Session #5: Comparing Architectures

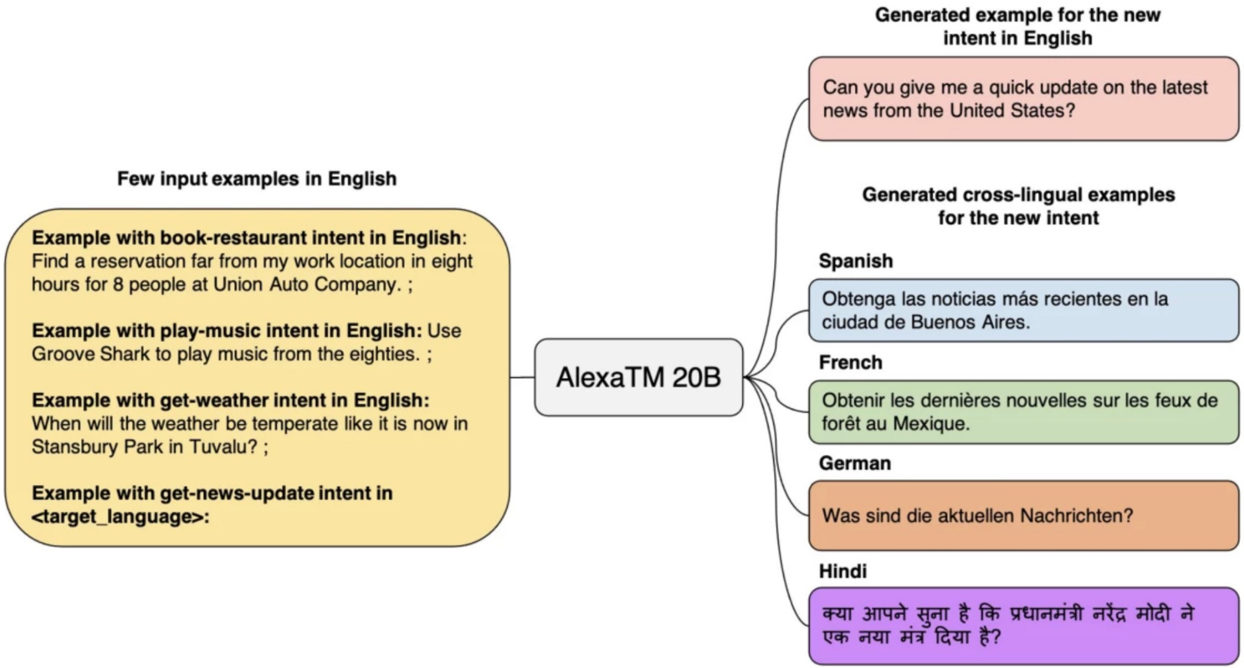
Thursday, Sept 8

CSCI 601.771: Self-supervised Statistical Models

JOHNS HOPKINS UNIVERSITY

News: 20B-parameter Alexa model

"With an encoder-decoder architecture — rather than decoder only — the Alexa Teacher Model excels other large language models on few-shot tasks such as summarization and machine translation."



Using AlexaTM 20B to generate annotated data for a new intent in different languages.

https://www.amazon.science/blog/20b-parameter-alexa-model-sets-new-marks-in-few-shot-learning?utm_source=email&utm_medium=newsletter&utm_campaign=science_news&utm_term=august_2022&utm_content=20B-parameter&mkt_tok=MzIzLU1GUio1NjgAAAGGj9mj5p7IZnN1QvKvIUlieyofl2txXmQLKXbjCtGejsnWt9FtUhlWH9wDIvpVhUZtWkm8Z2AYkmnVnJ1-XsV7hkVqZoBKfno1xY4

Week's prompt

This would have been a much better paper if ____

What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?

Thomas Wang [1 *] **Adam Roberts** [2 *] **Daniel Hesslow** [3] **Teven Le Scao** [1] **Hyung Won Chung** [2] **Iz Beltagy** [4]
Julien Launay [3 5 †] **Colin Raffel** [1 †]

By: Ayo Ajayi and Muhannad Muzammil Godil

Zero-shot Generalization

- What is Zero-shot Generalization?

✍️: Ayo Ajayi and Muhannad Muzammil Godil

Source

Zero-shot Generalization

- What is Zero-shot Generalization?
  - “Zero-shot” allows a model to recognize what it hasn’t seen before.

