

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, pretrained 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 is 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 large Arabic corpora 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. We evaluate different size variants of ARAGPT2 using the perplexity measure, where ARAGPT2-MEGA achieves a perplexity of 29.8 on held-out articles from Wikipedia. Pretrained variants of ARAGPT2 (base, medium, large, mega) are publicly available on github.com/aub-mind/arabert/aragpt2 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., 2020), 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).

GPT-1 (Radford et al., 2018) showed that Causal Language Modeling¹ is an effective pretraining 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), a 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.

Work on Arabic language modeling have 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., 2020). In contrast, Arabic text generation or causal language modeling haven’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 learn-

¹This is the regular Language Modeling objective where the model learns the probability of a word given previous context. We make use of the CLM acronym to distinguish from masked language modeling (MLM).

ing 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, we develop the first advanced language generation models built from the grounds up on Arabic language. We describe the process of pretraining a GPT-2 transformer model for the Arabic language, and which we name ARAGPT2. The model comes in 4 size variants: **base** (135M²), **medium** (370M), **large** (792M) and **mega** (1.46B³), which allows the exploration of ARAGPT2 in multiple applications with different data availability and computational constraints. We use the perplexity measure as a way to evaluate ARAGPT2. We make the models available on github.com/aub-mind/arabert/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 pretrain a billion-size GPT2 model on a large-scale Arabic corpus.
- Publicly releasing four variants of ARAGPT2 on popular NLP libraries.

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, and evaluation procedures and results. We finally conclude in Section 5.

2 Related Works

Work on Arabic causal language modeling has been mostly limited to automatic speech recognition (ASR) systems. Since the language modeling component in an ASR systems is a key module which ensures that the output text adheres with the statistical structure of language. Work on Arabic language models in ASR systems have 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 Recurrent neural network as a 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 pretraining 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-LSTM⁴ 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 pretraining 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 predict the missing word at the end of a sequence. MLM was used in models such as ARABERT (Antoun et al., 2020), Arabic-BERT (Safaya et al., 2020), Arabic-ALBERT⁵, GigaBERT (Lan et al., 2020), MarBERT (Abdul-Mageed et al., 2020), and QARiB (Chowdhury et al., 2020). Only two works have attempted to create a 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.

²Million Parameters

³Billion Parameters

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@70K	1792	1.27e-3	70K	18H	59.6
Base	1792	1.27e-3	120K	1.5	55.8
Medium	1152	1e-3	85K	1.5	49.9
Large	256	1e-4	220K	3	36.6
Mega@450K	256	1e-4	450K	5	33.7
Mega	256	1e-4	780K	9	29.8

Table 2: ARAGPT2 training details and validation perplexity. **Base@70K** and **Mega@450K** models were trained for approx. half the total steps

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. 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. We overcame this limitations by following the training pro-

cedure of the GROVER model, where we make use of 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. We also make use of the GROVER architecture 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). We apply thorough filtering to the corpus to remove low quality, and short articles, in addition to removing text with inappropriate nature.
- 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 Assafir newspaper.

The total dataset size is 77GB with 8.8B words⁶. 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 English words i.e. named entities, which we keep without lower-casing.

⁴ASGD Weight-Dropped LSTM

⁵<https://github.com/KUIS-AI-Lab/Arabic-ALBERT/>

⁶word count was done after preprocessing, where a white space is inserted before and after punctuations, brackets, numbers... which increased the total word count

Preprocessing We first filter inappropriate text (short articles, articles with repeated sentences, pronographic content). Then replace links, emails and user mentions with special tokens. We remove diacritics, elongations and insert white-spaces before and after each punctuation and special characters. We also append the '`<|endoftext|>`' token to the end of each article. We then train a Byte-level byte-pair-encoding (BPE) tokenizer with 64000 vocabulary size on all of our preprocessed dataset, using the optimized BPE implementation from the HuggingFace library (Wolf et al., 2020). Finally we apply BPE encoding on the preprocessed dataset, which results in a total of 9.7M training examples with 1024 sub-word tokens each.

4 Experiments and Evaluation

4.1 Pretraining Setup

All models were trained on a TPUv3-128 slice⁷ with different batch sizes and 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 Evaluation

We chose the perplexity score as an evaluation metric, since it measures the degree of 'uncertainty' a model has assigning probabilities to the test text. To make sure that the model wasn't being tested on a text from the training set, we use Arabic Wikipedia articles that were published after August 2020, since older articles were included in the September Wikipedia dump.

Table 2 shows that, unsurprisingly, validation perplexity keeps improving with larger model sizes and longer training time. In fact, we are still underfitting the validation set from Wikipedia. We illustrate the generation capabilities of the different variants of ARAGPT2 through the selected examples in Appendix A.

4.3 Discussion

During our initial testing of the **base** and **mega** models, we noticed that the **base** model tends to repeat phrases more than the **mega** model. We solved this issue by applying a repetition penalty as in the CTRL paper (Keskar et al., 2019), which

penalizes the probability scores of previously generated tokens. We also employ a 'no repeat tri-gram' strategy that inhibits the model from generating the same tri-gram more than once. These measures were also necessary with the **mega** model but only when generating very long text from short prompts.

5 Conclusion

ARAGPT2 is the first advanced Arabic language generation model based on the transformer architecture. The model was trained on the largest 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 it is coherent, grammatically correct and syntactically sound.

We hope that publicly releasing ARAGPT2 will open up new research possibilities for Arabic NLP. We are also aware of the ethical implications and dangers of the malicious use of such powerful models i.e. automatic fake news generation, modeling the dataset inherent biases... Hence we are currently working on developing and releasing a detector model that is tasked to detect output generated by all ARAGPT2 variants.

