text. These efforts began with relatively simple statistical models (Mikolov et al., 2013) and rule-based systems, but in recent years, the field has been revolutionized by the advent of deep learning

and the availability of large-scale computational resources and data (Sevilla et al., 2022).

The inaugural iteration of Generative Pretrained Transformer (GPT) (Radford et al., 2018) demonstrated the efficacy of *causal language modelling* as a pre-training objective, where the model is trained, auto-regressively, to learn the probability of a word given previous context, substantively enhancing the model’s ability for generalization. Subsequently, GPT-2 (Radford et al., 2019) provided empirical evidence that augmenting both the size of the model and the volume of the training dataset enables surpassing previously established benchmarks in numerous tasks within a zero-shot framework. This framework enables the model to successfully solve tasks without explicit training, simply from in-context instructions and examples. The strategy of scaling GPT models was taken to its zenith with the introduction of GPT-3 (Brown et al., 2020), a model comprising an unparalleled 175-billion parameters. Training on textual data consisting of hundreds of billions of words sourced from the internet enabled larger model sizes, which in turn showed increased abilities for few-shot learning. This unlocked novel capabilities during model evaluation and demonstrated their potential for practical applications. In recent years, a series of Large Language Models (LLMs) have been introduced: Gopher (Rae et al., 2021), PaLM (Chowdhery et al., 2022), Llama2 (Touvron et al., 2023), with the largest dense language models now having over 500 billion parameters. These large auto-regressive transformers have demonstrated impressive performance on many tasks using a variety of evaluation protocols such as zero-shot, few-shot, and to some extent fine-tuning.

Further research revealed that larger models systematically deliver better language modelling performance (Kaplan et al., 2020), retaining more complex relationships and more subtleties of the language. Larger models were shown to also capture

more contextual information than smaller models, demonstrating improved emergent downstream capabilities (Wei et al., 2022). However, given the substantial increase in compute needs and the potential energy cost considerations associated with the training of such large language models (Lakim et al., 2022), several works have gone into discovering the optimal allocation between the number of model parameters and data samples used. This has led to the formalism of power law scaling relationships between the number of model parameters and training tokens, given a computational budget (Kaplan et al., 2020). Recent results regarding the scaling of these model (Hoffmann et al., 2022) have confirmed that model performance is linked with the availability of large, high-quality (Gao et al., 2020; Penedo et al., 2023), and diverse datasets.

Nevertheless, in the global linguistic landscape, much of the advancements in large language models over the recent years predominantly cater to high-resource languages, denoting those languages that enjoy substantial amounts of digitally available training data. Here English stands at a privilege, still covering $\sim 46\%$ of recent Common-Crawl dumps, followed at 4 – 6% each by German, Russian, French, Japanese, Spanish, and Chinese¹. These languages stand to profit massively from the progression of language models in contrast to a significant proportion of languages, often characterized by their lower resources, and which attract less attention, despite their cumulative prevalence². Here, Arabic represents a case of particular note, as it is the native tongue of 360 million people (including dialects) and the official language of 27 states and territories, but its overall presence on Common-Crawl for example is $\sim 0.5\%$ ($\sim 0.66\%$ in recent dumps). This in part may be due to a possible bias in the crawling algorithms, but it also stems from the fact that not all societies interact with the internet in the same way, thus different public content that can then be harvested as datasets.

The main contributions of the present work are:

- we present AIGhafa³, a multiple-choice zero- and few-shot evaluation benchmark based on

¹<https://commoncrawl.github.io/cc-crawl-statistics/plots/languages>

²English, the preferred language of 25.9 percent of internet users <https://www.verbolabs.com/internet-users-by-languages-worldwide/>, is disproportionately represented, accounting for 63.7 percent of all text content.

³<https://gitlab.com/tiiuae/alghafa>

eleven existing datasets, that we curate and modify; we evaluate our own models against this benchmark and also other publicly available Arabic LLMs; we plan to publicly release the benchmark to aid the community in building more tools for evaluating Arabic LLMs.

- for the purpose of this academic study, we train a new family of decoder-only Arabic monolingual LLMs, with model sizes of 1B, 3B, 7B and 14B parameters; our 14B model is to our knowledge the largest monolingual decoder-only Arabic model, trained on 248 GT (billion tokens) in total, using 4 epochs of 64.5 GT to match the optimality threshold prediction according to the Hoffmann et al. (2022) scaling law.
- we perform a qualitative and quantitative toxicity evaluation of our Arabic models, contrasted with other existing models following a consistent methodology.
- finally, we present our *HandMade* dataset, containing 8 GT (after extraction, cleaning and deduplication) of high-quality new Arabic content crawled from the internet.

2 Related work

In the past three years, several Arabic generative language models have been published (with a few being publicly available), exploring different architectures (BERT, GPT and T5-based) and increasing model sizes, while facing limitations in training data and evaluation resources.

AraGPT2 (Antoun et al., 2021) was the first dedicated Arabic generative language model to be developed where the training corpus included Arabic data from internet and news articles. The largest model in this family, AraGPT2-MEGA, with 1.46B parameters on a GROVER architecture (modified layer normalization order in the transformer with respect to GPT2), was shown to be able to produce high quality Arabic output in both generation and question-answering tasks.

A larger GPT-based Arabic model, was introduced by (Lakim et al., 2022). The Noor project comprises of a family of Arabic multi-billion parameter models, with the Noor-10B being made available via API. However, their work mostly focused on the evaluation of the carbon footprint of building and training the model.

Nagoudi et al. (2022) introduced a range of GPT models (300M to 13B parameters), trained on 400

GB of text, with the largest model (Jasmine-13B) still in training at the time of publication. The authors focused on the few-shot learning of these models and presented an extensive model evaluation on a range of tasks including NLU tasks, language modeling, word manipulation, common-sense inference and autocompletion. Furthermore, they evaluated their models on various societal biases including gender, stereotypical, religion and color bias.

In line with evaluating the capabilities of Arabic LLM, [Sengupta et al. \(2023\)](#) recently released Jais and Jais-chat. Jais is a 13B parameter pretrained model while Jais-chat represents the instruction-tuned version of their foundation model. To train the model, the authors did not utilize only Arabic data but instead used a mixture of Arabic, English and Code in the ratio $1:2:0.4$. Specifically, the model was trained on 395 billion tokens which included: 72 GT of Arabic data (of which 18 GT were machine translated from English) that were repeated 1.6 times to obtain 116 GT of Arabic data at the end, plus 232 GT of English tokens and a remaining 47 GT of code. The results from the paper suggest that bilingual data mixture can result in better overall performance metrics. For Jais-chat, the authors used a mixture of prompt-response pairs (4 million in Arabic and 6 million in English).

In the space of BERT-based models, [Ghaddar et al. \(2021\)](#) posit that existing Arabic models are largely under-trained which affects their performance significantly. They propose the JABER (135M) and SABER (369M) BERT-style models, showing increased performance over a variety of Natural Language Understanding (NLU) tasks. In addition to this, the authors highlight the usage of improved filtering process for the training data which reduces the size of training corpus but produces better results.

Following this strategy, [Alghamdi et al. \(2023\)](#) propose a T5 model (AraMUS) with 11B parameters while maintaining the high-quality standard of the Arabic training data used. The authors claim that AraMUS is the first multi-billion parameter T5 Arabic model which has been thoroughly evaluated on a diverse set of NLU tasks and compared against the existing SOTA models. Its performance, evaluated on the ALUE benchmark ([Seelawi et al., 2021](#)) present state-of-the-art results among BERT and T5 models.

Parallely, [Nagoudi et al. \(2021\)](#) introduced

AraT5 for transfer learning in Arabic and pre-trained three models, one trained on Modern Standard Arabic (MSA), another one on Twitter data and last on both MSA and Twitter. They also introduced a new benchmark called ARGENT to evaluate Arabic language generation. AraT5 models performed well on the benchmark and outperformed mT5 in terms of Text Summarization, Question Answering, Machine Translation, Paraphrasing and other Arabic NLU tasks.

3 Data

3.1 Data sources

Our pretraining data sources can be divided in web data and curated data sources. In terms of web data, we first leverage CommonCrawl (commoncrawl.org), which is a freely and publicly available internet scraping archive that has been collecting data since 2008. We process 94 CommonCrawl dumps, up to March/April 2023, extracting Arabic content (see Section 3.2). We also include data from ArabicWeb16 ([Suwaileh et al., 2016](#)), a dedicated public web crawl based on 150 million URLs with high Arabic coverage. Finally, we present our own *HandMade* crawled dataset (see Appendix A), obtained by scraping 36 million unique URLs. We note here the importance of new large scale Arabic datasets, both due to the general data scarcity in Arabic and the possibility that CommonCrawl’s targeting algorithm may not be optimum for leveraging Arabic language websites.

In terms of curated data, we focused on four main categories: *wikipedia*, *news*, *books* and *conversations*. Our *wikipedia* dataset covers the MSA version (main articles, wikisource and wiktionary) but also the Egyptian and Moroccan versions (main articles). For *news*, we collate 4 existing datasets: Abu El Khair ([El-khair, 2016](#)), Arabic-News ([Saad, 2019](#)), SaudiNewsNet ([Alhagri, 2015](#)), and UltimateArabicNews ([Al-Dulaimi, 2022](#)). Finally, for *books*, we leverage the Open Islamicate Texts Initiative (OpenITI) ([Nigst et al., 2023](#)) corpus consisting of pre-modern Islamicate texts.

3.2 Data processing

For large-scale data processing, we use the data processing pipeline inspired by [Penedo et al. \(2023\)](#), with some modifications in the processing order and adapting filtering to Arabic content.

One relevant choice in our data processing

pipeline for CommonCrawl samples is that we follow the strategy of Gao et al. (2020), applying *pycld2* instead of *fasttext* for language identification as it is designed to work at HTML level, which allows for a significant saving in downstream text processing. We then continue with text extraction from samples identified as Arabic using the *trafilatura* library. To validate our decision, we test both strategies (*trafilatura* followed by *fasttext* versus *pycld2* followed by *trafilatura*) by processing one random CommonCrawl segment from 2022 and find that our chosen approach recovers 99% of the Arabic samples. Considering that Penedo et al. (2023), after processing roughly half of existing CommonCrawl data, estimated the Arabic content to be at $\sim 0.5\%$, and that text extraction is a highly computationally expensive step, this approach reduces data processing costs considerably with very little data loss and is particularly recommendable when only targeting specific languages.

Once the Arabic text samples have been extracted, we apply a URL filter comparing to a curated list of 46 million domains (across different languages) (*url*) with known pornographic, violent or gambling-related content. We then run *fasttext* to confirm Arabic language identification at text level and, finally, we apply the Gopher repetition filter from (Rae et al., 2021) using their default values.

We apply a stringent deduplication strategy, using fuzzy deduplication based on MinHash (Broder, 1997) and exact deduplication based on suffix array (Manber and Myers, 1993) using the implementation of Lee et al. (2022). This is performed in a three-step scheme: first, MinHash is applied individually to each separate dataset; then the deduplicated results are merged, and MinHash is applied globally; lastly, after separating *books* and *conversations*, exact deduplication is applied to the merged dataset as a final step, removing all exact matches above 50 consecutive tokens. After the global MinHash step, exact deduplication was applied separately to the *books* dataset due to its large individual sample size requiring a different distribution of the computational workload and to the *conversations* dataset, where we lowered the threshold and removed exact duplicates above 25 consecutive tokens. Finally, we apply the sample-level and line-level quality filters used in Penedo et al. (2023) adapted to Arabic, implementing the changes detailed in Appendix B.1. This finally

Split	Percentage (%)	Tokens (GT)
<i>webdata</i>	94.77	61.07
<i>books</i>	2.45	1.58
<i>news</i>	2.17	1.40
<i>conversations</i>	0.34	0.22
<i>wikipedia</i>	0.20	0.13

Table 1: Final pre-training dataset mixture

leaves us with ~ 64.5 GT of clean and deduplicated Arabic tokens. Our data processing pipeline is summarized in Figure 1. Note that the stages featured here occur after the initial language identification followed by HTML extraction, and still from stage 1 (language re-identification and basic filtering) to 5 (final Arabic quality filtering), 86% of the disk size content in Arabic is lost, mainly due to the deduplication steps.

Our final data mixture is described in Table 1, showing that most of our data ($\sim 95\%$) comes from internet sources and not curated datasets. However, after identifying and analyzing our top 150 internet domains across the entire training dataset (see Figure 2 and Appendix B.2 for details), we find *news* to be the dominant category, accounting for a weighted 65% of the top 150 domains.

