AraGPT2: Pre-Trained Transformer for Arabic Language Generation

Wissam Antoun and **Fady Baly** and **Hazem Hajj**
American University of Beirut
{wfa07, fbg06, hh63}@aub.edu.lb

Abstract

Recently, pre-trained transformer-based architectures have proven to be very efficient at language modeling and understanding, given that they are trained on a large enough corpus. Applications in language generation for Arabic are still lagging in comparison to other NLP advances primarily due to the lack of advanced Arabic language generation models. In this paper, we develop the first advanced Arabic language generation model, AraGPT2, trained from scratch on a large Arabic corpus of internet text and news articles. Our largest model, AraGPT2-mega, has 1.46 billion parameters, which makes it the largest Arabic language model available. The mega model was evaluated and showed success on different tasks including synthetic news generation, and zero-shot question answering. For text generation, our best model achieves a perplexity of 29.8 on held-out Wikipedia articles. A study conducted with human evaluators showed the significant success of AraGPT2-mega in generating news articles that are difficult to distinguish from articles written by humans. We thus develop and release an automatic discriminator model with a 98% percent accuracy in detecting model-generated text. The models are also publicly available[1], hoping to encourage new research directions and applications for Arabic NLP.

1 Introduction

Few years ago, Natural language processing (NLP) was revolutionized with the introduction of multi-head self-attention transformer architecture (Vaswani et al., 2017). The transformer achieved superior performance compared to recurrent neural networks several NLP tasks including machine translation, sentence classification with BERT (Devlin et al., 2019), and ELECTRA (Clark et al., 2020b), and sentence completion with GPT-2 (Radford et al., 2019), GROVER (Zellers et al., 2019), and CTRL (Keskar et al., 2019). Recent works have shown that larger models pre-trained on larger datasets can further improve performance i.e. RoBERTa (Liu et al., 2019), and XLM-R (Conneau et al., 2019).

On the other hand, work on Arabic language modeling has mostly targeted natural language understanding (NLU) by pre-training transformer-based models using the Masked Language Modeling (MLM) task i.e. AraBERT (Antoun et al., 2020a). In contrast, Arabic text generation or causal language modeling hasn't received much attention. Few works such as hULMonA (ElJundi et al., 2019) used next word prediction as a pre-training task in for transfer learning in Arabic text classification. (Khooli, 2020) and (Doiron, 2020) leveraged the existing GPT2 English model and adapted it for Arabic using text from the Arabic Wikipedia dumps, which is sub-optimal for Arabic.

In this paper, the first advanced language generation models built from the grounds up on Arabic language have been developed. The process of pre-training AraGPT2, a GPT-2 transformer model for the Arabic language is described. The model comes in 4 size variants: **base** (135M[2]), **medium** (370M), **large** (792M) and **mega** (1.46B[3]), which allows the exploration of AraGPT2 in multiple applications with different data availability and computational constraints. The perplexity measure is used to automatically evaluate AraGPT2. Furthermore, a human-based evaluation is provided, which highlights the ability of AraGPT2 to deceive human evaluators. Finally, an AraELECTRA (Antoun et al., 2020b) based detector is devel-

[1]Pretrained variants of AraGPT2 (base, medium, large, mega) and discriminator are publicly available on github.com/aub-mind/arabert/tree/master/aragpt2

[2]Million Parameters

[3]Billion Parameters

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 196–207
Kyiv, Ukraine (Virtual), April 19, 2021.

oped and released. It is able to consistently identify news articles written by AraGPT2. Making such powerful models publicly available to the Arabic research community enables research in rising Arabic NLP fields i.e Conversational Agents (Naous et al., 2020), Detection of Automatic News Generation Detection (Harrag et al., 2020)...

Our contributions can be summarized as follows:

- A methodology to pre-train a billion-size GPT2 model on a large-scale Arabic corpus.
- An automatic discriminator that achieves a 98% accuracy in detecting model-generated synthetic text.
- The four variants of AraGPT2 are released on popular NLP libraries, along with the automatic AraGPT2 discriminator.

The rest of the paper is structured as follows. Section 2 provides a concise review of previous literature on Arabic language modeling. Section 3 details the methodology used in developing AraGPT2. Section 4 describes the experimental setup, evaluation procedures and results. In addition, the approach to build a machine-generated text discriminator is presented in Section 5. Finally, a conclusion of the work and implications are mentioned in Section 6.

2 Related Works

2.1 English and Non-Arabic Language modeling

GPT-1 (Radford et al., 2018) showed that Causal Language Modeling[4] is an effective pre-training technique that improves a model's generalization capabilities. GPT-2 then showed that using a larger model trained on a larger dataset surpasses the state-of-the-art of many tasks in a zero-shot setting, where a model solves a task without receiving any training on that task. Taking the scaling approach to the extreme led to the creation of GPT-3 (Brown et al., 2020), with 175 billion parameter model, also trained with CLM using terabytes of internet text. GPT-3 explored the idea of few-shot learning, where a model is given examples from a new task as a text prompt, which unlocks new capabilities at test time. It was later shown that a carefully designed GPT-3 prompt allows the model to generate website designs, scramble/unscramble words...

The advantage of scaling model sizes and training datasets comes with drawbacks, particularly the high computational cost, in addition to the huge corpora required for pre-training. It was estimated that training GPT-2 and GPT-3 costs $43K and $4.6M respectively, without any hyper-parameter tuning. These drawbacks restricted the availability of large pre-trained models to English mainly and a handful of other languages i.e. ruGPT3[5] for Russian, and Chinese 1.5B GPT2 (Zhang, 2019).

2.2 Arabic Language modeling

Work on Arabic causal language modeling has been mostly limited to automatic speech recognition (ASR) systems. Since the language modeling component in ASR systems is a key module that ensures that the output text adheres with the statistical structure of language. Work on Arabic language models in ASR systems has mostly relied on N-grams language models. (Ali et al., 2014) built an N-grams language model (LM) using GALE training data transcripts of 1.4M words. More recent work in Arabic ASR implemented a recurrent neural network as an LM, using 130M tokens, and achieved a perplexity of 481 compared to 436 for a 4-gram LM (Khurana et al., 2019). Hamed et al. (2017) developed a code-switched Arabic-English language model using tri-gram LM and provided performance superior compared to two separate monolingual LMs. The code-switched LM was trained on 2.3M sentences or 13M words and achieved a perplexity of 275.

