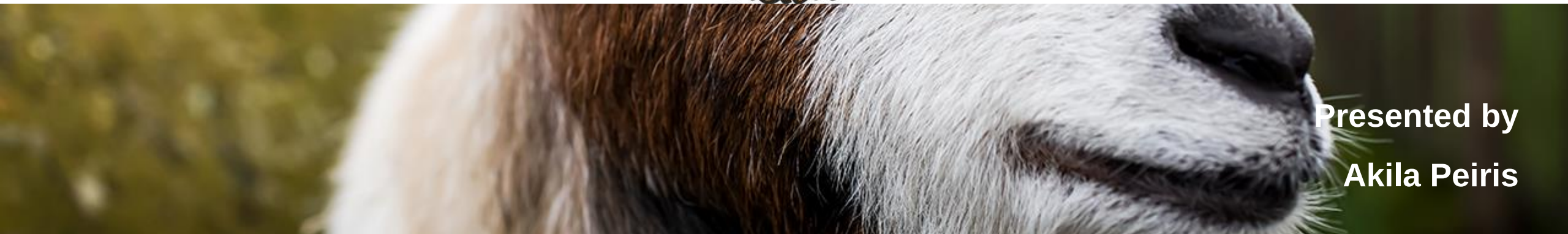
Yudhanjaya Wijeratne

Ishan Marikar

ELUWA

A MORE CONVERSATIONAL
OPT 2.7B

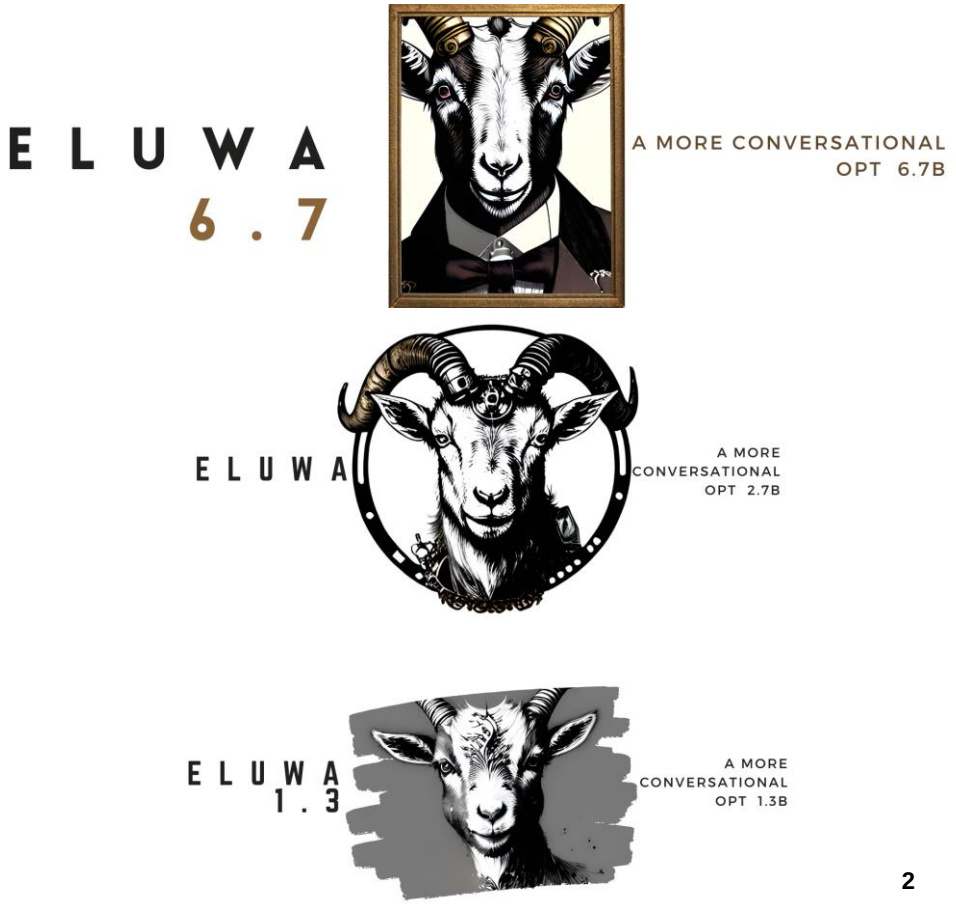


Presented by

Akila Peiris

Better Question-answering Models On A Budget

- https://arxiv.org/pdf/2304.12370.pdf
- Yudhanjaya Wijeratne, Ishan Marikar
- Backyard Labs Colombo, Sri Lanka
- Eluwa models (https://huggingface.co/BackyardLabs)
  - 1.3 B
  - 2.7 B
  - 6.7 B



What We Will Cover

- **Background**
- Methodology
- Testing
- Results

Backgroud: Low Rank Adaptation

- LoRA[1]
- Freeze a pre-trained model and inject smaller weights into it
- Reduces the number of parameters to train and the memory required to do so
- Possible to fine-tune LLM (like GPT-3) with low resources (on a budget)

[1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Backgroud: Self-instruct

- Self-Instruct paper [2] - make LMs better at instruction-following
- By bootstrapping it off its own responses.
- Elegantly create datasets for fine-tuning models

[2] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Han- naneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560, 2022.