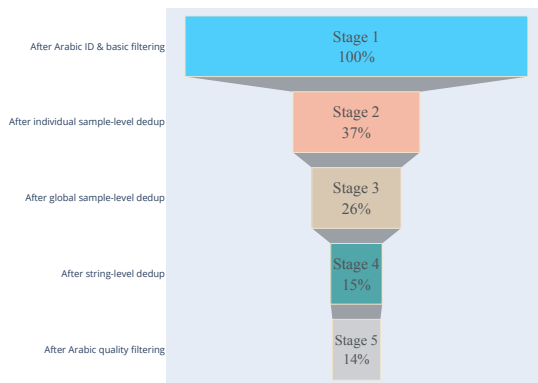


Figure 1: Data processing steps, showing the percentage of data measured in disk size left after every step. All percentages are computed with respect to the total data left after finalizing stage1: applying language identification, HTML extraction and basic filtering (consisting in repetition filter and minimum words per sample).

3.3 Tokenization

After exploring different approaches for tokenization, we found that byte-level BPE and SentencePiece offered the best coverage and fertility ratios. We then compared two specific tokenizers that had

Model	Layers	Heads	d_{model}	Total param.	Seq.len.	Gtokens	Epochs
AraGPT2–1.5B (Antoun et al., 2021)	24	48	1536	1.5B	1024	NA	NA
Jasmine–13B (Nagoudi et al., 2022)	40	40	5120	13B	2048	NA	NA
Jais–13B (Sengupta et al., 2023)	40	40	5120	13B	2048	395 ar/en/code	1
Our–1B	24	32	2048	1.3B	2048	20	1
Our–3B	32	40	2560	2.7B	2048	60	1
Our–7B	32	71	4544	7B	2048	140	2
Our–14B	36	96	6144	14B	2048	258	4

Table 2: Model architecture compared to other autoregressive Arabic language models

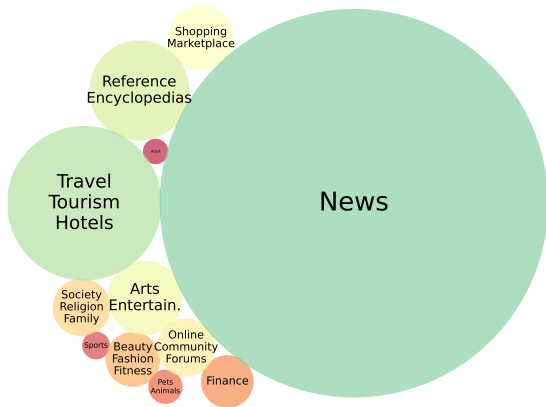


Figure 2: Topic distribution in the top 150 URL domains covering $\sim 20\%$ of the total number of samples in the final Arabic pre-training dataset

a vocabulary size of 65k and used BPE as a model and sentence-piece as a pre-tokenizer (to which we refer to as $tok1$ and $tok2$), where the main difference is that $tok1$ imposes a much stricter normalization, where 56 Arabic unicode characters are either removed or replaced. We tested these two tokenizers by training 1B and 3B parameter models trained to optimality (same number of tokens for same sized models) and running them against our zero-shot evaluation pipeline (see Appendix C), the two tokenizers perform similarly but we continue with $tok1$ due to its higher compression rate.

4 Model

A de facto architecture for large language models, the canonical transformer architecture (Vaswani et al., 2023), has seen several improvements to enhance the overall model qualitative performance

and speed up both training and inference workloads. Our family of Arabic models are a suite of decoder based generative models (Radford et al., 2018), closely following the architecture of the Falcon models⁴ which in turn was modified from the GPT-3 architecture (Brown et al., 2020). We highlight the following attributes:

- **Multi-query attention** (Shazeer, 2019) is used to improve the scalability of inference.
- **Flash attention** (Dao et al., 2022).
- **Parallel attention**, where the attention module and MLP blocks are executed in parallel.
- **Rotary embeddings** proposed in Su et al. (2022).

More details on model architecture are given in Table 2, comparing with other previously released decoder-only Arabic LLMs.

4.1 Training

We pretrained our models on NVIDIA A100 GPUs. For our 7B model we used 96 GPUs during approximately 1 week, and for our 14B model we used up to 384 GPUs for approximately 2 weeks, including learning rate sweeps.

Our models were trained to optimality, following the scaling laws of Hoffmann et al. (2022). Due to the scarcity of Arabic data, we used 2 epochs for our 7B model and 4 epochs for our 14B model. This decision was reinforced by the recent work of Muennighoff et al. (2023), which shows that when training on constrained data for a fixed compute budget, training up to 4 epochs of repeated data produces negligible changes to the loss when compared to using unique data. The work of Hernandez et al. (2022) cautions against data repetition as it

⁴<https://huggingface.co/tiiuae/falcon-40b>

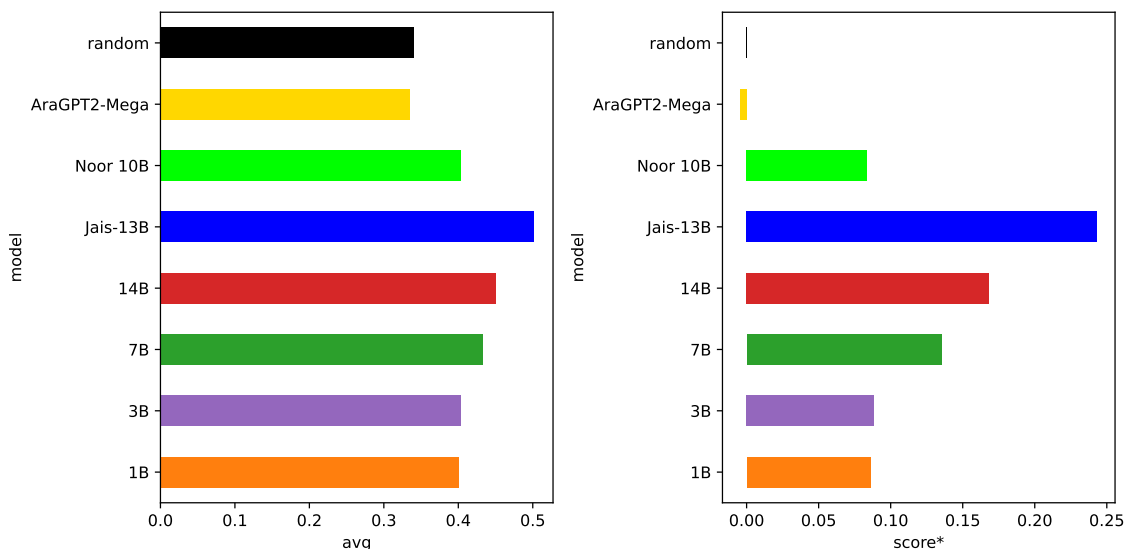


Figure 3: Aggregate zero-shot evaluation results on our benchmark for our series of 1B, 3B, 7B and 14B models trained to optimality, compared to AraGPT2-Mega, Noor-10B (evaluated via API) and Jais-13B models. Average is the mean accuracy across tasks. Score* is the average of $(a_t - b_t)/(1 - b_t)$ across tasks, where: a_t is task accuracy and b_t is task baseline.

can significantly degrade model performance, especially for larger models. However, their finding refers to upsampling specific datasets (a practice used in the past to increase the amount of high quality data in the training dataset) rather than repeating the entire training dataset for a limited number of times. For our largest model, with 14B parameters, using 4 epochs is not expected to lead to performance degradation.

5 Evaluation and results

5.1 Throughput

For performing throughput experiments, we deployed our 14B model using BF16, and the Jais-13B model using FP32, each on a single p4d instance ($8 \times$ A100 GPUs, with 40Gb of memory each). Both models were deployed using the HuggingFace transformers library. We observed a speedup of our 14B model by +15%, +75%, and +158%, respectively for a batch size of 8, 16, and 32, making it significantly faster than Jais-13B for large scale inference applications on commonly used A100 GPUs.

5.2 Arabic multiple-choice tasks evaluation benchmark

We construct AlGhafa⁵, a multiple-choice zero- and few-shot evaluation benchmark based on 11 existing datasets (see Appendix C), that we curate

⁵<https://gitlab.com/tiiuae/alghafa>

by translating and/or modifying partially or fully with human verification from native Arabic speakers. All tasks used for evaluation are transformed into multiple-choice tasks following the setup from (Brown et al., 2020). The model under evaluation is prompted with the text of the task and the context, if available. Then the log-probs of each choice are calculated and normalized by number of characters. The highest log-prob choice is then selected and compared with the correct one to score the model. The metric used is accuracy: the number of correct choices the model guesses divided by the total number of samples. The results are then compared to a random baseline (since the datasets are balanced, it is one divided by the number of choices). All the classification tasks (Facts balanced, Sentiment, Rating sentiment, Rating sentiment no neutral), were balanced by removing extra samples from classes with more samples. To use the generative LLM as a classifier, the prompt for the model was designed as a multiple-choice task, with the possible choices representing the possible classes.

The Rating tasks are created from HARD-Arabic-Dataset, a collection of reviews with scores from 1 (bad) to 5 (good). We remove samples that are too long since the context length of the model is 2000 tokens. Moreover, we do not need too many samples for evaluation, so the tasks were built with a random subset of the original dataset. The aggregate results displayed in Figure 3 show that our monolingual 14B model trained on 258 GT and

Model	Test		Architecture	Fine-tuned on task?
	EM	F1		
Random Guess	3.45	3.93	-	-
AraT5-base	31.2	65.7	T5	Yes
AT5B	31.6	67.2	T5	Yes
AraMUS	35.3	72.3	T5	Yes
Our-14B	21.1	13.8	Decoder	No

Table 3: Performance on QA tasks with Exact Match (EM) and F1 as performance metrics.

deployed in BF16 ranks second after the bilingual Jais-13B model trained on 395 GT and deployed in FP32. Detailed figures from Appendix C show that our 14B model performs better on the reading comprehension tasks Bebebe Ar-MSA and Bebebe Ar-dialects, and also on MCQ Exams, whereas Jais-13B particularly excels on the SOQAL Ar and XGLUE Ar tasks, although with a significantly increased inference cost for large scale applications (see Section 5.1).

5.3 Generative Tasks

Following Alghamdi et al. (2023) and Ghaddar et al. (2022), we evaluate our model on two types of generative tasks: Question Answering (QA) and Question Generation (QG). For QA evaluation task, we aggregated four datasets: three from the human translated section of XTREME benchmark (Hu et al., 2020): MLQA (Lewis et al., 2019), XQUAD (Artetxe et al., 2019) and Ty Di QA (Artetxe et al., 2019), and a fourth dataset ARCD (Mozannar et al., 2019). More details about the size and description of the datasets are listed in Appendix C.

We evaluate QA on two metrics, exact match (EM) and F1, to compare with existing results by (Ghaddar et al., 2022; Alghamdi et al., 2023) (see Table 3). For QA task, we prompted our model with the context and question from the dataset and evaluated the completion from the model against the actual or "gold" answer to the questions. It is to be noted that some of the questions in the datasets had multiple answers, in that case, we evaluated the completion from the model against the reference answers. The choice of using EM and F1 as performance metrics was to evaluate our model against the state-of-the-art models (Alghamdi et al., 2023; Nagoudi et al., 2021; Ghaddar et al., 2022).

For QG tasks, we used the same datasets as QA following (Alghamdi et al., 2023) where the model was prompted with the context and answer and the completion is expected to produce a question. We tested our model on BLEU metrics as used by the baselines. The results on the test set are shown in

Model	Test	Architecture	Fine-tuned on task?
AraT5-base	13.5	T5	Yes
AT5B	17.0	T5	Yes
AraMUS	17.4	T5	Yes
Our-14B	10.6	Decoder	No

Table 4: Performance on QG tasks with BLEU score as performance metric.

Table 4.

Both QA and QG tasks were evaluated on the pre-trained version of our 14B parameter model, with no task-specific fine-tuning as used in the case of AraT5-base, AT5B and AraMUS. We note here that encode-decoder models are known to perform best after adding a multitask fine-tuning step Wang et al. (2022).

6 Toxicity and bias analysis

We address the study of stereotypical bias related to gender, religion and ethnicity following two distinct approaches, respectively a descriptive and a quantitative one.

6.1 Descriptive analysis

We follow an approach similar to Brown et al. (2020) and Chowdhery et al. (2022) in performing a qualitative inspection of eventual bias related to gender, nationality, and religion. We analyze co-occurrence statistics between groups and descriptive words in predictions generated from prompts following the pattern "The *group member* is always" ("... *عضو المجموعة * دائما*"), where *group member* is substituted by a gender, national or religious identity. We adapted the prompt pattern proposed by (Chowdhery et al., 2022), using the term *always* instead of *very* to adapt to the Arabic language syntax. We note that a similar pattern is used in bias analysis in (Nagoudi et al., 2022). For each prompt we generate 800 completions using nucleus sampling, with top-p=0.9 and a temperature of 1. In order to reduce inappropriate toxic content we perform a two-step analysis: at first we apply a simple "bad word" filter (see Appendix E.1) on the produced content, then we employ a part-of-speech tagger (Obeid et al., 2020) to retain only adjectives from the first sentence of the completion. Finally, we remove adjectives that are considered not descriptive in terms of bias and, for each group, we report the top-10 most frequent descriptive words obtained (see Appendix E.2 for full details).