With the rising popularity of transfer learning in NLP, Arabic CLM was used as a pre-training task for an Arabic universal LM, hULMonA (ElJundi et al., 2019). The model was then fine-tuned on different downstream text classification tasks. hULMonA is a 3 stack of AWD-LSTM[6] layers (Howard and Ruder, 2018), trained on 600K Wikipedia article pre-segmented using the MADAMIRA Arabic morphological analyzer and disambiguator (Pasha et al., 2014).

Masked Language Modeling (MLM) has been useful as a pre-training task for several Arabic NLU models. Masked Language Modeling (MLM) is a slightly different objective than CLM that requires a system to predict a masked word within a sequence compared to CLM which predicts the missing word at the end of a sequence. MLM

[4]This is the regular Language Modeling objective where the model learns the probability of a word given the previous context. The CLM acronym is used to distinguish from masked language modeling (MLM).

[5]https://github.com/sberbank-ai/ru-gpts/

[6]ASGD Weight-Dropped LSTM

was used in models such as AraBERT (Antoun et al., 2020a), Arabic-BERT (Safaya et al., 2020), Arabic-ALBERT[7], GigaBERT (Lan et al., 2020), MarBERT (Abdul-Mageed et al., 2020), and QARiB (Chowdhury et al., 2020). Only two works have attempted to create an Arabic transformer causal language model. Khooli (2020) and Doiron (2020) finetuned the OpenAI GPT2-base model on Arabic Wikipedia, which was mainly trained on English text. Doiron (2020) also continued training on a collection of dialectal Arabic datasets, in order to create a dialectal Arabic GPT2. While this approach has shown the capability to generate Arabic text, it is sub-optimal for Arabic and is useful in cases where the training data is scarce.

Our proposed model is hence, the first Arabic transformer-based causal language model trained from scratch on the largest Arabic corpora available at the time of writing.

3 AraGPT2: Methodology

AraGPT2 is a stacked transformer-decoder model trained using the causal language modeling objective. The model is trained on 77GB of Arabic text. AraGPT2 comes in four variants as detailed in Table 1, with the smallest model, **base**, having the same size as AraBERT-base which makes it accessible for the larger part of researchers. Larger model variants (**medium**, **large**, **xlarge**) offer improved performance but are harder to fine-tune and computationally more expensive. The AraGPT2-detector is based on the pre-trained AraELECTRA model fine-tuned on the synthetically generated dataset. More details on the training procedure and dataset are provided in the following sections.

3.1 Model

AraGPT2 closely follows GPT2's variant architectures and training procedure. Table 1 shows the different model sizes, number of heads, number of layers, parameter count, and optimizer used for each model variant. All models are trained with context sizes of 1024 tokens. The LAMB (You et al., 2019) optimizer is used in the **base** and **medium** models only, since it allows using large batch sizes without worrying about training divergence. Using LAMB and Adam (Kingma and Ba, 2014) to train the **large** and **mega** variants isn't possible on TPUv3 due to the optimizer's high memory requirements, since memory cost scales linearly with the number of parameters. The limitations were overcome by following the training procedure of the GROVER model (Zellers et al., 2019) by using the Adafactor optimizer (Shazeer and Stern, 2018), which reduces memory requirements by factoring the second-order momentum parameters into a tensor product of two vectors. The GROVER architecture was also used instead of GPT2's, in which the layer normalization order in the transformer block is changed.

3.2 Dataset

The training dataset is a collection of the publicly available Arabic corpora listed below:

- The unshuffled OSCAR corpus (Ortiz Suárez et al., 2020).
- The Arabic Wikipedia dump from September 2020.
- The 1.5B words Arabic Corpus (El-Khair, 2016).
- The OSIAN corpus (Zeroual et al., 2019).
- News articles provided by As-safir newspaper.

Preprocessing First, the corpus was filtered by removing short documents with less than 3 sentences, and documents with more than 20% repeated sentences. URLs, emails, and user mentions were also replaced with special tokens. All diacritics, and elongations were removed as well, while punctuation and non-alphabetic characters were padded with white-spaces. Moreover, the '<|endoftext|>' token is appended at the end of each document. The total dataset size is 77GB with 8.8B words[8]. The majority of the training data is comprised of Arabic news article, which is mostly written in MSA. The corpus also contains a small set of English words i.e. named entities, which are kept without lower-casing. Subsequently, a Byte-level byte-pair-encoding (BPE) tokenizer is trained with 64000 vocabulary size on all of our preprocessed dataset, using the optimized BPE implementation from the HuggingFace library (Wolf et al., 2020). Finally, the BPE encoding is applied on the preprocessed dataset, which results in a total of 9.7M training examples with 1024 sub-word tokens each.

[7] https://github.com/KUIS-AI-Lab/Arabic-ALBERT/

[8] Word count was done after preprocessing, where white space is inserted before and after punctuations, brackets, numbers... which increased the total word count

Model	Size	Architecture	Context Size	Emb. Size	Heads	Layers	Optimizer
Base	135M	GPT2	1024	768	12	12	LAMB
Medium	370M	GPT2	1024	1024	16	24	LAMB
Large	792M	GROVER	1024	1280	20	36	Adafactor
Mega	1.46B	GROVER	1024	1536	24	48	Adafactor

Table 1: AraGPT2 model variants with sizes, architecture and optimizer

Model	Batch Size	Learning Rate	Steps	Time (days)	PPL
Base	1792	1.27e-3	120K	1.5	55.8
Medium*	80	3e-4	1M	23	45.7
Large	256	1e-4	220K	3	36.6
Mega	256	1e-4	780K	9	**29.8**

Table 2: AraGPT2 training details and validation perplexity. ***Medium** was trained on a TPUv3-8 with a small batch size, since the model was not converging with a large batch size

4 Experiments and Evaluation

4.1 Pre-training Setup

All models were trained on a TPUv3-128 slice[9] with different batch sizes and the total number of steps as shown in Table 2. **Base** and **mega** were trained for approximately 20 epochs, while **medium** and **large** were trained for 10 and 6 epochs respectively, due to TPU access limitations.