Backgroud: Meta AI OPT

- Meta AI's Open Pre-trained Transformer (OPT) models[3]
  - GPT-3 alternative
  - Expensive (~ 992 80GB Nvidia A100 GPUs worth of processing)
- Large Language Model Meta AI (LLaMA)[4]
  - Variants: 7B to 65B parameters
  - 13B model outperformed GPT-3 175B

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

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Bap- tiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

Backgroud: Stanford’s Alpaca

- Stanford Alpaca project[5]
  - 52,000 self-instruct[2] question-answer pairs
  - LLaMA 7B model improved to GPT3’s text-davinci-003 level
- Coincided with 8 and 4-bit quantization improvements[6][7]
  - Less RAM and VRAM
  - Huggingface Transformer library compatible

[2] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Han- naneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560, 2022.

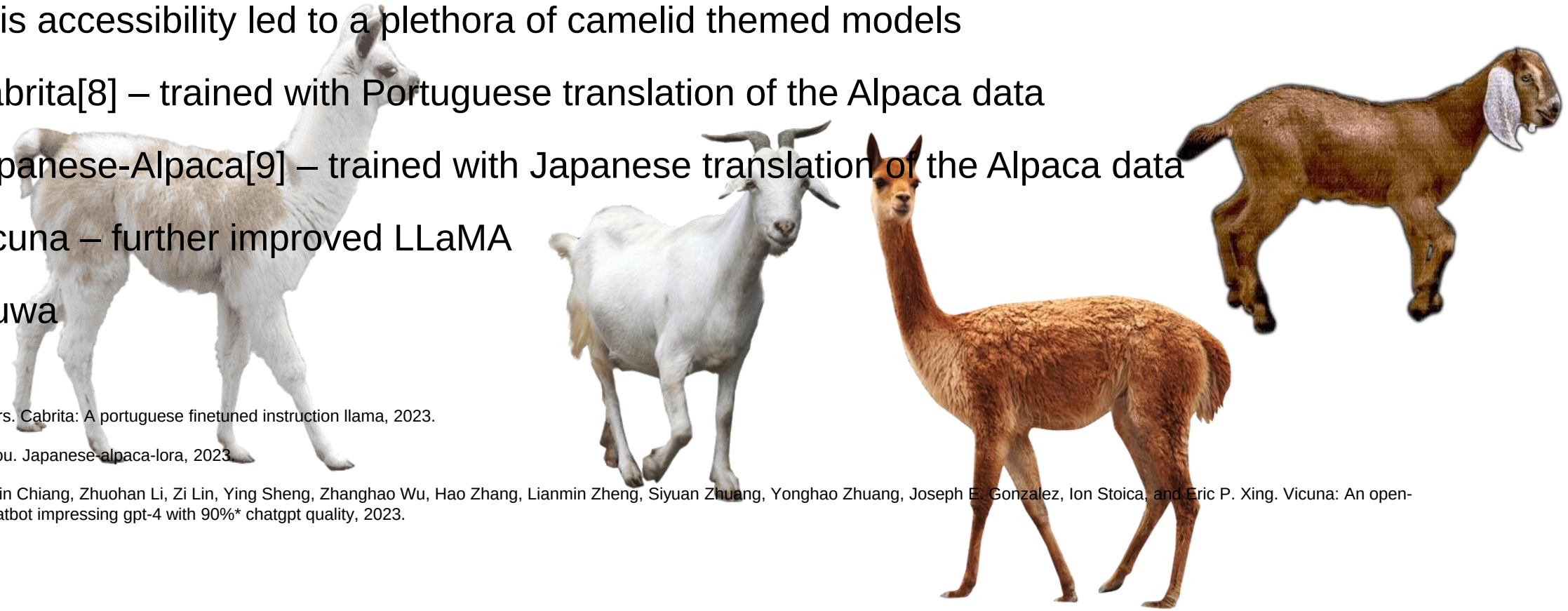
[5] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[6] Tim Dettmers. 8-bit approximations for parallelism in deep learning. arXiv preprint arXiv:1511.04561, 2015.

[7] Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. arXiv preprint arXiv:2212.09720, 2022.

Backgroud: Ruminants and camelids

- This accessibility led to a plethora of camelid themed models
- Cabrita[8] – trained with Portuguese translation of the Alpaca data
- Japanese-Alpaca[9] – trained with Japanese translation of the Alpaca data
- Vicuna – further improved LLaMA
- Eluwa



[8] 22 hours. Cabrita: A portuguese finetuned instruction llama, 2023.

[9] kunishou. Japanese-alpaca-lora, 2023.