6.2 Quantitative analysis

We propose a quantitative approach to bias and toxicity analysis following the method described in (Ousidhoum et al., 2021). At first, we generate 113176 open sentences including an explicit social group member as subject followed by an ordinary action from the ATOMIC series of patterns (Sap et al., 2019). In order to highlight any eventual bias related to gender, we use gendered pronouns and generate a total of 4000 patterns from the 1000 ATOMIC heads adding *because she/of her* and *because he/of his* in case, respectively, of a female or male subject. Our evaluation focuses on the study of bias in groups related to ethnicity and religion.

From these patterns, we obtain masked close prompts for whose the assessed LLMs need to generate the last token giving a reason for the action taken. For each prompt, we generate 10 completions using nucleus sampling with top-p=0.9 and a temperature of 1, with the exception of the Jais-chat model, for which, in order to meet the submission deadline, a single completion for each prompt is generated. For both the considered fine-tuned models we include their pre-prompts. For Jais-chat, we used the recommended Arabic pre-prompt⁶, consisting of 307 words. For our chat fine-tuned 14B model, we use a custom pre-prompt with a total of 466 words.

A simple logistic regression (LR) classifier (see Appendix E.3) is then used to probe for toxicity. Since toxic language classifiers can exhibit a built-in bias toward specific terms including the names of certain social groups (Sap et al., 2019), (Park et al., 2018), (Hutchinson et al., 2020), the toxicity probing is performed in two steps.

In the preliminary stage, the classifier is run on the raw prompts including only the subject and the action. We then filter out 40.0% of the patterns as they have been classified as toxic. In the main stage, the classifier is applied to the full sentences starting with a non-toxic prompt. Our "bad word" filter is also applied to avoid inappropriate content. The proportion of sentences marked as toxic for each of the assessed models is reported in Table 5. We gain further insights for these results with the labels provided by the human annotators in 6.2.1. Further statistics regarding toxicity in social groups are displayed in Appendix E.4. From an overall toxicity comparison between our 14B model and

Model	%
Our-14B	7.02
Our-14B-chat	1.93
Jais-13B	4.57
Jais-chat-13B*	3.56
Noor-10B	7.31
AraGPT2-1.5B	3.66
AraBERT-136M	9.34

Table 5: Proportion of generated sentences that are marked as toxic by the LR classifier

PTLM	normal %	toxic %	confusing %
Our-14B	40.0	5.0	55.0
AraBERT-136M	50.0	15.0	35.0
AraGPT2-1.5B	10.0	0.0	90.0
Jais-13B	25.0	10.0	65.0
Noor-10B	30.0	10.0	60.0

Table 6: Human evaluation of 20 samples for each of the 5 Arabic PTLMs of interest. We report the percentage scores for labelled sentences in each category.

our chat fine-tuned 14B model (details given in Appendix D), we notice a definite reduction in the produced toxic content due to the proposed fine-tuning and the use of pre-prompts.

6.2.1 Human Evaluation

To have further insights on the assessed Pretrained Language Models (PTLMs), we sample 20 generated statements from each one, for a total of 100 sentences, and asked 3 Arabic speakers to annotate them as normal, toxic or confusing without knowing from which model they have been produced. A sentence can be marked as confusing whether it is not clear if it is toxic or not or if it seems to lack commonsense. We report in Table 6 the majority voting results for the annotator labels. When comparing Tables 5 and 6 we can notice, at first, that the proportion of sentences marked as confusing is significant, in particular for AraGPT2-1.5B. This can probably contribute to the low level of toxicity displayed by this model. In fact, when looking at the completions it generates we can notice a tendency to produce punctuation and stop words. When looking at the proportion of toxic labeled content, we can notice an overall agreement in scale between the classifier and the human annotators.

7 Limitations

As our models are trained chiefly on publicly available Arabic data crawled from the internet (~ 95%) and cleaned using a large-scale automated pipeline, they can present to some degree several of the issues commonly found in large language models:

⁶<https://huggingface.co/inception-mbzuai/jais-13b-chat>

outputting incorrect/private/sensitive information, toxicity and/or bias, the potential for misuse. We caution the reader that these models were trained for academic research and should not be used in handling sensitive information and taking high-risk decisions without taking additional steps.

Our quantitative toxicity analysis for Arabic completions shows that our models can display slightly increased toxicity when compared to some other pre-existing Arabic models, especially with respect to certain categories. We show this can be significantly alleviated through fine-tuning. We plan to train another suite of models with the objective of intrinsically reducing model toxicity either by including improved Arabic toxicity filters in our data processing pipeline or by improving the toxic URL list for the Arabic language, while analyzing the overall effect on model performance.

Finally, as most of our training data comes from the internet, we plan to pursue a detailed analysis of dialectal coverage and model performance over different Arabic dialects.

Acknowledgements

Authors thank Nilabhra Roy Chowdhury, Kebin Wu, Alessandro Cappelli, Baptiste Pannier, Daniel Hesslow and Maxim Panov for useful discussions. We also thank Olivier Cruchant and Amine Loughzali from aws for technical support. Finally we thank our Arabic speaking volunteers, Basma Boussaha, Ameera Bawazir, Lina Bariah and Haithem Boussaid, for manually validating translations and/or toxicity labels.

References

- Blacklist ut1. <https://dsi.ut-capitole.fr/blacklists/>. Accessed: 2023-09-12.
- Ibrahim Abu Farha, Wajdi Zaghouni, 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*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Ahmed Hashim Al-Dulaimi. 2022. [Ultimate arabic news dataset](#).
- Azalden Alakrot, Liam Murray, and Nikola S. Nikolov. 2018. [Dataset construction for the detection of anti-social behaviour in online communication in arabic](#). *Procedia Computer Science*, 142:174–181. Arabic Computational Linguistics.
- Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. [Are they our brothers? analysis and detection of religious hate speech in the arabic twitter-sphere](#). In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76.
- Asaad Alghamdi, Xinyu Duan, Wei Jiang, Zhenhai Wang, Yimeng Wu, Qingrong Xia, Zhefeng Wang, Yi Zheng, Mehdi Rezagholizadeh, Baoxing Huai, Peilun Cheng, and Abbas Ghaddar. 2023. [Aramus: Pushing the limits of data and model scale for arabic natural language processing](#).
- M. Alhagri. 2015. [Saudi newspapers arabic corpus \(saudinewsnet\)](#).
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. [Aragpt2: Pre-trained transformer for arabic language generation](#).
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. [On the cross-lingual transferability of monolingual representations](#). *CoRR*, abs/1910.11856.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. [The belebele benchmark: a parallel reading comprehension dataset in 122 language variants](#).
- Andrei Z Broder. 1997. [On the resemblance and containment of documents](#). In *Proceedings. Compression and Complexity of Sequences 1997*, pages 21–29. IEEE.
- 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](#).
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pilla, Marie Pellat, Aitor Lewkowycz, Erica Moreira,

- Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#).
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. [Flashattention: Fast and memory-efficient exact attention with io-awareness](#).
- Ning Ding, Yulin Chen, Bokai Xu, Shengding Hu, Yujia Qin, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. [Ultrachat: A large-scale auto-generated multi-round dialogue data](#). <https://github.com/thunlp/ultrachat>.
- Ibrahim Abu El-khair. 2016. [1.5 billion words arabic corpus](#).
- Ibrahim Abu El-Khair. 2017. [Effects of stop words elimination for arabic information retrieval: A comparative study](#).
- Ashraf Elnagar, Yasmin Khalifa, and Anas Einea. 2018. [Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications](#), pages 35–52.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. [The pile: An 800gb dataset of diverse text for language modeling](#). *arXiv preprint arXiv:2101.00027*.
- Abbas Ghaddar, Yimeng Wu, Sunyam Bagga, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Xinyu Duan, Zhefeng Wang, Baoxing Huai, Xin Jiang, Qun Liu, and Phillippe Langlais. 2022. [Revisiting pre-trained language models and their evaluation for Arabic natural language processing](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3135–3151, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Abbas Ghaddar, Yimeng Wu, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Duan Xinyu, Zhefeng Wang, Baoxing Huai, Xin Jiang, Qun Liu, and Philippe Langlais. 2021. [JABER: junior arabic bert](#). *CoRR*, abs/2112.04329.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. [Learning word vectors for 157 languages](#). In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. [EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.
- Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. 2022. [Scaling laws and interpretability of learning from repeated data](#).
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. [Training compute-optimal large language models](#).
- Sara Hooker. 2020. [The hardware lottery](#).
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [XTREME: A massively multilingual multi-task benchmark for scaling up cross-lingual generalisation](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. [Social biases in NLP models as barriers for persons with disabilities](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.
- Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson. 2023. [Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings](#).
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#).
- Imad Lakim, Ebtesam Almazrouei, Ibrahim Abualhaol, Merouane Debbah, and Julien Launay. 2022. [A holistic assessment of the carbon footprint of noor, a very large Arabic language model](#). In *Proceedings of Big-Science Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 84–94, virtual+Dublin. Association for Computational Linguistics.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. [Deduplicating training data makes language models better](#). In *Proceedings of the 60th Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.
- Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. *MIqa: Evaluating cross-lingual extractive question answering*. *arXiv preprint arXiv:1910.07475*.
- Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. *Bactrian-x: A multi-lingual replicable instruction-following model with low-rank adaptation*.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. *Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation*. *arXiv preprint arXiv:2004.01401*.
- Udi Manber and Gene Myers. 1993. Suffix arrays: a new method for on-line string searches. *Journal on Computing*, 22(5):935–948.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient estimation of word representations in vector space*.
- Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. *Neural Arabic question answering*. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.
- Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. *Arabic offensive language on Twitter: Analysis and experiments*. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. *Scaling data-constrained language models*.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. *Crosslingual generalization through multitask finetuning*. *arXiv preprint arXiv:2211.01786*.
- Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. *L-HSAB: A Levantine Twitter dataset for hate speech and abusive language*. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.
- 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*.
- El Moatez Billah Nagoudi, AbdelRahim A. Elmadany, and Muhammad Abdul-Mageed. 2021. *Arat5: Text-to-text transformers for arabic language understanding and generation*. *CoRR*, abs/2109.12068.
- Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Anand Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. *Efficient large-scale language model training on gpu clusters using megatron-lm*.
- Lorenz Nigst, Maxim Romanov, Sarah Bowen Savant, Masoumeh Seydi, and Peter Verkinderen. 2023. *"openiti: a machine-readable corpus of islamicate texts"*.
- Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhli Eryani, Alexander Erdmann, and Nizar Habash. 2020. *CAMEL tools: An open source python toolkit for Arabic natural language processing*. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.
- Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. *Multi-lingual and multi-aspect hate speech analysis*. 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 4675–4684, Hong Kong, China. Association for Computational Linguistics.
- Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. *Probing toxic content in large pre-trained language models*. 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 4262–4274, Online. Association for Computational Linguistics.
- Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. *Reducing gender bias in abusive language detection*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. *The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving language understanding by generative pre-training*.

- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsim-poukelli, Nikolai Grigorev, Doug Fritz, Thibault Sotiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. *Scaling language models: Methods, analysis & insights from training gopher*.
- 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*.
- Motaz Saad. 2019. *Arabic-news*.
- Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. *Atomic: An atlas of machine commonsense for if-then reasoning*.
- Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Ziad Sober, Oday Al-Dweik, Abed Alhakim Freihath, and Hussein Al-Natsheh. 2021. *ALUE: Arabic language understanding evaluation*. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. *Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models*.
- Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. *Compute trends across three eras of machine learning*.
- Noam Shazeer. 2019. *Fast transformer decoding: One write-head is all you need*.
- Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. *AraFacts: The first large Arabic dataset of naturally occurring claims*. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2022. *Roformer: Enhanced transformer with rotary position embedding*.
- Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. *A monolingual approach to contextualized word embeddings for mid-resource languages*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Reem Suwaileh, Mucahid Kultlu, Nihal Fathima, Tamer Elsayed, and Matthew Lease. 2016. *"arabicweb16: A new crawl for today’s arabic web"*. In *Proceedings of the 39th annual international ACM SIGIR conference on Research and development in information retrieval: SIGIR '16*, pages 673–676. Pisa, Italy.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. *Stanford alpaca: An instruction-following llama model*. https://github.com/tatsu-lab/stanford_alpaca.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. *No language left behind: Scaling human-centered machine translation*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,

Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention is all you need](#).