Acknowledgments

This research was supported by the TFRC program, which we thank for the free access to cloud TPUs. We thank As-Safir newspaper for the data access, and also thank Nick Doiron for the insightful discussions.

References

- Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diagglossic 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. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

⁷TPUv3-128 has a total 2TB of HBM memory with 16GB per core. TPUs were freely provided by the TFRC program.

- 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.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. 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.
- 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*.
- 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*.
- 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).
- 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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz 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.

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 الوباء ، 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 الأمريكية (FDA) لقاحها المضاد لفيروس زيكا الذي تم تطويره بالتعاون مع شركة " غلاكسو سميثكلين " 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 لعلاج . لكن مراكز السيطرة على الأمراض والوقاية منها بالولايات المتحدة الأمريكية 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 الوافي
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 باسنتور " (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 " غلاكسو سميث 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 لعلاج . لكن هناك العديد من الإجراءات الاحتياطات 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 العطس أو تغطية 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 1: Random unseen context about coronavirus vaccine(top). Followed by two generated samples by 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 بعثك 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 ثمنًا للبئر ، لكن القاضي رفض الشكوى وقال للرجل المخادع : لقد اشتريت منك البئر لكنك لم تشتتر مني الماء .وفي يوم من الأيام وبينما كان أحد المزارعين يجلس على إحدى شرفات منزله المطلّة على حديقة مليئة بالأشجار رأى رجلا قادمًا نحوه ويده كيس مليء بالحلوى فسلم عليه وسأله عن سبب مجيئه إليه فأخبره الرجل بأنه يريد أن يعطيه بعض الحلوى التي يجلبها ، فقال له المزارع : حسنًا سأعطيك ما تريد بشرط واحد وهو أن لا تخبر أحدا بهذا الأمر أبداً ، وافق الرجل على شرط المزارع وأعطاه الكيس المملوء بالحلوى ، وقبل أن يغادر المزرعة سأله المزارع قائلاً : هل تعلم يا سيدي بأثني 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 قليلاً 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 حلاً لمشكلته ، ذهب جاره المزارع إلى القاضي وأخبره بما حدث مع جاره ، فسأله القاضي مستغرباً : لماذا فعلت ذلك ؟ أجابه الجار : لأنني أريد أن أعرف كيف حصلت على الحلوى دون أن تأخذها مني ودون أن تعطيني لأحد غيرك !وبالفعل عندما وصل الجار إلى القاضي طلب منه القاضي أن يريه الصندوق الذي يحتوي على الحلوى ، نظر القاضي إلى الصندوق فوجد بداخله صندوقاً آخر كبيراً مليئاً بالحلوى ، تعجب القاضي من محتويات الصندوق لكنه سرعان ما قرر أن يفتح الصندوق ليرى ماذا يوجد بداخله ، فتح القاضي الصندوق فوجده مملوءاً 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 صندوقاً ثالثاً مليئاً بالحلوى أيضاً ، سأل القاضي صاحب الصندوقين عما يحتويه كل صندوق فأجاب صاحب الصندوق الثالث قائلاً : 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 نانما ، 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 غاضباً ، فوضع قبضته الضخمة فوق الفأر ، وزمجر وفتح فمه بنوي ابتلاع الفأر الصغير بلقمة واحدة ، صاح الفأر عندها بصوت يرتجف من الخوف راجباً أن يعفو الأسد عنه ، وقال :
Generated Text 1
" سامحني يا ملك الغابة ، أرجوك سامحني ، أنا لا أستطيع العيش بدونك " .أجاب الأسد قائلاً : " 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 ، وسأتي 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 ، ولكنني لم أكن أعرف أنك ملك الغابة وسيدها ، وأن لك سلطاناً على الحيوانات كلها ، لذلك أرجو منك أن تسمح لي بأن أعيش معك تحت سقف واحد وتحت حماية سلطانك وملكك " .فرح الأسد كثيراً بالفكرة ، وطلب من الفأر أن يذهب معه إلى بيته ليعيشا معاً حياة سعيدة هانئة ، وبالفعل ذهب الفأر مع الأسد إلى بيت الأسد الذي رحب به أجمل ترحيب ، وأعطاه قطعة كبيرة من الخشب ليصنع منها سريراً كبيراً له ولزوجته وأولاده ، وهكذا عاش الفأر مع الأسد في سعادة وهناء .وبعد فترة قصيرة من الزمن ، مرض الأسد مرضاً شديداً ، ولم يعد قادراً على تحمل المزيد من الأعباء الملقاة على عاتقه ، فطلب من الفأر أن يساعده في حمل سريريه الثقيل والذهاب به إلى الطبيب ، فوافق الفأر على الفور ، وذهب مسرعاً إلى الطبيب ليخبره بأنه مصاب بمرض خطير يحتاج إلى جراحة عاجلة ، لكنه رفض إجراء الجراحة خوفاً 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 مبتسماً وقال له : " هيا بنا يا صديقي العزيز لنذهب

Figure 2: 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 ونقله فيه الخبز حتى يصبح ذهبيا ومقرمشا .
Generated Text
نوزع شرائح الفليفلة الخضراء والحمراء والخضراء والبنندورة 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 محمي 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 جبلكم لعبة حلوة كثير 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 موطننا 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 الباقي 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 3: Random unseen contexts on miscellaneous topics. Followed by a generated sample by ARAGPT2-mega with $top_p = 0.95$