4.2 Numerical Evaluation

For the validation dataset, the Arabic Wikipedia articles that were published after August 2020 were used, since older articles were included in the September Wikipedia dump. The perplexity score was selected as a numerical evaluation metric since it measures the degree of 'uncertainty' a model has assigning probabilities to the test text. Table 2 shows that, unsurprisingly, validation perplexity keeps improving with larger model sizes. In fact, the model is still under-fitting the validation set from Wikipedia. The generation capabilities of the different variants of AraGPT2 is illustrated through the selected examples in Appendix A.

4.3 Zero-Shot Evaluation

During zero-shot task evaluation, the model is only given a natural language instruction to motivate and ground the task, without any back-propagation happening. The task of searching and finding the best input prompt, also known as "prompt engineering", is hard. Since the search space is practically infinite, and the performance is highly sensitive to changes in the prompt. The zero-shot performance of AraGPT2-Mega is evaluated on two tasks, question answering, and translation. AraGPT2-mega correctly answers 25% of the trivia questions but fails in English-to-Arabic translation. Details on the datasets, prompts, and evaluation are presented in Appendix B.

4.4 Evaluating the Human Ability to Detect Machine-Generated Text

The gold standard for evaluating a model's language generation capability is human evaluation. We presented 74 Arabic-speaking subjects from various social media with a survey designed to test the average-human ability to distinguish between machine-generated and human-written text and thus testing the model's ability to deceive a human subject. The survey had a total of 8 news articles, 4 machine-generated using AraGPT2-Mega and 4 written by humans. Each category was split into long and short text, which allows us to test the long-term generation coherency. In addition, the human evaluators are allowed to add justification for each answer.

The survey results, Figure 1, show that AraGPT2-Mega successfully fooled approx. 60% of the respondents, with longer passages having a higher error rate than short passages. In the provided explanations, some subjects relied on punctuation mistakes, coherence, and repetition issues, while others spotted factual inaccuracies. However, the results also show that humans were misclassifying human-written 50% the time (chance level performance), while also citing factual inconsistencies, grammatical errors, and unusual writing styles[10].

These surprising results show that AraGPT2 can accurately generate human-like text while

[9] TPUv3-128 has a total of 2TB of HBM memory with 16GB per core. TPUs were freely provided by the TFRC program.

[10] Survey results are available on our GitHub repository.

Figure 1: Survey results showing human error rates on machine generated *(left)* and human written text *(right)*

maintaining grammatical correctness that can fool the average reader. It should be noted that there exist some tools, i.e. the Giant Language model Test Room (GLTR) (Gehrmann et al., 2019), that allows humans to study the statistical distributional differences in text generated by GPT2-based models and human-written text. Figure 5 in Appendix C displays a visualization of token-level information created by GLTR with text generated by AraGPT2 and on human-written articles.

5 Automatic Detection of Machine Generated Text

Large language models could have a significant societal impact if used for malicious purposes, such as automating the generation of misleading news articles, fake reviews, or high-quality phishing messages. The survey in Section 4.4, showcases the failure of the average-human to consistently detect machine-generated text, which motivates the problem of automatic detection of AraGPT2-generated text. Related work on the detection of machine-generated text by Jawahar et al. (2020) indicates that automatic detectors like the GROVER-detector (Zellers et al., 2019) and the RoBERTA-detector (Solaiman et al., 2019) have better success than human evaluators. In addition, previous work on detecting Arabic GPT2 (Khooli, 2020) auto-generated tweets, achieved 98.7% accuracy, by fine-tuning an AraBERTv0.1 (Antoun et al., 2020a) based classifier (Harrag et al., 2020).

Our detector is based on the pre-trained AraELECTRA (Antoun et al., 2020b) model, which we fine-tuned on a dataset created by combining 1500 human-written news articles, with 1500 articles generated by AraGPT2-Mega. For article generation, we only provided the model with a short prompt of 25 words. We created two versions of the dataset, one with short texts (150 tokens) and one with long texts (500 tokens), in order to evaluate the impact of the text's length.

Fine-tuned AraELECTRA achieves 98.7% and 94.9% F1-score on long and short text respectively[11], which indicates that longer text is easier to detect than short text. The high scores achieved by AraELECTRA can be explained by the fact that machine-generated text tends to be more predictable compared to human-written text (see Appendix C, Fig. 5). The difference in text predictability can be easily exploited by a language model to detect machine-generated text. Another contributing factor is that AraELECTRA was pre-trained on the exact same dataset as AraGPT2.

6 Conclusion

AraGPT2 is the first advanced Arabic language generation model based on the transformer architecture. The model was trained on the largest publicly available collection of filtered Arabic corpora. The model was evaluated using the perplexity measure which measures how well a probability model predicts a sample. Results show that AraGPT2 is able to produce high quality Arabic text that is coherent, grammatically correct and syntactically sound.

It is important to note that AraGPT2, like many ML models, has ethical implications and can be used maliciously i.e. automatic fake news generation, modeling the dataset inherent biases... To help detect misuse of the model, a detector model that is tasked to detect output generated by AraGPT2 is also released. More importantly, our hopes that publicly releasing AraGPT2 will open up doors for new research possibilities for the Arabic NLP community.

Acknowledgments

This research was supported by the University Research Board (URB) at the American University of Beirut (AUB), and by the TFRC program for providing free access to cloud TPUs. Many thanks to As-Safir newspaper for the data access, and also thanks to Nick Doiron for the insightful discussions.

[11] The trained model will be publicly available in our repository

References

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. *arXiv preprint arXiv:2010.04900*.