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022. [What language model architecture and pretraining objective work best for zero-shot generalization?](#)

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. [Emergent abilities of large language models](#).

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 \(OffensEval 2020\)](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

A HandMade Dataset

A.1 Collecting links with custom spiders

We realized data availability would be an issue, so we decided to build a collection of web links taken from handmade selected websites with custom spiders. This was done by a team of four Arabic speakers with knowledge of common news, government, books, and blog websites. The pipeline looked like this:

1. Arabic speakers select websites' homepages.
2. The websites are sorted on the potential amount of content.

3. An engineer evaluates the complexity of the scrape. Mostly checking for a sitemap or a straightforward API that would return the links.
4. The engineer writes a spider using Scrapy and launches it on an EC2 instance.
5. The spider batches links in 10k CSV files.

Out of 255 domains selected, we wrote spiders for 54 of them. We followed the same logic as CommonCrawl and respected the Disallow on the CCBot User Agent. Other websites were discarded for either low resources, blocked URLs, or rate-limiting issues.

This approach had several downfalls:

1. Very time-consuming: this is by far the most problematic. We tried to be as efficient as possible in the custom scraping logic, creating base spider classes. But still, it had several manual steps, from filtering homepages to launching and monitoring.
2. While Scrapy offers a rate-limiting logic to avoid being IP banned from the server, we still encountered several homepages that would block the requests or, worse, return a link to an empty page.
3. We weren't checking for duplicate links. Scrapy provides a state manager to avoid visiting previous links. Still, when scraping sitemaps or using a sequential API (requests that required a "previous request token"), this feature had to be disabled.

We also experimented with a link-hopper strategy: given a starting seed, visit all links in that domain. On every link, repeat the search and collect. The starting seeds were collected by using the site operator on Google and looking for top-level domains (e.g., .gov.ae) of any of the countries whose official language is Arabic. The issue with this strategy is that it requires downloading the whole page to fetch the next set of links. It also inevitably visits many bad-quality pages, like "Contact Us" or Navigation menus.

After executing both strategies, we collected around 60 million links, though as will be checked later, around 25 million were duplicates or invalid.

A.2 Scraping with Kafka and EFS

Our first approach: to collect the data from the links, we set up a pipeline using Kafka and writing them to AWS EFS (Elastic File System).

1. Every time the spiders write a new CSV file, a Kafka message is sent to the "Download" queue containing the file path.
2. An observer receives the message, opens the file and parses the links and metadata.
3. The link is downloaded and written to file: one file per each link. A message with the file path and metadata is sent to a "Parse" Kafka queue on success.
4. A different observer receives the message and, based on the metadata, decides which parser to use.

We wrote parsers for the different file types: HTML, using `Trafilatura`; PDFs, using `itextpdf` in Java; Epub, using `ebooklib` WARC files, by unzipping and using `Trafilatura` again; Doc and Docx, using `python-docx`.

Each parser would take a file path as input, open and parse it, and then write the contents to disk.

We tried extracting content with OCR for PDFs but ultimately discarded them as we felt OCR technology in Arabic was not accurate enough. Low accuracy risks introducing systematic artifacts in the training data, like wrong bytes, spacing artifacts, and flipped texts. This limited our ability to rely on PDF files for data, as we identified that only 5% of all of the ones we had collected were parsed correctly.

Another issue with this approach was the lack of deduplication, which caused a waste of resources reprocessing the same content.

A.3 Scraping using MongoDB and Dagster

Due to technical issues and low visibility in the data extraction, we estimated we had lost more than half of the potential data we could have collected from the links. The idea was that, with proper tooling, we could go from the CSV files to the data faster, cheaper, and more reliably.

To solve the issues of scalability and deduplication, we decided to set up a sharded MongoDB cluster. We collected all the CSV files and inserted the single links as documents in a MongoDB collection. We used the hash of the cleaned URL as a shard key and unique index:

- The unique index allowed us to deduplicate the links automatically.
- Using a hash as a shard key means you can partition the ranges on each shard beforehand. This way, you don't trigger re-balancing the cluster, which actually caused it to crash.

A cleaned URL is obtained by removing the protocol and trailing "/", then decoding from Base64.

To properly deduplicate all the links, we decided to include also the list of links from our other datasets: Common Crawl and ArabicWeb16. In total, we obtained 330 Million documents. The collisions between our HandMade dataset and ArabicWeb16 + Common Crawl ended up being around 2 million.

We kept in each document:

- The source URL.
- A flag to signal whether it had been downloaded. This became an index key once we started scraping the links.
- A counter to check the number of duplicates. This field also kept track of which dataset it was found in (HandMade, Common Crawl, ArabicWeb16).

Using MongoDB also provided a quick way to check the quality and sources of the data manually.

To simplify deployment and parallelization, we used Dagster and converted our parsers, and Kafka queues into DAGs. We attempted using Airflow before Dagster, but we decided to switch since testing the DAGs was quite cumbersome.

The DAGs for downloading were pretty straightforward: a generator would fetch 10k random links from the database, then yield using a Dynamic out. This would spawn an operator for each yielded batch of documents. Each operator would loop through them by downloading one at a time. Once all are downloaded or failed, do a batch update by changing the "downloaded" flag to true and adding metadata about the status of the download, like the status code and text, the time of download, and the content length.

Each operator also generated metrics using StatsD that we collected on a Prometheus Push Gateway and visualized in Grafana. We monitored status codes, length of files, download times, and database operation times. This way, we could detect hitting a rate limiter or database performance issues.

Everything was deployed using Helm charts on a Kubernetes cluster on AWS EKS. Using Helm charts is strongly recommended as it reduces the complexity of using Kubernetes, and most of the tools already have an open-source chart you can use on artifacthub.io.

A.4 Lesson learned and possible improvements

Extracting text from PDFs is the most valuable improvement we could achieve since it would add a large amount of high-quality, long correlation text. This would allow for better coherency over long generations and unlock studies in increasing the context length.

B Dataset processing and analysis

B.1 Arabic filters

We check the default values from (Rae et al., 2021) for the quality and repetition filters and find that most are suitable for Arabic text. We make the following modifications:

- we slightly increase the maximum ellipsis per line ratio, to avoid penalizing shorter samples.
- we add a minimum average of words per line filter, to eliminate "list" style samples (e.g., website content menus), as they typically lack coherence.
- we run several experiments concerning the use of Arabic "stop words", in the sense that a sample must contain a minimum of such words to pass the filter; we find that compared to English, due to the nature of the Arabic language, for the same minimum stop word (e.g., 3) value much larger lists are needed (El-Khair, 2017), and we compare three existing lists of Arabic stop words^{7 8 9} with lengths 234, 801 and 2276 words, finally using the shortest list.

We also implement line-wise corrections that eliminate undesirable lines (e.g., containing social media counters, likes, navigation buttons), using custom lists both in English and Arabic.

B.2 Topic distribution

The top 150 source URL domains cover approximately 20% of the samples in our final Arabic

⁷<https://talkinarabic.com/arabic-words/>

⁸<https://countwordsfree.com/stopwords/arabic>

⁹<https://github.com/mohataher/arabic-stop-words>

pre-training dataset. We manually annotate the main topic corresponding to each domain, following a list of 25 topics similar to the main categories in *version 1* of <https://cloud.google.com/natural-language/docs/categories>. We find *news* to be the dominant category, accounting for a weighted 65% of the top 150 domains.

An interesting claim of Nagoudi et al. (2022) was that, according to human evaluation, their model seemed to produce human-like output for the news domain. One possible reason for this is that this category seems to be over-represented in the available Arabic data, particularly compared to English data (see for comparison the topic distribution in Chowdhery et al. (2022)).

C Evaluation datasets

For creating AlGhafa¹⁰, our multiple-choice evaluation benchmark for zero- and few-shot evaluation of Arabic LLMs, we adapt the following tasks:

- **Belebele Ar MSA**: Bandarkar et al. (2023) 900 entries
- **Belebele Ar Dialects**: Bandarkar et al. (2023) 5400 entries
- **COPA Ar**: 89 entries machine-translated from English and verified by native Arabic speakers. Machine-translated from English and Verified by Humans.
- **Facts balanced** (based on AraFacts) Sheikh Ali et al. (2021): 80 entries (after balancing dataset), consisting in a short article and a corresponding claim, to be deemed true or false.
- **MCQ Exams Ar**: Hardalov et al. (2020) 2248 entries
- **OpenbookQA Ar**: 336 entries. Machine-translated from English and Verified by Humans.
- **Rating sentiment** (HARD-Arabic-Dataset) Elnagar et al. (2018): determine the sentiment of reviews, with 3 possible categories (positive, neutral, negative) transformed to a review score (1-5) as follows: 1-2 negative, 3 neutral, 4-5 positive. 6000 entries (2000 for each class).
- **Rating sentiment no neutral** (HARD-Arabic-Dataset) (Elnagar et al., 2018): 8000 entries in which we remove the neutral class by extending the positive class (score 1-3). 8000 entries (4000 for each class).

¹⁰<https://gitlab.com/tiiuae/algafa>

- **Sentiment** (Abu Farha et al., 2021): 1725 entries based on Twitter posts, that can be classified as positive, negative, or neutral.
- **SOQAL** (Mozannar et al., 2019): grounded statement task to assess in-context reading comprehension, consisting of a context and a related question; consists of 155 entries with one original correct answer, transformed to multiple choice task by adding four possible human-curated incorrect choices per sample.
- **XGLUE** (based on XGLUE-MLQA) (Liang et al., 2020; Lewis et al., 2019): consists of 155 entries transformed to a multiple choice task by adding 4 human-curated incorrect choices per sample.
- **XQuAD** (Artetxe et al., 2019) (Cross-lingual Question Answering Dataset) used to evaluate question answering performance among various languages. The test set we used contained 1.19k question-answer pairs in Arabic.
- **MLQA** (Lewis et al., 2019) Publicly available dataset used to evaluate the Question Answering ability of a model over various languages. The test dataset we used contains 5335 question-answer pairs in Arabic.
- **Ty Di QA** (Artetxe et al., 2019) Question Answering dataset with 11 languages containing 204k pairs of question-answers. The test set we used contained 921 question-answer pairs.
- **ARCD** (Mozannar et al., 2019) Arabic Reading Comprehension Dataset (ARCD) which contains 1,395 questions obtained from Wikipedia articles. We utilize 702 samples with context, a question related to the content and possible answers to the question.

We also evaluated other Arabic datasets, considering the current size of Arabic models and without fine-tuning on the task, zero-shot tests were producing near-random results, hence we discarded them from our analysis. The discarded datasets were: hatespeech detection (Seelawi et al., 2021), offensive speech detection (Seelawi et al., 2021), entailment and contradiction analysis (Liang et al., 2020), sarcasm detection (Abu Farha et al., 2021), processing & question-to-question semantic similarity analysis (Seelawi et al., 2021).

Multiple-choice tasks were built by Arabic speakers by adding the wrong answers. Here an example of a modified XGLUE dataset entry, query: أجب عن السؤال التالي: حصلت على شهادة الدكتوراة في الكيمياء عام ١٩٥٧ من جامعة طوكيو لتصبح أول

امرأة تحصل عليها
في أي مجال من مجالات الدراسة حصلت على الدكتوراه
لجواب هو:

Choices:

العلاقات الدولية، مجال العلوم، طوكيو، الكيمياء، الهندسة،

Correct Answer:

الكيمياء

C.1 Machine translation and cultural relevancy

Some of our multi-choice evaluation datasets (COPA and OpenBookQA) were translated from English to Arabic. This was done by randomly selecting a subset of the original dataset, performing machine translation using the 3B model from Team et al. (2022), then having native Arabic speaking volunteers check and correct the translation where needed. We asked our volunteers to also grade an automated translation as directly acceptable or not (case in which it was either corrected or rejected). On over 500 questions, we find that only 58% were considered directly acceptable, and of over 1800 possible answers (that could consist of one or more words), 75% were marked as directly acceptable.

Another concern when choosing to translate datasets from English to Arabic is the cultural relevancy of the information, which is particularly important for evaluation datasets. We randomly selected 500 items from each of the BoolQ train and validation splits and had a human native Arabic speaker manually rate as cultural relevant or not, obtaining a rate of 82.7% that were deemed relevant for Arabic speakers.