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

Methodology

- OPT models (1.3B, 2.3B and 6.7B) and trained on Alpaca using LoRA
- 2 LoRA models for each OPT model
  - Trained for 1000 iterations
  - Trained 2 full epochs
- Deliberately used low-resource parameters
- Trained on Google Colab
- All models fit within a GTX 1080 Ti

Methodology: Training parameters

- per_device_train_batch_size=8
- gradient_accumulation_steps=4
- warmup_steps=100
- learning_rate=2e-4
- fp16=True logging_steps=10

- Used learning rate from Alpaca project as varying between 1e-4 to 3e-4 did not make much of a difference to the loss curve
- Attempts at creating a Sinhala Alpaca using the same methods have failed

Testing: Vicuna Benchmark & GPT-4 scoring

- Adopted Vicuna benchmark: 80 questions under various categories
- Answers scored by GPT-4
- On a single GTX 1080Ti
- Following parameters
- top_p=0.70
  top_k=0
  temperature=1.0
  repetition_penalty=1.1
  typical_p=1.0
  num_beams=1

Testing: Vicuna Benchmark & GPT-4 scoring

Model	OPT 1.3b base	Eluwa 1.3b 1000 iter	Eluwa 1.3b 2 epoch	OPT 2.7b base	Eluwa 2.7b 1000 iter	Eluwa 2.7b 2 epoch	OPT 6.7b base	Eluwa 6.7b 1000 iter	Eluwa 6.7b 2 epoch
Generic	53	58	59	22	44	57	65	74	77
Knowledge	71	69	78	35	60	72	57	89	72
Roleplay	32	44	54	29	38	58	49	70	80
Common sense	41	55	57	20	48	50	67	80	86
Fermi	21	19	17	4	28	23	27	16	26
Counterfactual	28	29	42	5	24	23	30	43	71
Coding	16	7	8	2	7	7	9	9	7
Math	3	3	3	0	3	3	3	3	10
Writing	22	11	54	8	19	19	13	24	10
Total	287	295	372	125	271	312	320	408	439

Testing: Wikitext-2 results

- Tested on Wikitext-2[11]
- On Google Colab
- Following parameters
- temperature=0.1
  top_p=0.8
  top_k=40
  num_beams=1
  repetition_penalty=1.2

[11] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843, 2016.

Testing: Vicuna Benchmark & GPT-4 scoring

Model	Wikitext Perplexity
Eluwa 1.3b 2epoch	14.296875
Eluwa 2.7b 2epoch	12.0859375
Eluwa 6.7b 2epoch	10.6015625
OPT 6.7b base	10.28125
OPT 2.7 base	11.78125
OPT 1.3 base	13.8828125

This benchmark shows a slight increase in perplexity after training
indicating that the new models have a slightly more difficult time predicting the tokens of Wikitext-2

Results

- The models dramatically improved model outputs
  - 30-50% improvement in question-answering behavious
  - Eluwa 2.7B performed better than OPT 6.7b base model
  - 1.3b line is almost anomalously good
- Observable improvement in creative tasks (e.g. roleplaying, counterfactual thinking)
- Coding, Fermi (reasoning), mathematics and vague writing tasks did not improve as much
- Training and testing was completed for a total of about 40 USD

References

[1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[2] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Han- naneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560, 2022.

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

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Bap- tiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[5] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

References

[6] Tim Dettmers. 8-bit approximations for parallelism in deep learning. arXiv preprint arXiv:1511.04561, 2015.

[7] Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. arXiv preprint arXiv:2212.09720, 2022.

[8] 22 hours. Cabrita: A portuguese finetuned instruction llama, 2023.

[9] kunishou. Japanese-alpaca-lora, 2023.

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[11] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843, 2016.

GPT-3 models

LATEST MODEL	DESCRIPTION	MAX REQUEST	TRAINING DATA
`gpt-3.5-turbo`	Most capable GPT-3.5 model and optimized for chat at 1/10th the cost of `text-davinci-003`. Will be updated with our latest model iteration.	4,096 tokens	Up to Sep 2021
`gpt-3.5-turbo-0301`	Snapshot of `gpt-3.5-turbo` from March 1st 2023. Unlike `gpt-3.5-turbo`, this model will not receive updates, and will only be supported for a three month period ending on June 1st 2023.	4,096 tokens	Up to Sep 2021
`text-davinci-003`	Can do any language task with better quality, longer output, and consistent instruction-following than the curie, babbage, or ada models. Also supports inserting completions within text.	4,097 tokens	Up to Jun 2021
`text-davinci-002`	Similar capabilities to `text-davinci-003` but trained with supervised fine-tuning instead of reinforcement learning	4,097 tokens	Up to Jun 2021
`code-davinci-002`	Optimized for code-completion tasks	8,001 tokens	Up to Jun 2021

- https://neoteric.eu/blog/gpt-4-vs-gpt-3-openai-models-comparison/