Ahmed Ali, Yifan Zhang, Patrick Cardinal, Najim Dahak, Stephan Vogel, and James Glass. 2014. A complete kaldi recipe for building arabic speech recognition systems. In *2014 IEEE spoken language technology workshop (SLT)*, pages 525–529. IEEE.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020a. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. Araelectra: Pre-training text discriminators for arabic language understanding.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Shammur Absar Chowdhury, Ahmed Abdelali, Kareem Darwish, Jung Soon-Gyo, Joni Salminen, and Bernard J. Jansen. 2020. Improving Arabic text categorization using transformer training diversification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 226–236, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020b. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Nick Doiron. 2020. Making a mini gpt-2 with dialect prompts.

Ibrahim Abu El-Khair. 2016. 1.5 billion words arabic corpus. *arXiv preprint arXiv:1611.04033*.

Obeida ElJundi, Wissam Antoun, Nour El Droubi, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. 2019. hulmona: The universal language model in arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 68–77.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Injy Hamed, Mohamed Elmahdy, and Slim Abdennadher. 2017. Building a first language model for code-switch arabic-english. *Procedia Computer Science*, 117:208–216.

Fouzi Harrag, Maria Dabbah, Kareem Darwish, and Ahmed Abdelali. 2020. Bert transformer model for detecting Arabic GPT2 auto-generated tweets. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 207–214, Barcelona, Spain (Online). Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Abed Khooli. 2020. gpt2-small-arabic.

Sameer Khurana, Ahmed Ali, and James Glass. 2019. Darts: Dialectal arabic transcription system. *arXiv preprint arXiv:1909.12163*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. Gigabert: Zero-shot transfer learning from english to arabic. In *Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

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

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.

Tarek Naous, Christian Hokayem, and Hazem Hajj. 2020. Empathy-driven Arabic conversational chatbot. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 58–68, Barcelona, Spain (Online). Association for Computational Linguistics.

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*, pages 1703–1714, Online. Association for Computational Linguistics.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv preprint arXiv:1804.04235*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Jörgrg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in neural information processing systems*, pages 9054–9065.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. Osian: Open source international arabic news corpus-preparation and integration into the clarin-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182.

Zhibo Zhang. 2019. Gpt2-ml: Gpt-2 for multiple languages. `https://github.com/imcaspar/gpt2-ml`.

A Generated Samples from ARAGPT2

Prompt
أحدث حصول لقاح شركة " أوكسفورد " البريطانية على الموافقة في موطنه ، حالة من الارتياح في العالم ، نظرا إلى فعاليته في وقاية الجسم من وباء كورونا ، فضلا عن سعره المناسب وسهولة تخزينه اللافتة مقارنة بالتطعيمات الأخرى المتاحة في السوق الدولية . وبحسب شبكة " سكاي نيوز " البريطانية فإن هذا الموافقة على هذا اللقاح تعني الشيء الكثير للعالم وليست مجرد خبر عادي .
Generated Text 1
وقالت الشبكة إن هذه هي المرة الأولى التي يحصل فيها لقاح مضاد لكورونا على موافقة منظمة الصحة العالمية ، كما أنها المرة الأولى التي تحصل فيها شركة بريطانية على مثل هذه الموافقة منذ أكثر من 20 عاما . وأضافت أن الشركة حصلت أيضا على موافقة إدارة الغذاء والدواء الأميركية (FDA) لقاحها المضاد لفيروس زيكا الذي تم تطويره بالتعاون مع شركة " غلاكسو سميثكلاين " للأدوية وشركة " سانوفي أفنتيس " الفرنسية للصناعات الدوائية . وأشارت إلى أنه لم يتم حتى الآن الإعلان عن أي حالات إصابة بكورونا بين البشر في الولايات المتحدة أو غيرها من دول العالم . ونقلت الشبكة عن المدير التنفيذي لشركة " جلاكسو سميث كلاين " قوله : " نحن سعداء للغاية بحصولنا على هذه الموافقة لأن ذلك يعني أننا تمكنا من تحقيق هدفنا المتمثل في حماية أكبر عدد ممكن من الناس من الإصابة بفيروس كورونا " . وأضاف : " نأمل أن نتمكن من إنتاج المزيد من اللقاحات المضادة لهذا الفيروس وغيره من الأمراض المعدية المنتشرة في جميع أنحاء العالم خلال السنوات القليلة المقبلة " يذكر أن فيروس كورونا المسبب لمتلازمة الشرق الأوسط التنفسية هو أحد الفيروسات التي تصيب الجهاز التنفسي ، ولا توجد حتى الآن على مستوى العالم معلومات دقيقة عن مصدر هذا الفيروس ولا طرق انتقاله ، كما لا يوجد تطعيم وقائي أو مضاد حيوي لعلاجه . لكن مراكز السيطرة على الأمراض والوقاية منها بالولايات المتحدة الأميركية كانت قد أعلنت في وقت سابق من الشهر الجاري أن لقاحا تجريبيا أنتجته شركة " نوفارتس " السويسرية أثبت فاعليته في الوقاية من مرض متلازمة الشرق الأوسط التنفسية (MERS - CoV) لدى الأطفال والبالغين الذين يعانون من أعراض شبيهة بأعراض الانفلونزا . وكانت منظمة الصحة العالمية قد أعلنت في شهر سبتمبر أيلول الماضي تسجيل أول حالة وفاة ناجمة عن الإصابة بفيروس كورونا في المملكة العربية السعودية ، حيث توفي رجل يبلغ من العمر 69 عاما كان يعاني من عدة أمراض مزمنة جراء إصابته بهذا الفيروس . وكان الرجل قد نقل إلى مستشفى الملك فيصل التخصصي ومركز الأبحاث في مدينة الرياض بعد شعوره بأعراض تنفسية حادة أدت إلى دخوله في غيبوبة وتوفي بعد يومين من إدخاله المستشفى . وقال الدكتور علاء العلوان المدير العام للمكتب التنفيذي لمجلس وزراء الصحة لدول مجلس التعاون ورئيس اللجنة الخليجية لمكافحة الأمراض المعدية إنه بناء على ما أعلنته منظمة الصحة العالمية فقد تمت الموافقة على طلب وزارة الصحة بالمملكة العربية السعودية لتزويدها باللقاح الواقي
Generated Text 2
وقالت الشبكة إن هذه هي المرة الأولى التي يحصل فيها لقاح مضاد لفيروس كوفيد - 19 على موافقة منظمة الصحة العالمية ، ما يعني أنه سيكون متاحا للجميع خلال الأسابيع القليلة المقبلة .وأضافت أن هذا اللقاح الذي يحمل اسم " تاميفلو " (Tamiflu) تم تطويره بالتعاون مع شركة " غلاكسو سميث كلاين " (GlaxoSmithKline) وشركة " سانوفي باستور " (Sanofi Pasteur) الفرنسية ، وهما شركتان تابعتان لمجموعة " أسترا زينيكا " (AstraZeneca) الدوائية .وقال رئيس قسم الأمراض المعدية في منظمة الصحة العالمية ديفيد نابارو : " نحن مسرورون للغاية بحصول هذا اللقاح على الموافقة النهائية من قبل إدارة الأغذية والعقاقير الأمريكية (FDA) ومنظمة الصحة العالمية (WHO) " . وأضاف : " إنه نبأ سار لجميع العاملين في مجال الرعاية الصحية والمرضى الذين سيستفيدون من هذا اللقاح " .من جهته ، قال الرئيس التنفيذي لشركة " غلاكسوسميث كلاين " ، بروس أرمسترونغ : " إننا سعداء جدا بهذا الخبر السار ، ونحن واثقون من أننا سنتمكن من إنتاج كميات كبيرة من هذا اللقاح لتلبية الطلب العالمي المتزايد عليه " .وتابع قائلا : " لقد عملنا بشكل وثيق مع وزارة الصحة العامة في المملكة المتحدة لضمان الحصول على الموافقات التنظيمية اللازمة لهذا اللقاح ، ونأمل أن نتمكن من إنتاجه بكميات كافية لتلبية احتياجات المرضى في جميع أنحاء العالم " .يشار إلى أن فيروس كورونا المسبب لمتلازمة الشرق الأوسط التنفسية (MERS - CoV) هو أحد الفيروسات التي تصيب الجهاز التنفسي ، ولا توجد حتى الآن على مستوى العالم معلومات دقيقة عن مصدر هذا الفيروس ولا طرق انتقاله ، كما لا يوجد تطعيم وقائي أو مضاد حيوي لعلاجه . لكن هناك العديد من الإجراءات والاحتياطات التي يمكن اتخاذها للحد من خطر الإصابة به ، مثل الإقلال من مخالطة المرضى وغسل اليدين بانتظام وتجنب لمس العين والأنف والفم واستخدام المناديل عند السعال أو العطس وتغطية الفم والأنف أثناء السعال أو العطاس وتغطية الأنف والفم عند العطس أو السعال باستخدام منديل والتخلص منه مباشرة في سلة النفايات وعدم مشاركة الآخرين في أدوات الطعام والشراب غير المستخدمة بعد غسلها جيدا والحرص على غسل اليدين بالصابون والماء الدافئ لمدة 20 ثانية على الأقل ثم شطفها بالماء البارد وتجفيف اليدين بمنشفة نظيفة والابتعاد عن أماكن الازدحام والتجمعات البشرية قدر الإمكان .يذكر أن منظمة الصحة العالمية كانت قد أعلنت الأسبوع الماضي أن عدد حالات الإصابة المؤكدة بفيروس كورونا الجديد ارتفع إلى 991 حالة ، توفي منها 563 شخصا منذ اكتشاف المرض في شهر سبتمبر أيلول 2012 .