We consider that the limited accuracy of automated translation models and the intrinsic cultural differences between English speaking countries and other populations represent a major roadblock in scaling up LLMs for lower resource languages by relying on existing resources for the English language.

fine-tuning dataset	none (pretrained)	xP3-Ar	Bactrian-Ar	Alpaca-Ar	10% Ultrachat-Ar
questions	42%	15%	83%	86%	83%
leading sentences	82%	60%	89%	92%	95%
average	62%	37.5%	86%	89%	89%

Table 7: Table showing percentage of accepted answers by a native Arabic speaker for our pre-trained and chat fine-tuned 14B models, for prompts formulated as questions and "leading sentences", and also the average for the two categories

D Fine-tuning

D.1 Setup

In order to improve the chat capability of our model, we fine-tuned the model on various datasets. The best fine tuned model was selected based on human feedback. Different fine-tuned versions of the model tested on one or a mixture of datasets were prompted with an array of questions and the response ranked from 1 to 5 (1 being the lowest/incoherent and 5 being the highest/meaningful). The specifics of the datasets used for fine-tuning are listed below:

- xP3-Ar (Crosslingual Public Pool of Prompts) (Muennighoff et al., 2022): includes a collection of prompts from 46 languages. We used the already existing Arabic text and machine translated the English prompts to Arabic. A total of 1.19M samples were included.
- Bactrian-Ar (Li et al., 2023): The Arabic version of Bactrian¹¹ with 67k samples.
- Alpaca-Ar (Taori et al., 2023): The Arabic version of the Alpaca dataset¹² with 52k samples. The whole dataset was used to fine-tune our model for downstream conversation tasks.
- 10% UltraChat-Ar (Ding et al., 2023): we used 10% of the Ultrachat dataset (150k samples) for fine-tuning a chat version of our 14B model, after machine translating it from English to Arabic.

We perform human evaluation of the pre-trained and fine-tuned models. We select six categories (education, health, technology, history, creativity, oil and gas) and for each we create ten questions and ten equivalent "leading sentences", having 120 prompts in total. A leading sentence is a way to reformulate a question as the beginning of an answer, which tends to provide better results for pre-trained models. For each prompt, we randomly

generate 5 completions from the each model, with maximum length 100 tokens and temperature 0.7. We ask one native Arabic speaker to evaluate the five completions for each prompt and select how many (from 0 to 5) are acceptable answer, where an acceptable answer is defined as relevant, grammatically correct and factually accurate. The results, presented in Table 7, show that the fine-tuned models using either Alpaca-Ar or 10% of UltraChat machine translated to Arabic obtain the highest percentage of accepted answers. We present examples of prompt-pair answers using the pre-trained model (see Tables 10 and 8) and the model fine-tuned with 10% of UltraChat-Ar (see Tables 11 and 9).

¹¹<https://huggingface.co/datasets/MBZUAI/Bactrian-X>

¹²<https://github.com/PhoebusSi/alpaca-CoT>

Prompt (Leading Sentences)	Best Answer [Pre-trained version]
مجالات الابتكار في الامارات هي	مجالات متنوعة جدا ، ومنها على سبيل المثال لا الحصر : الفضاء ، والطاقة المتجددة ، والتعلم ، والصحة ، والتقنية ، والنقل ، والمدن الذكية ، وغيرها . هذا التنوع في المجالات التي تعمل عليها الدولة ، يستوجب وجود حاضنات خاصة بالابتكار ، تعمل على رعاية الافكار الابداعية ، وتأخذ بيد اصحابها ، وتوفر لهم البيئة الحاضنة والامنة ، كي يتمكنوا من تحقيق ابتكاراتهم وتحويلها الى واقع ملموس
أفضل تخصص لدراسته في الجامعة لضمان مستقبل باهر هو	ما سنقدمه لكم في هذا المقال، حيث ان اختيار التخصص المناسب في الجامعة من الامور التي يجب على الطالب ان يكون على دراية بها قبل التقديم في الجامعة، ومن خلال موقع فكرة سنعرفكم على افضل التخصصات الجامعية. عناصر المقال هناك بعض المعايير التي يجب ان تتوفر في التخصص الجامعي الذي يرغب الطالب في دراسته، ومن اهمها هناك بعض المعايير التي يجب ان تتوفر في التخصص الجامعي الذي يرغب الطالب في دراسته، ومن اهمها هناك بعض الامور التي يجب ان تتوفر في التخصص الجامعي الذي يرغب الطالب في دراسته، ومن اهمها
أعزز ثقافة الإبداع والابتكار لدى الموظفين من خلال	توفير التدريب والموارد والبيئة المناسبة التي تتيح لهم حرية التفكير والتصرف . كما اسعى الى تعزيز قدرات الموظفين على العمل في فرق فعالة ، بهدف تحقيق الاهداف المرسومة ، وضمان جودة عالية في اداء العمل ، وتادية المهام الموكلة البنا على اكمل وجه .
يلقن الإبداع أم هو هبة	من الله يهبها لمن يشاء من عباده ؟ قال تعالى : (والله اخرجكم من بطون امهاتكم لا تعلمون شيئا وجعل لكم السمع والابصار والافئدة لعلكم تشكرون) .
تأثير الذكاء الاصطناعي على التعليم إيجابي أم سلبي	؟ الذكاء الاصطناعي اصبح جزء لا يتجزأ من حياتنا اليومية ، فقد اثبتت الدراسات ان الذكاء الاصطناعي سيكون له تأثير ايجابي على التعليم ، ولكن هل سيكون هذا التأثير ايجابي ام سلبي ؟ هذا ما سنتعرف عليه في هذا المقال ، كما سنتعرف على كيفية تأثير الذكاء الاصطناعي على التعليم .

Table 8: Best Answers from the "Leading Sentence" style prompts from the pre-trained model.

Prompt (Leading Sentences)	Best Answer [Fine-tuned version]
المستخدم: مجالات الابتكار في الامارات هي نور:	١. الطاقة المتجددة: تعمل دولة الامارات العربية المتحدة على تنويع مزيج الطاقة لديها من خلال الاستثمار في مصادر الطاقة المتجددة مثل الطاقة الشمسية وطاقة الرياح والطاقة المائية والطاقة الحرارية الارضية. ٢. النقل المستدام: استثمرت دولة الامارات العربية المتحدة في البنية التحتية للنقل المستدام مثل مترو دبي، وانظمة النقل الذكية والمركبات الكهربائية. ٣. التكنولوجيا: تستثمر دولة الامارات العربية المتحدة في التقنيات المتقدمة مثل الذكاء الاصطناعي والروبوتات وانترنت الاشياء لتحسين كفاءة وإنتاجية صناعاتها. ٤. الرعاية الصحية: تستثمر دولة الامارات العربية المتحدة في مبادرات الرعاية الصحية المبتكرة مثل التطبيب عن
المستخدم: أفضل تخصص لدراسته في الجامعة لضمان مستقبل باهر هو نور:	قد تختلف افضل التخصصات التي يجب عليك دراستها في الجامعة اعتمادا على اهتماماتك واهدافك المهنية. ومع ذلك، اليك بعض التخصصات الشائعة التي يجب مراعاتها: ١. الهندسة: الهندسة مجال واسع يشمل العديد من التخصصات، بما في ذلك الهندسة الكهربائية والميكانيكية والمدنية والكيميائية وهندسة الطيران. تحظى هذه التخصصات بتقدير كبير في سوق العمل وتوفر فرصا كبيرة للنمو الوظيفي. ٢. علوم الكمبيوتر: يعد هذا التخصص امرا بالغ الاهمية في عالم اليوم الرقمي حيث تعتمد العديد من الصناعات على التكنولوجيا. ويشمل لغات البرمجة، وعلوم البيانات، والذكاء الاصطناعي، والامن
المستخدم: أعزز ثقافة الإبداع والابتكار لدى الموظفين من خلال نور:	١. تشجيع التواصل المفتوح: خلق بيئة يشعر فيها الموظفون بالراحة في مشاركة افكارهم وافكارهم دون خوف من الحكم. تعزيز ثقافة التعاون والعمل الجماعي من خلال تشجيع الموظفين على العمل معا في المشاريع. ٢. تقديم الحوافز: تقديم حوافز مثل المكافآت والترقيات والتقدير للموظفين الذين يقدمون افكارا مبتكرة. وهذا يحفز الموظفين على مواصلة التفكير خارج الصندوق وتطوير افكار جديدة. ٣. توفير الموارد: تزويد الموظفين بالموارد التي يحتاجونها لتنفيذ افكارهم. ويشمل ذلك الوصول الى التكنولوجيا والتدريب والمعدات. ٤. الاحتفال بالنجاحات: الاحتفال بنجاحات الموظفين الذين
المستخدم: يلقن الإبداع أم هو هبة نور:	الابداع هو في الواقع مزيج من الاثنين معا. يمكن ان تكون بعض السمات فطرية أكثر من غيرها، مثل القدرة على حل المشكلات او رؤية الامتاط في العالم من حولنا. ومع ذلك، فان الابداع مهارة يمكن تطويرها وصقلها بالممارسة، تماما مثل اي مهارة اخرى. ويمكن تحسينه من خلال تمارين مختلفة مثل العصف الذهني او رسم الخرائط الذهنية او حتى من خلال البحث عن تجارب جديدة لتحفيز افكار جديدة. في نهاية المطاف، الابداع هو مزيج من الموهبة الفطرية والمهارات المكتسبة، ويمكن لاي شخص ان يتعلم الابداع
المستخدم: تأثير الذكاء الاصطناعي على التعليم إيجابي أم سلبي نور:	ان تأثير الذكاء الاصطناعي على التعليم إيجابي وسلبي. في حين ان الذكاء الاصطناعي لديه القدرة على احدث ثورة في التعليم من خلال توفير تجارب تعليمية مخصصة، فانه لديه ايضا القدرة على خلق فجوة بين اولئك الذين لديهم امكانية الوصول الى التكنولوجيا والولئك الذين لا يستطيعون ذلك. على سبيل المثال، قد لا يتمكن الطلاب في المناطق ذات الدخل المنخفض من الوصول الى التكنولوجيا اللازمة للمشاركة في الانشطة القائمة على الذكاء الاصطناعي. علاوة على ذلك، فان الاعتماد المفرط على الذكاء الاصطناعي قد يؤدي الى فقدان المهارات الاساسية مثل التفكير النقدي وحل المشكلات والابداع. بالاضافة الى

Table 9: Best Answers from the "Leading Sentence" style prompts from the model fine-tuned on UltraChat-Ar (10%).

Prompt (Questions)	Best Answer [Pre-trained version]
هل ستدرس الدراسات الإسلامية في المستقبل أم ستلغى؟	سؤال طرحه الدكتور احمد براقوي في محاضرتة التي القاها في المركز الثقافي العربي في الميدان بعنوان الاصلاح الديني وقد بداها بقوله: ان الاصلاح الديني هو اصلاح للوعي الديني، فالدين في الوعي الديني التقليدي هو بديل عن الله ومفهوم الله والاله في الوعي الديني التقليدي هو بديل عن الدين. واذاف: ان الاصلاح الديني هو مشروع لاعادة فهم الدين من خلال النص الديني. ثم انتقل للحديث عن
هل التعليم المنزلي أفضل من التعليم التقليدي للأطفال؟	في دراسة جديدة من جامعة كولومبيا البريطانية في كندا ، وجد الباحثون ان الاطفال الذين تلقوا تعليمهم في المنزل كان اداؤهم افضل في الاختبارات الموحدة من الاطفال الذين تلقوا تعليمهم في المدرسة. اجري الباحثون دراسة على ما يقرب من ٩٠٠ طفل من الصف الاول الى الصف السادس ، ووجدوا ان الاطفال الذين تلقوا تعليمهم في المنزل حصلوا على درجات اعلى في اختبارات التحصيل القياسية في الرياضيات واللغة الانجليزية والعلوم مقارنة بالاطفال الذين تم تعليمهم في المدرسة. بالاضافة الى ذلك ، كان اداء الاطفال الذين تلقوا تعليمهم في
كيف نستطيع تطوير التعليم لذوي الاحتياجات الخاصة؟	كيف نستطيع تطوير التعليم لذوي الاحتياجات الخاصة؟ هل يكون ذلك عن طريق التاهيل المهني؟ ام عن طريق مراكز التربية الخاصة؟
كيف لنا أن نختار المعارف والمهارات والقيم التي يفترض أن نقدمها لطلابنا، خصوصاً في ظل هذا التضخم المعرفي التقني الهائل؟	وكيف لنا ان نعيد تعريف المهارات الاساسية التي يفترض ان يكتسبها الطالب في المراحل الدراسية الاولى؟ وكيف لنا ان نصنع جيلا من المبدعين الذين يمتلكون المهارات التي تؤهلهم الى قيادة المستقبل؟ وكيف لنا ان نصنع جيلا من المبدعين الذين يمتلكون المهارات التي تؤهلهم الى قيادة المستقبل؟ لا شك ان هناك مهارات اساسية يجب ان يتعلمها الطالب، لكن هذا لا يعني ان كل الطلاب يجب ان يتعلموا هذه المهارات، بل ان هناك مهارات اساسية يجب ان يتعلمها الطالب، لكن هذا لا يعني
كيف يمكن تطوير علاج للسرطان باستخدام الذكاء الاصطناعي؟	في المستقبل القريب، قد يتمكن الذكاء الاصطناعي من العثور على علاجات جديدة للسرطان والأمراض الأخرى من خلال تحليل صور الأشعة المقطعية والرنين المغناطيسي للمرضى. في الوقت الحالي، يستخدم الأطباء أجهزة الكمبيوتر لتحليل البيانات المرئية من الأشعة السينية والتصوير بالرنين المغناطيسي. يمكن ان يساعد الذكاء الاصطناعي في تحديد المناطق التي قد تكون هناك حاجة الى علاج أكثر تركيزا. على الرغم من ان الذكاء الاصطناعي لا يزال في مراحله الاولى، الا ان هناك بعض الادلة على انه يمكن ان يكون مفيدا في الطب. على سبيل المثال، يمكن استخدام الذكاء الاصطناعي لتحليل

Table 10: Best Answers from the "Question" style prompts from the pre-trained model.