✍️: Ayo Ajayi and Muhannad Muzammil Godil

Source

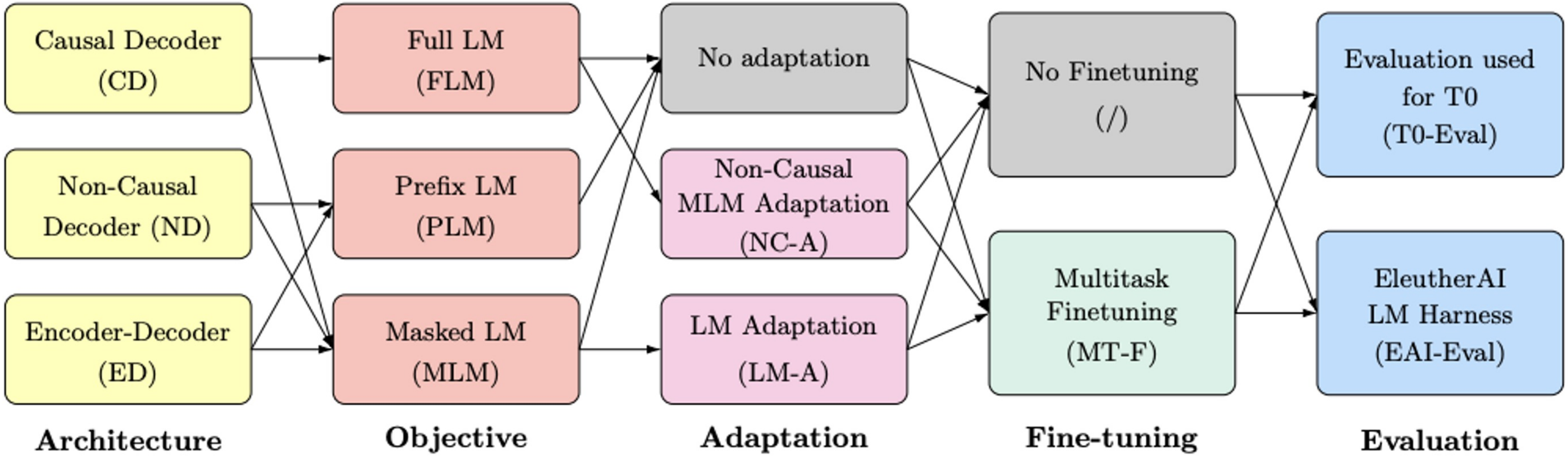
Zero-shot Generalization

- What is Zero-shot Generalization?
  - "Zero-shot" allows a model to recognize what it hasn't seen before.
  - The capability of large language models pretrained on unstructured data to perform tasks without additional training.
  - The ability of large language models to perform a wide variety of tasks that they were not explicitly trained on.

✍️: Ayo Ajayi and Muhannad Muzammil Godil

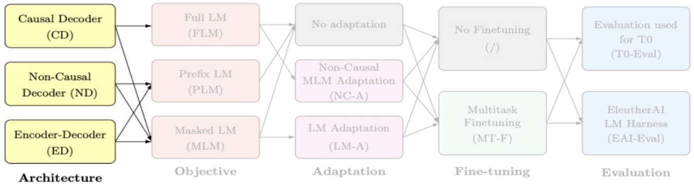
Source

Motivation

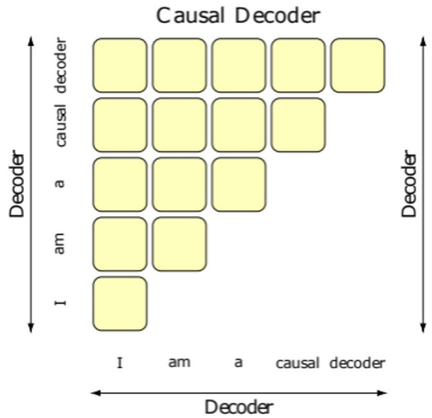


What is the perfect formula for attaining zero-shot generalization?

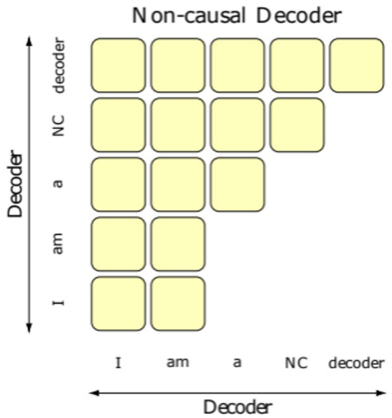
✍: Ayo Ajayi and Muhannad Muzammil Godil



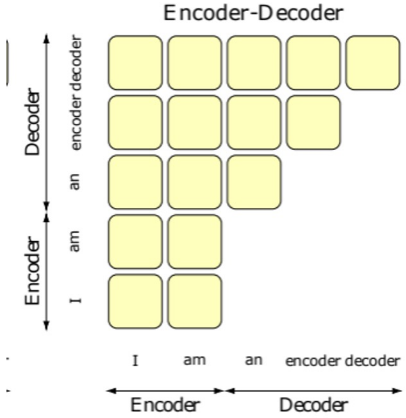
Architecture



Each token attends to previous tokens only



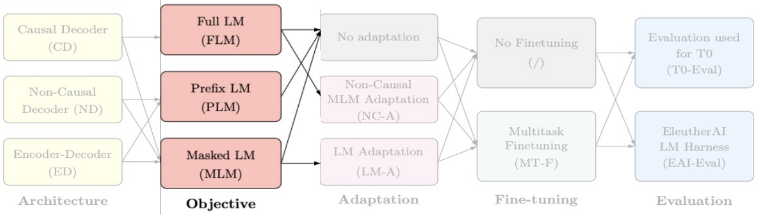
Attention is allowed to be bidirectional on any conditioning information



Attention is allowed to be bidirectional on any conditioning information fed into the encoder.