Figure 2: Random unseen context about coronavirus vaccine(top). Followed by two generated samples bu ARAGPT2-mega. Generated text 1 ($top_p = 0.95$), Generated text 2 ($top_p = 1$)

Prompt
يحكى أن مزارعا مخادعا قام ببيع بئر الماء الموجود في أرضه لجاره مقابل مبلغ كبير من المال ، وعندما جاء المزارع الذي اشترى البئر ليستخدم الماء الموجود فيه في اليوم التالي قال له الرجل المخادع : اذهب من هنا أيها الرجل فأنا قد بعتك البئر لكنني لم أبعك الماء الموجود فيه ، دهش الرجل مما سمع وتوجه إلى القاضي ليشتكي المزارع المخادع له بعد محاولات عديدة لإقناعه بأن البئر والماء الذي فيه من حقه
Generated Text 1
وأن عليه أن يدفع المبلغ الذي اشتراه به ثمنا للبئر ، لكن القاضي رفض الشكوى وقال للرجل المخادع : لقد اشتريت منك البئر لكنك لم تشتر مني الماء .وفي يوم من الأيام وبينما كان أحد المزارعين يجلس على إحدى شرفات منزله المطلة على حديقة مليئة بالأشجار رأى رجلا قادما نحوه وبيده كيس مليء بالحلوى فسلم عليه وسأله عن سبب مجيئه إليه فأخبره الرجل بأنه يريد أن يعطيه بعض الحلوى التي يحبها ، فقال له المزارع : حسنا سأعطيك ما تريد بشرط واحد وهو أن لا تخبر أحدا بهذا الأمر أبدا ، وافق الرجل على شرط المزارع وأعطاه الكيس المملوء بالحلوى ، وقبل أن يغادر المزرعة سأله المزارع قائلا : هل تعلم يا سيدي بأنني رجل فقير ولا أملك مالا لأحصل على الحلوى ؟ فأجابه الرجل : نعم أعلم ذلك ولكن المشكلة هي أنني لا أستطيع أن أعطيك الحلوى دون أن تخبرني بذلك ، ضحك المزارع قليلا ثم قال للرجل : إذا أخبرتني قبل أن تعطيني الحلوى فلن أسمح لك بأخذ أي شيء منها حتى لو كان هذا الشيء هو ثمن الحلوى التي أعطيتك إياها .أخبر المزارع جاره بما حدث معه وطلب منه أن يوصله إلى القاضي لعله يجد حلا لمشكلته ، ذهب جار المزارع إلى القاضي وأخبره بما حدث مع جاره ، فسأله القاضي مستغربا : لماذا فعلت ذلك ؟ أجابه الجار : لأنني أريد أن أعرف كيف حصلت على الحلوى دون أن تأخذها مني ودون أن تعطيها لأحد غيرك !وبالفعل عندما وصل الجار إلى القاضي طلب منه القاضي أن يريه الصندوق الذي يحتوي على الحلوى ، نظر القاضي إلى الصندوق فوجد بداخله صندوقا آخر كبيرا مليئا بالحلوى ، تعجب القاضي من محتويات الصندوق لكنه سرعان ما قرر أن يفتح الصندوق ليرى ماذا يوجد بداخله ، فتح القاضي الصندوق فوجده مملوءا بالحلوى وقد كتب عليه عبارة تقول : " هذه الحلوى لمن يستطيع الحصول عليها فقط " ، استغرب القاضي من محتوى الصندوق وقرر أن يعيد فتحه مرة أخرى ليجد بداخله صندوقا ثالثا مليئا بالحلوى أيضا ، سأل القاضي صاحب الصندوقين عما يحتويه كل صندوق فأجاب صاحب الصندوق الثالث قائلا : إن هذين الصندوقين مخصصان للفقراء الذين لا يستطيعون الحصول على الحلوى ، أما الصندوقين الآخرين فهما مخصصان للأغنياء الذين يستطيعون الحصول على تلك الحلوى .