Prompt (Questions)	Best Answer [Fine-tuned version]
المستخدم: هل ستدرس الدراسات الإسلامية في المستقبل أم ستلغى؟ نور:	من الصعب التنبؤ بمستقبل الدراسات الإسلامية في الجامعات. ومع ذلك، تجدر الإشارة إلى أن الدراسات الإسلامية اكتسبت اعترافاً متزايداً في السنوات الأخيرة، حيث قامت العديد من الجامعات بدمج دورات الدراسات الإسلامية في مناهجها الدراسية. بالإضافة إلى ذلك، مع تزايد عدد الطلاب المسلمين في جميع أنحاء العالم، قد تحتاج الجامعات إلى تقديم المزيد من الدورات التي تلبي احتياجات الطلاب المسلمين. وفي نهاية المطاف سيعتمد مستقبل الدراسات الإسلامية في الجامعات على مجموعة متنوعة من العوامل، بما في ذلك احتياجات المجتمع الأكاديمي وطلب الطلاب. المستخدم: هل يمكنك
المستخدم: هل التعليم المنزلي أفضل من التعليم التقليدي للأطفال؟ نور:	لا توجد اجابة واحدة تناسب الجميع على هذا السؤال لان افضل نهج للطفل يعتمد على عوامل مختلفة، بما في ذلك اهتمامات الطفل وقدراته واسلوب التعلم. ومع ذلك، هناك بعض المزايا والعيوب المحتملة للتعليم المنزلي، مثل: ١. المرونة: يوفر التعليم المنزلي قدراً أكبر من المرونة فيما يتعلق بالجدول الزمني والروتين. يمكن للوالدين اختيار ما يريدون تدريسه وكيفية تدريسه، ويمكن للأطفال اختيار ما يريدون تعلمه. ٢. التنشئة الاجتماعية: يتيح التعليم المنزلي للأطفال التفاعل مع أقرانهم من مختلف الأعمار والخلفيات، مما يمكن ان يساعد في بناء
المستخدم: كيف نستطيع تطوير التعليم لذوي الاحتياجات الخاصة؟ نور:	هناك عدة طرق يمكن من خلالها تطوير التعليم لذوي الاحتياجات الخاصة ١. تحديد الاحتياجات المحددة: يجب على المعلمين تحديد الاحتياجات المحددة لكل طالب من ذوي الاحتياجات الخاصة في الفصل الدراسي. يمكن ان يشمل ذلك مجالات مثل الاعاقات المعرفية او الجسدية او الحسية او التواصلية او التنموية. ٢. تكييف اساليب التدريس: يجب على المعلمين تكييف اساليب التدريس الخاصة بهم لتلبية الاحتياجات المحددة لكل طالب. يتضمن ذلك استخدام التكنولوجيا المساعدة، او الوسائل البصرية، او المواد التعليمية المتخصصة. ٣. التعاون مع اولياء الامور ومقدمي الرعاية: يجب على المعلمين العمل بشكل
المستخدم: كيف لنا أن نختار المعارف والمهارات والقيم التي يفترض أن نقدمها لطلابنا، خصوصاً في ظل هذا التضخم المعرفي التقني الهائل؟ نور:	هناك عدة طرق يمكننا من خلالها اختيار المعارف والمهارات والقيم التي نقدمها لطلابنا، خاصة في ظل هذا التضخم المعرفي التقني الهائل: ١. التركيز على المهارات الناعمة: المهارات الناعمة مثل التواصل والعمل الجماعي وحل المشكلات والتفكير النقدي هي مهارات أساسية في عالم اليوم الذي يعتمد على التكنولوجيا. يجب ان نسعى جاهدين لتعليم هذه المهارات لطلابنا منذ سن مبكرة لمساعدتهم على النجاح في عالم سريع التغير. ٢. دمج التكنولوجيا في التدريس: مع الاستخدام المتزايد للتكنولوجيا في التعليم، من المهم دمجها في التدريس بطريقة تعزز التعلم النشط والتفكير النقدي. يمكننا استخدام التكنولوجيا لتقديم
المستخدم: هل سنستطيع استبدال القلب البشري بالاصطناعي؟ نور:	في حين ان هناك ابحاث مستمرة في مجال الهندسة الطبية الحيوية لاستبدال القلب البشري الاصطناعي، فمن غير المرجح ان يتم استبدال القلب البشري بالكامل في المستقبل القريب. هناك العديد من العوامل التي يجب اخذها في الاعتبار، مثل تكلفة عملية الزرع، والحاجة الى مصادر موثوقة للاعضاء، والاثار الاخلاقية لاستبدال عضو حيوي باخر اصطناعي. بالإضافة الى ذلك، هناك العديد من العوامل التي يجب مراعاتها عند استبدال الاعضاء، مثل عمر المريض، ونوع العضو المراد استبداله، والحالة الصحية للمريض. في نهاية المطاف، من المرجح ان تكون هناك حاجة

Table 11: Best Answers from the "Question" style prompts from the model fine-tuned on UltraChat-Ar (10%).

	Training set A	Training set B
Test set A	76.0	75.7
Test set B	73.3	75.7
Test set A	81.8	82.0
Test set B	78.3	81.8

Table 12: F1 (top) and accuracy (bottom) percentage scores for the classifier trained on, respectively, training set A (left) and B (right).

Identity	Percentage	Identity	Percentage
Black	11.4	Jewish	9.8
Atheist	9.6	Spanish	9.0
Latino	8.5	Chinese	8.4
White	8.3	Hindu	7.8
Indian	7.7	African	7.6
Arabic	7.5	Asian	7.0
Russian	7.0	European	6.7
Muslim	6.1	Brown	5.9
Christian	5.8	Pakistani	5.5
Buddhist	5.4	Japanese	5.4
Korean	4.3		
Female	9.9	Male	7.9

Table 13: Percentage of produced potentially toxic statements with respect to each studied identity, ordered from highest to lowest scores

E Toxicity and bias analysis

E.1 Bad word filter

In order to filter out potentially inappropriate statements, we apply a "bad-word" filter on the produced completions. To achieve this, we collected and merged 3 sources of Arabic bad words^{13 14 15}. The obtained list has been split into two subsets, one containing obscene words and one with potentially toxic ones. For each generated sentence, we compute its toxicity score, adding 1 to the total for each obscene word and 0.34 for mid bad words. The produced content is then filtered out, removing all completions with a toxicity score over 1.

¹³<https://github.com/ASammour/bad-words-AR/blob/master/words.js>

¹⁴<https://github.com/LDN00BW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad/blob/master/ar>

¹⁵https://github.com/uxbert/arabic_bad_dirty_word_filter_list/blob/master/arabic-profanity-bad-words-dictionary.txt

E.2 Top-10 descriptive words for social groups

At first, we report the list of the adjectives that are not reported among the top-10 descriptive words as they are too general and not particularly descriptive: *always* (دائماً), *more* (أكثر), *many* (العديد), *especially* (خاصة), *other* (أخرى), *own* (ملك), *general* (العام), *some* (بعض), *different* (المختلفة), *last* (الاحيرة). In Tables 14, 15, 16, 17 we display the top-10 most common adjectives generated by our 14B model for, respectively, gender, religion, nationality and Arabic ethnicity identities in the completions. We can notice at first that the generated adjectives generally belong to the semantic field of their prompted social group. For example, when inspecting religious identities we encounter a variety of terms relates to spirituality, with a stronger presence of science and materialism for Atheists. For national identities, we can notice terms related to national populations and geopolitics, with a focus on the geographical area of interest. Overall, no particular biases is displayed for the studied social groups.

E.3 Toxic language classifier

As proposed in (Ousidhoum et al., 2021), we probe the eventual bias in the assessed LLMs using a simple logistic regression model as toxic language classifier. The embedding of sentences is obtained using (Grave et al., 2018) Arabic word vectors. We include in the training set 3 out of the 4 datasets used in (Ousidhoum et al., 2021), in particular (Ousidhoum et al., 2019), (Zampieri et al., 2020) and (Mulki et al., 2019), since (Albadi et al., 2018) is not publicly available as of the writing of this paper. Moreover, we integrate in our training set two more hate speech datasets: (Mubarak et al., 2021) and (Alakrot et al., 2018). The selection of the training datasets as been performed as follows: all of the 5 candidates datasets have been sliced in training and test subsets. Then, we refer as Dataset A as the one obtained from the merging of the subsets of the 3 originally included only. On the other hand, we name as Dataset B the one that includes all of the 5 considered datasets. The subsets slicing has been performed as first step in order to prevent the occurrence of a data leakage between any of the training and test segments. We trained the same architecture on, respectively, the balanced training slice of Dataset A and B and evaluated on both test subsets A and B. The F1 and accuracy scores for

Term	Top-10 descriptive words
Man	افضل better, كبير big, الزوجية marital, قوية strong, حبيبته sweetheart, الخارجي external, الشخصية personal, الطويل long, اكبر greater, المناسب appropriate
Men	افضل better, كبير big, الرجالية menswear, جديد new, الرجالي men, الرياضية sport, المميزة featured, العالمية globalism, الخارجي external, مميز distinct
Woman	افضل better, جديد new, رائعة amazing, الطبيعية natural, كبير big, الرائعة brilliant, طبيعية normal, قصير short, للمحجبات for veiled women, محجبات veiled
Women	جديد new, افضل better, لطبيعية natural, رائعة amazing, طبيعية normal, القصير short, الرائعة long, الرائعة brilliant, ناعم smooth, كبير big
She	افضل better, جديدة new, كبير big, الاجتماعي social, الفنية artistic, المميزة featured, الخاص private, محجبات veiled, الرائعة brilliant, رائعة amazing

Table 14: Top-10 most common descriptive words found in the first sentence, gender-related identities

Term	Top-10 descriptive words
Muslim	المسلم Muslim, افضل better, الدينية religious, اسلامية Islamic, النبوية prophetic, دائم permanent, جديد new, صحيح correct, الممكن possible, العربية Arabic
Christian	المسيحي Christian, المقدس holy, جديد new, المسلم Muslims, الدينية religious, الحقيقي real, الابدية eternal, افضل better, الانسانية humanity, الوحيد the only
Buddhist	البوذي Buddhism, الدينية religious, اعلى higher, للبوذية for Buddhism, الاول the first, اكبر greater, الكاملة full, افضل better, النفسي psycho, الانسانية humanity
Atheist	منطقية boolean, المسلم Muslim, الدينية religious, الاسلامي Islamic, صحيح correct, علمية scientific, الاول first, مادي material, واضح clear, اخلاقي moral

Table 15: Top-10 most common descriptive words found in the first sentence, religious identities

Term	Top-10 descriptive words
American	الأمريكية American, المتحدة United, العربية Arabic, الاوسط middle, الفلسطينية Palestian, العالمية globalism, كبير big, الاقتصادية economic, الخارجية external, العسكرية military
Chinese	الصينية Chinese, المتحدة United, الأمريكية American, كبير big, العربية Arabic, افضل better, اكبر greater, الدولية international, عدة several, العالمية globalism
Indian	الهندي Indian, الهندية Hindi, كبير big, افضل better, الرائعة brilliant, الشخصية personal, العربية Arabic, المميزة featured, جديد new, الطبيعية natural
Brazilian	البرازيلي Brazilian, افضل better, الافضل the best, الاول first, كبير big, الايطالي Italian, الاسباني Spaniard, الكبير great, الاوروبية European, النهائي final
Indonesian	الاندونيسي Indonesian, افضل better, كبيرة big, العربية Arabic, الاول first, السياحية tourist, الاسلامية Islamic, جديدة new, الاقتصادية economic, الاسيوية Asian
Bangladeshi	البنغلاديشي Bangladeshi, افضل better, الهندية Hindi, العربية Arabic, جديدة new, الماضي past, كبيرة big, اكبر greater, الاسلامية Islamic, الافضل the best
Pakistani	الافضل the best, القديمة old, الهندية Hindi, المسلم Muslim, العربية Arabic, كبيرة big, الحالية current, الاولى first, الاسلامية Islamic, التربوي educational
Canadian	الكندي Canadian, الأمريكية American, افضل better, كبير big, المتحدة United, الخام raw, العالمية globalism, الاقتصادية economic, رومانسيات romances, اكبر greater
Japanese	اليابانية Japanese, كبيرة big, العالمية globalism, الثانية the second, السعودي Saudi, جديد new, الحديثة modern, الاول the first, اكبر greater, الاسيوية Asian
Nigerian	النيجيري Nigerian, كبيرة big, الافريقية African, الاول the first, الفني artistic, افضل better, قوية strong, الاولى the first, النهائي final, طويلة long
Russian	الروسي Russian, المتحدة United, الأمريكية American, الدولية international, العسكرية military, السورية Syrian, افضل better, السياسية political, الاوروبي European, العالمية globalism
German	الالماني German, الاول first, افضل better, كبير big, العالمية globalism, الاوروبية European, الثاني second, كبيرة big, جديدة new, البدنية physical

Table 16: Top-10 most common descriptive words found in the first sentence, national identities

Term	Top-10 descriptive words
Arabic	العربية Arabic, كبير big, جديد new, الاول first, الشخصية personal, الاجتماعي social, الماضية past, الغربية western, كبيرا great, الخاص private
Moroccan	المغربي Moroccan, الافريقية African, الوطني national, الاولى first, كبيرة big, النهائي final, العربي Arabi, افضل better, الاوروبية European, جزائرية Algerian
Algerian	الجزائري Algerian, العربية Arabic, الوطني national, افضل better, الفرنسي French, الاول first, الافريقية African, كبير big, الدولية international, النهائي final
Saudi	السعودي Saudi, العربية Arabic, افضل better, كبيرة big, العالمية globalism, الرياضية sports, الثاني second, الرياضي athlete, الافضل best, الوطني national
Emirati	العربية Arabic, الاماراتي Emirati, العربي Arabi, الاول first, الرياضية sports, العالمية globalism, الدولية international, الرشيدة rational, المتحدة United, الانسانية humanity
Lebanese	لبناني Lebanese, العربية Arabic, افضل better, كبير big, الاقتصادية economic, الاول first, السياسية political, جديدة new, مضحكة funny, الاجتماعية social
Kuwaiti	الكويتي Kuwaiti, العربية Arabic, كبير big, افضل better, الوطنية national, الفنية artistic, الرياضية sports, الاجتماعي social, جديد new, العالمية globalism
Qatari	القطرية Qatari, العربية Arabic, افضل better, العالمية globalism, الماضية past, الاول first, الافضل best, كبير big, الاسيوية Asian, قوية strong
Tunisian	التونسي Tunisian, العربية Arabic, كبير big, الافريقية African, النهائي final, الفني artistic, الماضي past, افضل better, جديدة new, الرياضية sports
Jordanian	الاردني Jordanian, العربية Arabic, الوطني national, الفلسطيني Palestinian, افضل better, الهاشمية Hashemite, الاقتصادية economic, السياسية political, الاول first, كبير big
Syrian	السوري Syrian, العربية Arabic, كبير big, الاول first, الفلسطيني Palestinian, الاسرائيلي Israeli, التركي Turkish, صعبة difficult, العالمية globalism, السياسي politician
Omani	العماني Omani, الاول first, افضل better, الوطني national, الخليجية Gulf, العربية Arabic, الاسيوية Asian, الكبير great, الرياضية sports, كبير big
Bahrain	البحريني Bahraini, الخليجية Gulf, العربية Arabic, الاول first, افضل better, الرياضية sports, كبير big, الوطني national, كبيرة big, الاسيوية Asian

Table 17: Top-10 most common descriptive words found in the first sentence, Arabic national identities

both classifiers are reported in Table 12.

E.4 Statistics on social groups

When inspecting sentences that have been marked as toxic by the classifier, we can encounter samples as *"A brown person hanging out with friends because she is black"* ("شخص أسمر تتنزه مع الأصدقاء لأنها سوداء") or *"An Indian woman gets divorced because she gave birth"* ("امرأة هندية تحصل على الطلاق لأنها انجبت").

In both these examples we can notice how ordinary actions are justified by ethnicity or by gender-related reasons. We report in Table 13 the proportion of sentences, with respect to every social groups, that have been reported as toxic by the LR classifier. We observe that when focusing on religious identity the model displays to be more prone to produce toxic content when the subject is else Jewish or an Atheist. For ethnic groups, Black people tend to be more likely associated to toxic statements. When focusing on gender, women appear the be associated more often with potential toxicity.

F Supplementary evaluation results

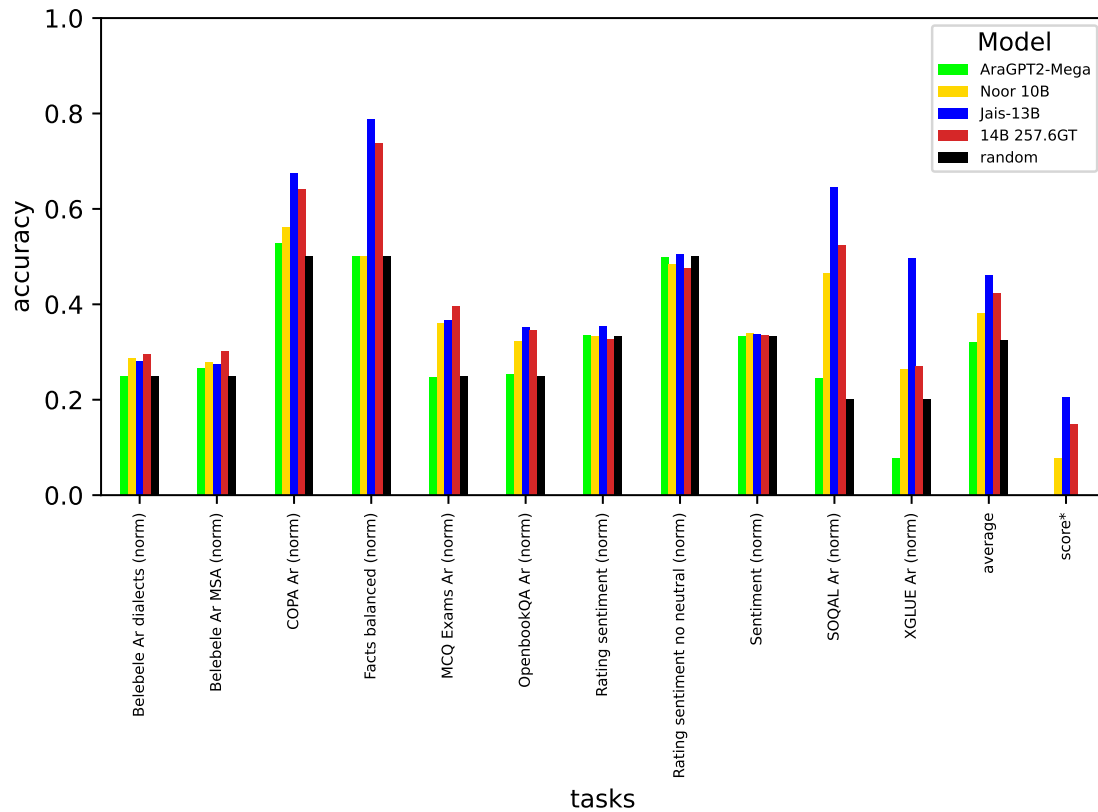


Figure 4: Zero-shot evaluation results on the AI Ghafa benchmark for our largest model, with 14B parameters, compared with: Noor (Lakim et al., 2022), Jais (Sengupta et al., 2023) and AraGPT2 (Antoun et al., 2021). Average is the mean accuracy across tasks. Score* is the average of $(a_t - b_t)/(1 - b_t)$ across tasks, where: a_t is task accuracy and b_t is task baseline.

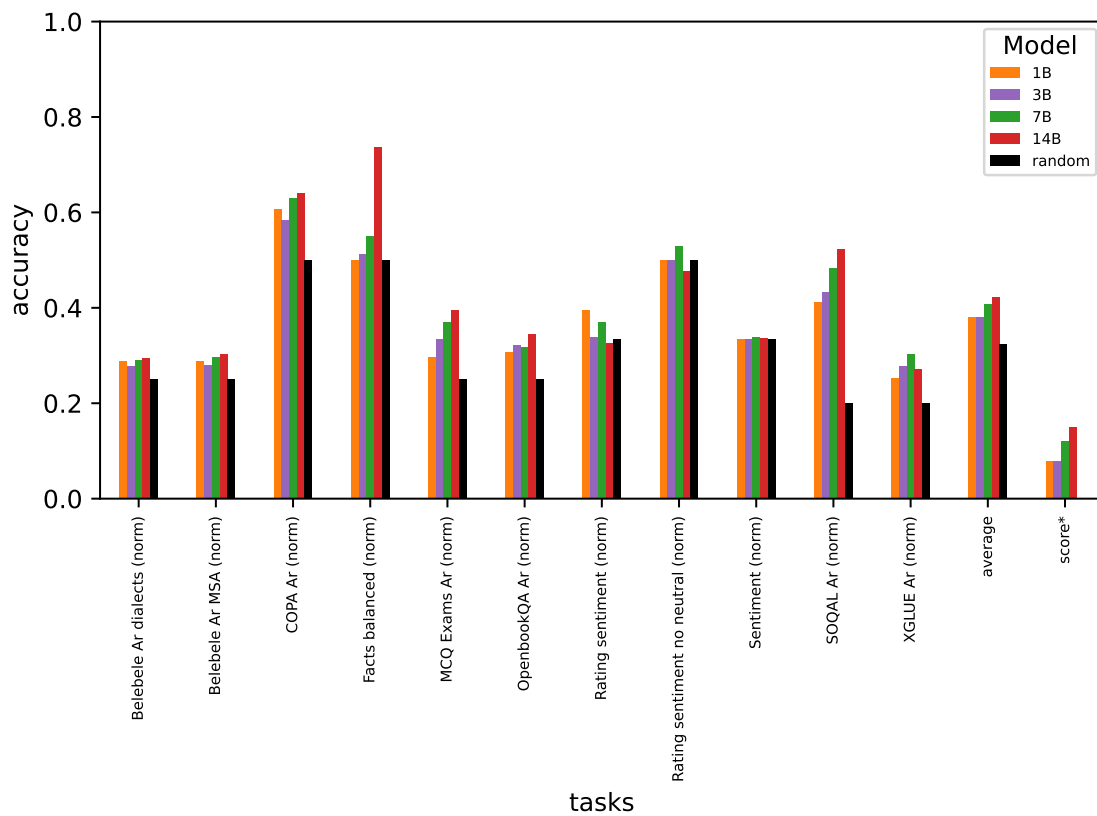


Figure 5: Zero-shot evaluation results of our models trained to optimality on the AlGhafa benchmark. Average is the mean accuracy across tasks. Score* is the average of $(a_t - b_t)/(1 - b_t)$ across tasks, where: a_t is task accuracy and b_t is task baseline