Prompt
يحكى أن الأسد ملك الغابة وأقوى سكانها كان ذات يوم نائما ، عندما بدأ فأر صغير يعيش في الغابة نفسها بالركض حوله والقفز فوقه وإصدار أصوات مزعجة ، مما أقلق نوم الأسد ودفعه للاستيقاظ ، وعندما قام الأسد من نومه كان غاضبا ، فوضع قبضته الضخمة فوق الفأر ، وزمجر وفتح فمه ينوي ابتلاع الفأر الصغير بلقمة واحدة ، صاح الفأر عندها بصوت يرتجف من الخوف راجيا أن يعفو الأسد عنه ، وقال :
Generated Text 1
" سامحني يا ملك الغابة ، أرجوك سامحني ، فأنا لا أستطيع العيش بدونك " .أجاب الأسد قائلا : " أنا ملك الغابة ، وأنت أيها الفأر الصغير مجرد تابع لي ، ولا تستطيع عمل أي شيء دون إذني ، أما أنا فأستطيع أن أعمل ما أشاء وأعيش كما أشاء ، ولكن بشرط واحد وهو أن تساعدني على القضاء على جميع الفئران التي تعيش في الغابة " .فقال الفأر : " نعم سيدي ، سأفعل كل ما تأمرني به ، وسأقضي على جميع الفئران الموجودة في الغابة ، وسآتي بك إلى هنا لتعيش معي وتاكل من طعامي وتشرب من مائي وتنام في فراشي " .ضحك الأسد ضحكة عالية ثم قال للفأر : " هل أنت موافق ؟ " ، فقال الفأر : " لقد سمعت هذا الكلام منذ قليل فقط ، ولكني لم أكن أعرف أنك ملك الغابة وسيدها ، وأن لك سلطانا علي الحيوانات كلها ، لذلك أرجو منك أن تسمح لي بأن أعيش معك تحت سقف واحد وتحت حماية سلطانك وملكك " .فرح الأسد كثيرا بالفكرة ، وطلب من الفأر أن يذهب معه إلى بيته ليعيشا معا حياة سعيدة هانئة ، وبالفعل ذهب الفأر مع الأسد إلى بيت الأسد الذي رحب به أجمل ترحيب ، وأعطاه قطعة كبيرة من الخشب ليصنع منها سريرا كبيرا له ولزوجته وأولاده ، وهكذا عاش الفأر مع الأسد في سعادة وهناء .وبعد فترة قصيرة من الزمن ، مرض الأسد مرضا شديدا ، ولم يعد قادرا على تحمل المزيد من الأعباء الملقاة على عاتقه ، فطلب من الفأر أن يساعده في حمل سريره الثقيل والذهاب به إلى الطبيب ، فوافق الفأر على الفور ، وذهب مسرعا إلى الطبيب ليخبره بأنه مصاب بمرض خطير يحتاج إلى جراحة عاجلة ، لكنه رفض إجراء الجراحة خوفا من غضب الأسد عليه وانتقامه منه ، لكن الفأر أصر على الذهاب إلى الطبيب لإجراء الجراحة ، وفي أثناء ذلك سمع صوت الأسد يناديه ويقول له : " ماذا تريد مني يا ملك الغابة ؟ " ، فأجابه الفأر : " أريد أن أحملك على ظهري لأذهب بك إلى الطبيب حتى يشخص حالتك ويصف لك العلاج المناسب لحالتك " .نظر الأسد إلى الفأر مبتسما وقال له : " هيا بنا يا صديقي العزيز لنذهب

Figure 3: Random unseen contexts about children stories. Followed by a generated sample by AraGPT2-mega with $top_p = 0.95$

Prompt 1 (Cooking Recipes)
فتة الدجاج اللبنانية طريقة التحضير نضع الدجاج في قدر من الماء مع البصل والهيل وورق الغار والملح والفلفل الأسود ونسلقه على نار متوسطة حتى ينضج تماما . صفي الدجاج من المرق ونفتته ونضعه جانبا ثم نصفي المرق ونضعه في وعاء . نحمي الزيت ونقلي فيه الخبز حتى يصبح ذهبيا ومقرمشا .
Generated Text
نوزع شرائح الفليفلة الخضراء والحمراء والخضراء والبندورة المقشرة والمفرومة فوق الخبز ، ثم نرش فوقها القليل من البقدونس المفروم والصنوبر المحمص . ندهن صينية فرن بالقليل من الزيت ونرشها بالقليل من الطحين ثم ندخل الصينية إلى فرن محمى مسبقا على حرارة 180 درجة مئوية لمدة تتراوح بين 15 و 20 دقيقة أو حتى يحمر وجه الصينية . نخرج الصينية من الفرن ونتركها لتبرد قليلا ثم نقطعها ونقدمها ساخنة .

Prompt 2 (Short Conversation - Greeting)
مرحبا كيف حالك ؟
Generated Text
ان شاء الله بخير . اليوم جبتلكم لعبة حلوة كتير و هي عبارة عن مجموعة من الكرات الملونة بألوان مختلفة , كل ما عليك فعله هو الضغط على الكرة التي تريد أن ترميها في المكان المخصص لها ثم اضغط عليها مرة أخرى

Prompt 3 (Wikipedia-style)
لبنان تعد الجمهورية اللبنانية دولة عربية عاصمتها بيروت ، وتقع على الشاطئ المطل على الجزء الشرقي للبحر الأبيض المتوسط ، وتعد واحدة من أكثر البلدان المكتظة بالسكان ، وبالعودة إلى تاريخها فهي تعتبر موطنا لأقدم الحضارات البشرية في الألفية الثالثة قبل الميلاد خاصة المنطقة الساحلية منها ، فكانت من أهم المراكز الثقافية والتجارية في ذلك الوقت ، وكان لتضاريسها الوعرة دورا مهما على مر التاريخ في لجوء عدة جماعات إليها ، وأعلنت كدولة جمهورية عام 1926 م ، وفي عام 1946 م حصلت لبنان على استقلالها بعد الانتداب الفرنسي .
Generated Text
[١][٢]تعد اللغة العربية هي اللغة الرسمية للبلاد ، بالإضافة إلى وجود لغتين رسميتين أخريين هما الإنجليزية والفرنسية ، أما الديانة السائدة في البلاد فهي الإسلام ، حيث يشكل المسلمون حوالي 96 % من إجمالي عدد السكان ، ويتوزع الباقون على الطوائف المسيحية والديانات الأخرى ، ومن أبرز المعالم السياحية في لبنان مدينة بيروت التي تعد عاصمة البلاد ومركزها الثقافي والاقتصادي ، كما يوجد بها العديد من المواقع الأثرية الهامة مثل مغارة جعيتا ، وقلعة بعلبك ، وبرج إيفل ، ومتحف سرسق ، وقصر بيت الدين الذي يعد واحدا من أقدم القصور التاريخية في العالم ، وغيرها الكثير من المعالم المميزة . [٣]تعتمد السياحة في لبنان بشكل أساسي على الطبيعة الخلابة والمناخ المعتدل صيفا وشتاء ، لذلك فإن أغلب السياح يفضلون قضاء عطلتهم الصيفية في المناطق الجبلية والساحلية للاستمتاع بالمناظر الطبيعية الخلابة والمناظر المائية الرائعة ، بينما يفضل البعض الآخر التوجه إلى المدن والمناطق الداخلية للتمتع بمناخها المعتدل والاستجمام في أحضان الطبيعة الساحرة

Figure 4: Random unseen contexts on miscellaneous topics. Followed by a generated sample by ARAGPT2-mega with $top_p = 0.95$

B Zero-Shot Learning

B.1 Question Answering

In zero-shot factoid question answering, the information contained within the language model can be queried. The model is tested on the Arabic examples from the TyDiQA (Clark et al., 2020a) validation dataset (921 examples), and on the test set of ARCD (Mozannar et al., 2019) (702 examples). Hence, the model os provided with the following prompt: *"Answer the following question:"* - "أجب عن السؤال التالي:" , followed by the question, then the phrase *"The answer is"* - "الجواب هو:". It is also possible to append the phrase *"in the year"* - "في عام", if the expected answer is a year, as shown in Table 3.

Table 3: The input prompt for question answering

أجب عن السؤال التالي : متى عرضت أول حلقة من مسلسل بافي قاتلة مصاصي الدماء ؟ الجواب هو في عام
Answer the following question: When was the first episode of the series Buffy the Vampire Slayer shown? The answer is in the year

The answer length is set to be the same as the gold answer length, and a repetition penalty is applied as in CTRL (Keskar et al., 2019), which penalizes the probability scores of previously generated tokens. A 'no repeat tri-gram' strategy that inhibits the model from generating the same tri-gram more than once has also been employed. Note that the context passage is not provided, which forces the model to rely only on the information gained during pretraining.

The model achieves a 3.93% exact-match score and an F1-score of 14.51% on TyDiQA, and 4.07% exact-match score and 13.88% F1-score on ARCD. Since exact-match and F1-score misses answers that are correct but are worded differently (as shown in Table 4). A subset of 500 answers from the best TyDiQA run is selected, and scored manually. Manual scoring shows that AraGPT2 correctly answered 24.6% of the questions. The model was particularly good in countries and capitals question, year of birth and death, and some geography. Yet it was failing mostly on questions about quantities i.e. population counts, area, age... The predefined answer length negatively affected the generated answers in some cases, which is a limitation of the current approach.

Table 4: Examples of correct answers that have zero exact match score.