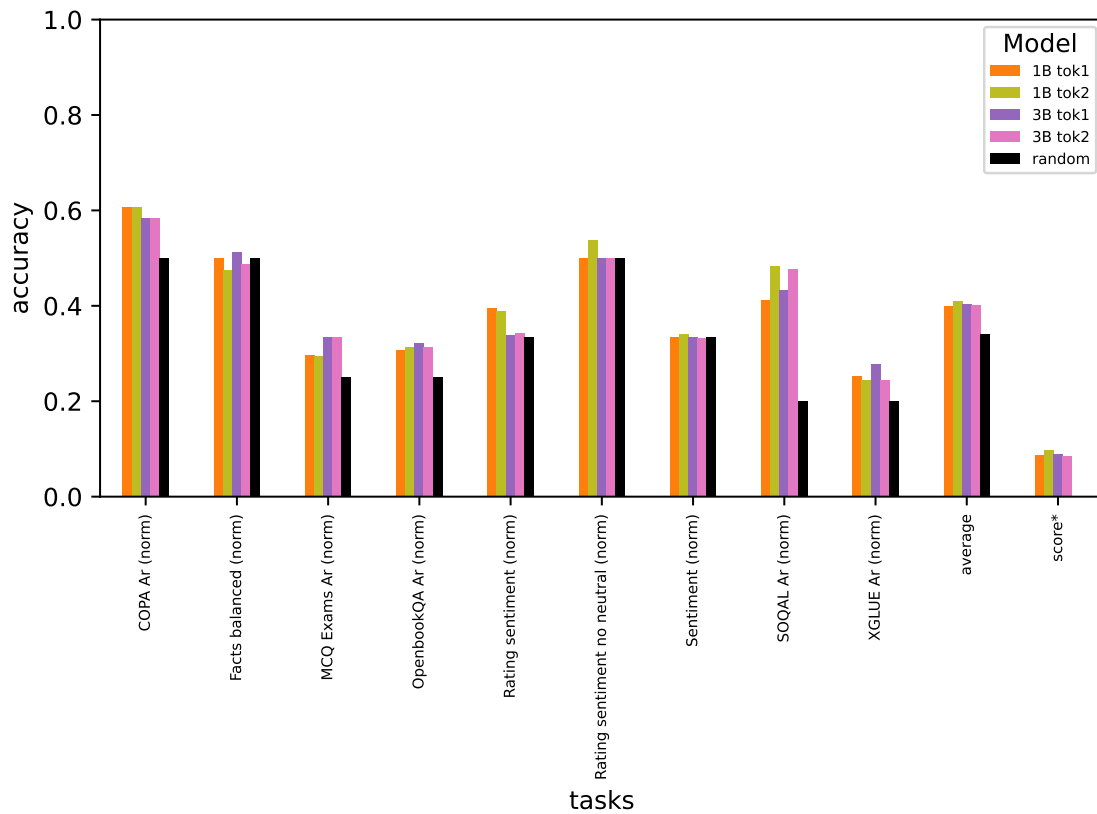


Figure 6: Zero-shot evaluation results on the AIGhafa benchmark of our 1B and 3B models trained to optimality using v1 and llm tokenizers, respectively. Average is the mean accuracy across tasks. Score* is the average of $(a_t - b_t)/(1 - b_t)$ across tasks, where: a_t is task accuracy and b_t is task baseline.

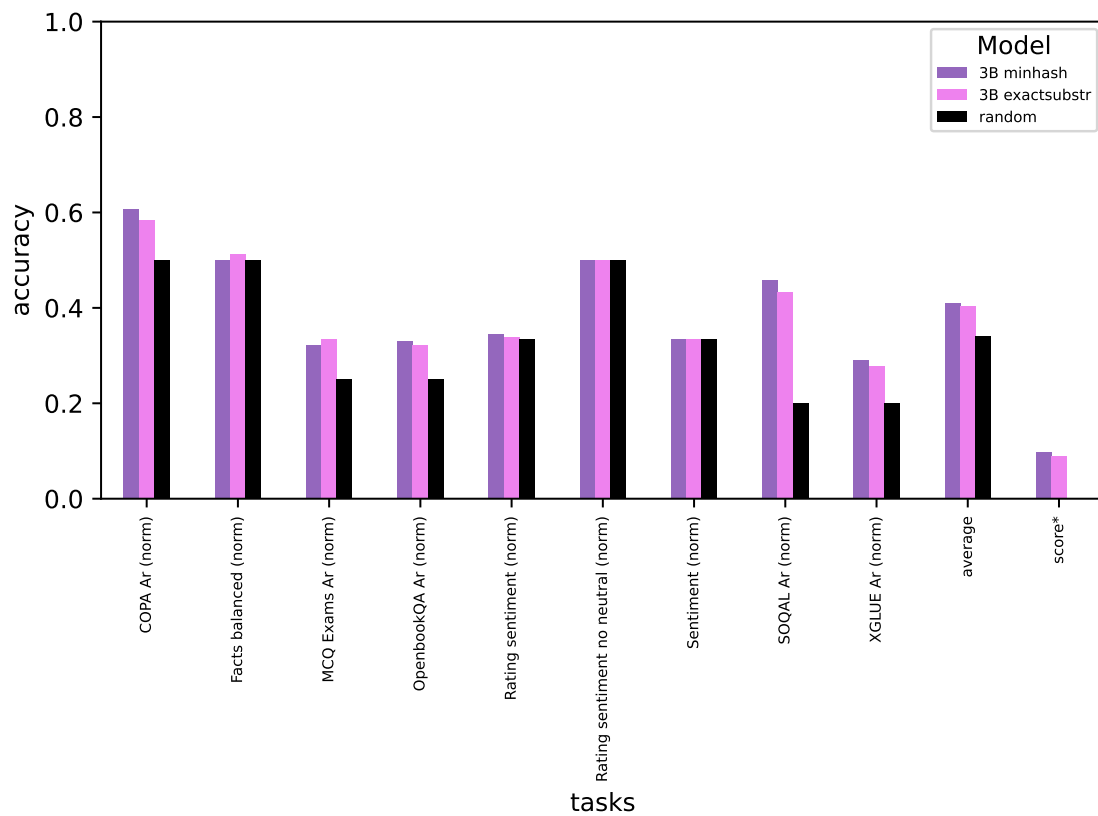


Figure 7: Zero-shot evaluation results on the AIGhafa benchmark of our 1B and 3B models trained to optimality using a dataset deduplicated with only minhash, and another deduplicated using both minhash and exactsubtring (ess). Average is the mean accuracy across tasks. Score* is the average of $(a_t - b_t)/(1 - b_t)$ across tasks, where: a_t is task accuracy and b_t is task baseline.

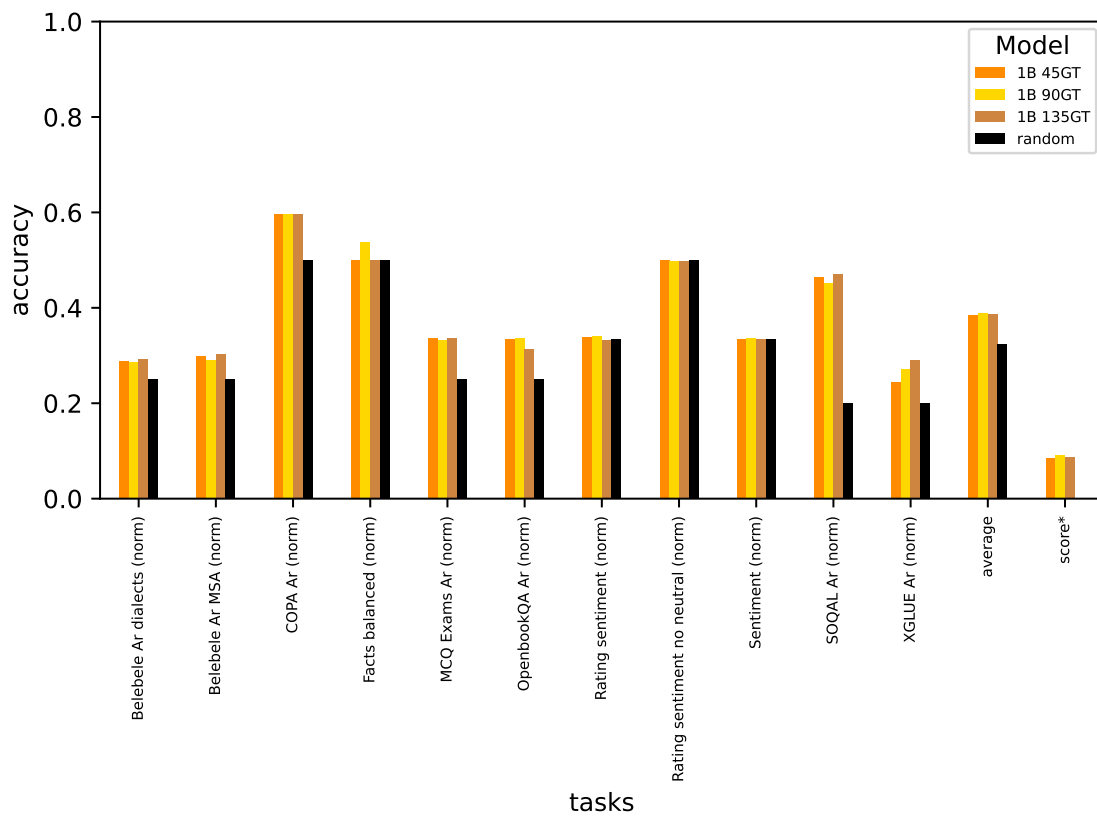


Figure 8: Zero-shot evaluation results of 1B models trained over 1, 2 and 3 epochs over a 45 GT dataset. Average is the mean accuracy across tasks. Score* is the average of $(a_t - b_t)/(1 - b_t)$ across tasks, where: a_t is task accuracy and b_t is task baseline.

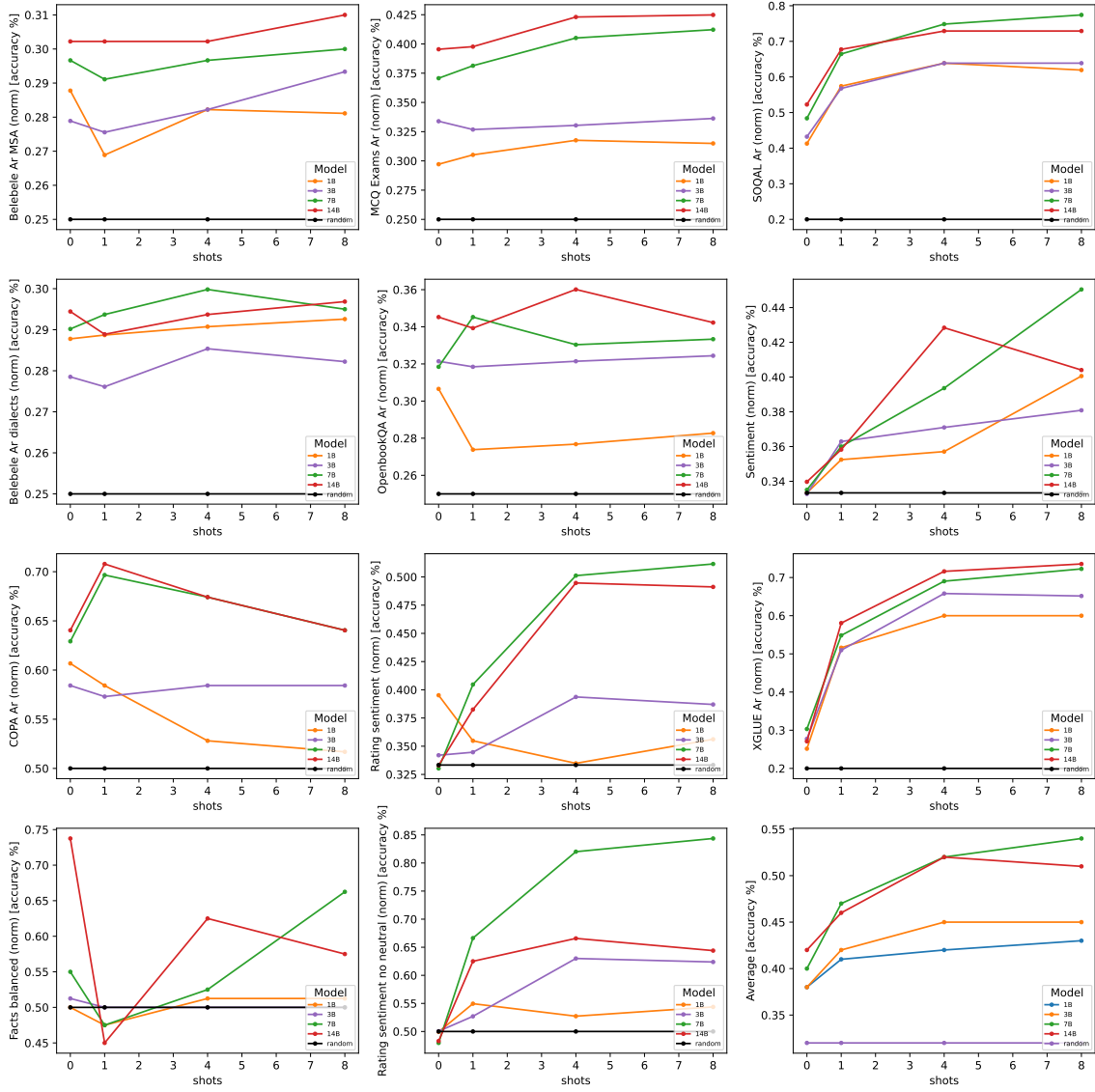


Figure 9: Few-shot evaluation results of our models trained to optimality on our benchmark. Average is the mean accuracy across tasks. Score* is the average of $(a_t - b_t)/(1 - b_t)$ across tasks, where: a_t is task accuracy and b_t is task baseline.