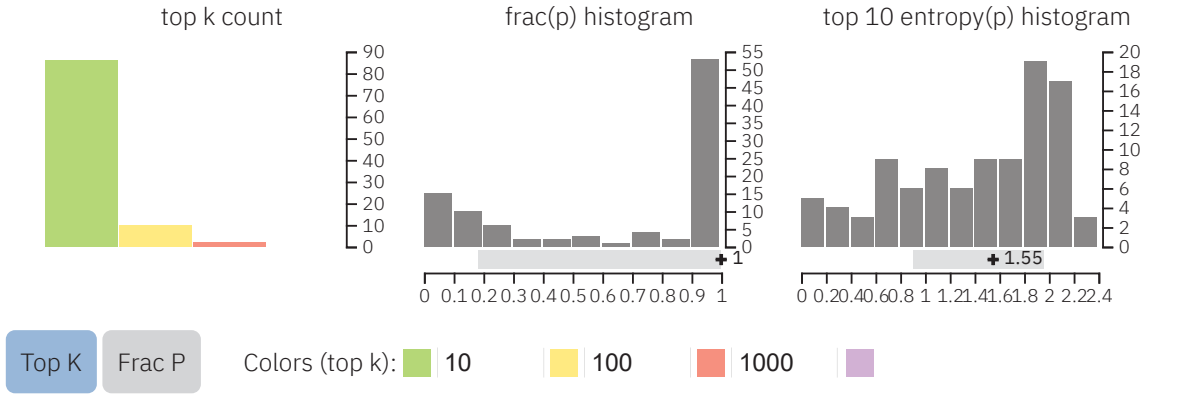
Question	من هو ألفرد نوبل ؟ *Who is Alfred Nobel?*
Predicted Answer	مخترع الديناميت ، ومخترع *Inventor of the dynamite, and the inventor of*
Ground Truth	مهندس ومخترع وكيميائي سويدي *An engineer and an inventor and a Swedish chemist*
Question	متى تاسس الاتحاد الدولي لكرة القدم ؟ *When was the FIFA founded?*
Predicted Answer	١٩٠٤ م . *1904 AD*
Ground Truth	٢١ مايو من العام ١٩٠٤ *21 May of the year 1904*
Question	من هو إدغار ديغا ؟ *Who is Edgar Degas?*
Predicted Answer	أنه فنان تشكيلي فرنسي ، ولد في باريس عام *He is a French visual artist, born in*
Ground Truth	فنان تشكيلي و رسام و نحات فرنسي *Visual artist and painter and sculptor*

B.2 Translation

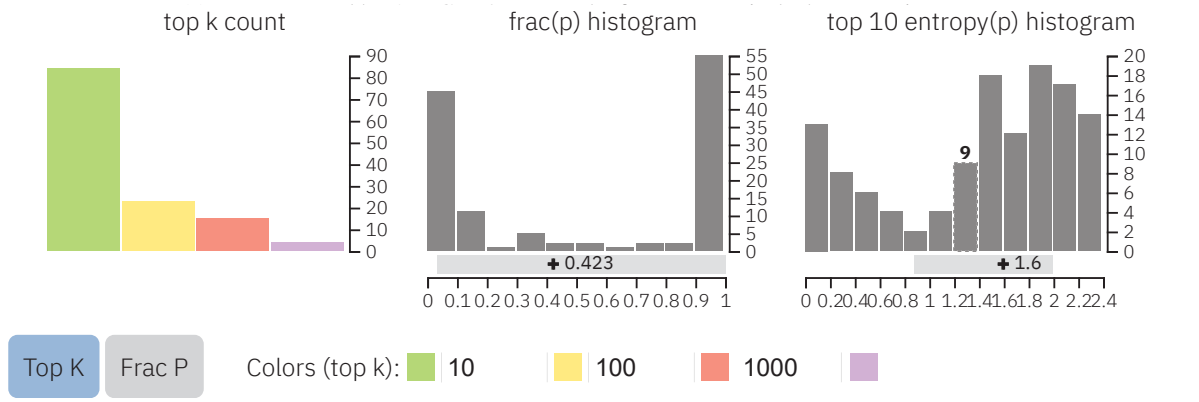
A experiments has also been conducted to test the translation capability of AraGPT2 by appending the prompt *"What is the translation of this sentence ?:"* - "ما هي ترجمة هذه الجملة ؟:" to the sentence from the source language, in order to induce the translation behavior of the model. We then apply greedy decoding to get the generated target sentence. Evaluation is performed on 5000 randomly selected pairs from the English-Arabic Tatoeba (Tiedemann, 2012) dataset. The model achieved only 1.32 BLEU score[12]. The low score is due to the scarce representation of English words in the vocabulary, since most words were split into single characters. Additionally, given that the prompt design greatly affects the model's zero-shot performance, our prompt design might have been suboptimal. Nevertheless, this negative result encourages research into prompt engineering for Arabic language models, which we leave as future work.

[12]Using the sacrebleu scorer (Post, 2018)

C GLTR Analysis and Visualizations



نفذ الشاب المقدسي محمد محمود السلايمة (23 عاما) من حي رأس العمود ، أمس ، عملية دهس في القدس المحتلة ، أصيب فيها شرطي إسرائيلي بجروح طفيفة . وقال المتحدث باسم الشرطة الإسرائيلية ميكي روزنفيلد إن سائق سيارة صدم بسيارته عددا من عناصر شرطة حرس الحدود الذين كانوا يقومون بدورية في شارع يافا في القدس الغربية ما أدى إلى إصابة أحدهم بجروح طفيفة نقل على أثرها إلى مستشفى هداسا عين كارم لتلقي العلاج . وذكرت وسائل الإعلام الإسرائيلية أن منفذ العملية هو شاب فلسطيني من سكان بلدة سلوان جنوب المسجد الأقصى .



واجه الأكراد سلسلة تحديات واختبارات في الفترة الماضية ، سواء مع الدولة أو المعارضة ، سياسية أو عسكرية ، غير أن الاختبار الأكثر جدية كان في مواجهتهم تنظيم الدولة الإسلامية في العراق والشام - داعش في عين العرب (كوباني) ، حيث تمكن مقاتلو وحدات حماية الشعب الكردية ، بالتعاون مع البشمركة وطيران التحالف الدولي ، من كسر تقدم داعش وإبعاده عن المنطقة ، بالتزامن مع تحصين مواقعهم القريبة ، سواء في القامشلي أو عامودا وصولا إلى رأس العين . أما عفرين فكانت محطة انطلاق العديد من المقاتلين الأكراد للتصدي ?? جبهة النصرة ، حتى أحياء في حلب ، مثل الشيخ مقصود والأشرفية ، ولعبوا دورا مهما في تخفيف الحصار عن نبل والزهراء .

(b) Human-Written Text

Figure 5: It is clear that the machine generated text in (a) is mostly green and yellow highlighted, while in the human-written text, (b), an increase in red and purple highlighted words can be noticed. P.S.: We use AraGPT2-base as the backend model in GLTR