✍: Ayo Ajayi and Muhannad Muzammil Godil

Objective

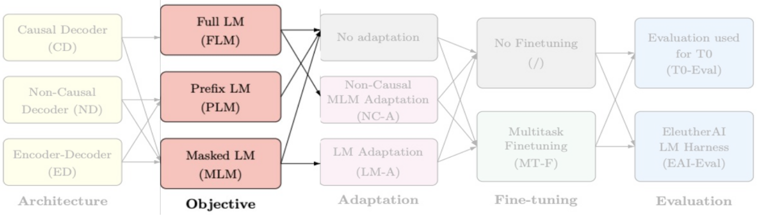


Full Language Modeling May **targets:** the force be with you

Given previous tokens, the model is tasked with predicting the following one. Large decoder-only models.

✍️: Ayo Ajayi and Muhannad Muzammil Godil

Objective



Full Language Modeling May [targets: the force be with you]

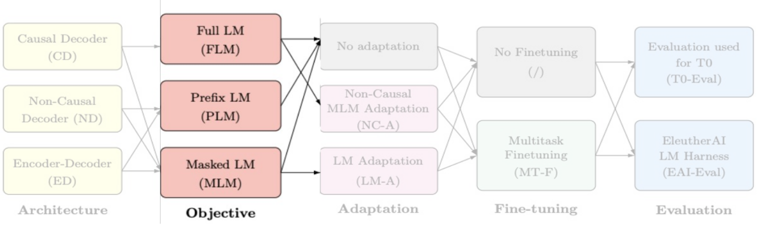
Given previous tokens, the model is tasked with predicting the following one. Large decoder-only models.

Prefix Language Modeling May the force [targets: be with you]

Predict each token outside the prefix given all previous tokens. Encoder-decoder and non-causal decoder-only models.

✍: Ayo Ajayi and Muhannad Muzammil Godil

Objective



Full Language Modeling May ~~~targets~~~

targets

May the force be with you

Given previous tokens, the model is tasked with predicting the following one. Large decoder-only models.

Prefix Language Modeling

targets

May the force be with you

Predict each token outside the prefix given all previous tokens. Encoder-decoder and non-causal decoder-only models.

Masked Language Modeling

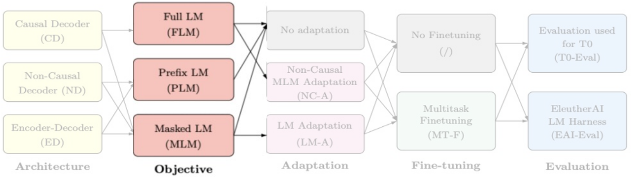
targets

May the force be with you

Replacing certain tokens with a mask. Encoder-only Models.

Objective - Modeling Configuration

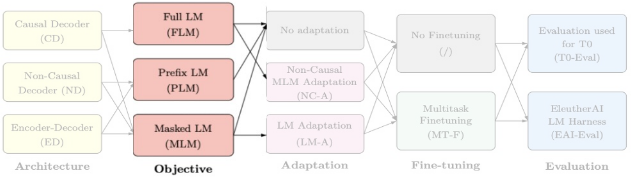
- For full language modeling, all tokens in a sequence are used during training.
- For prefix language modeling, we randomly select a prefix size, and hence only half of the tokens are used on average to derive the loss.
- For masked language modeling, we mask 15% of the tokens, in spans of 3 tokens on average.



✍️: Ayo Ajayi and Muhannad Muzammil Godil

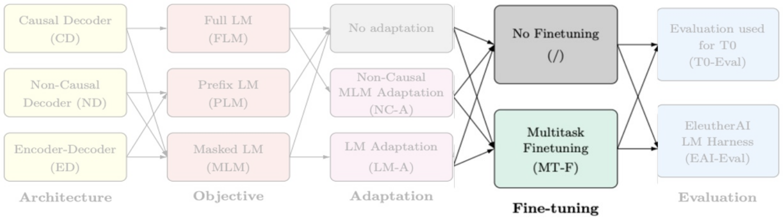
Self-Supervised Learning

- Self-supervised learning is **a machine learning process where the model trains itself to learn one part of the input from another part of the input.** It is also known as predictive or pretext learning.



✍️: Ayo Ajayi and Muhannad Muzammil Godil

Multitask Fine-tuning

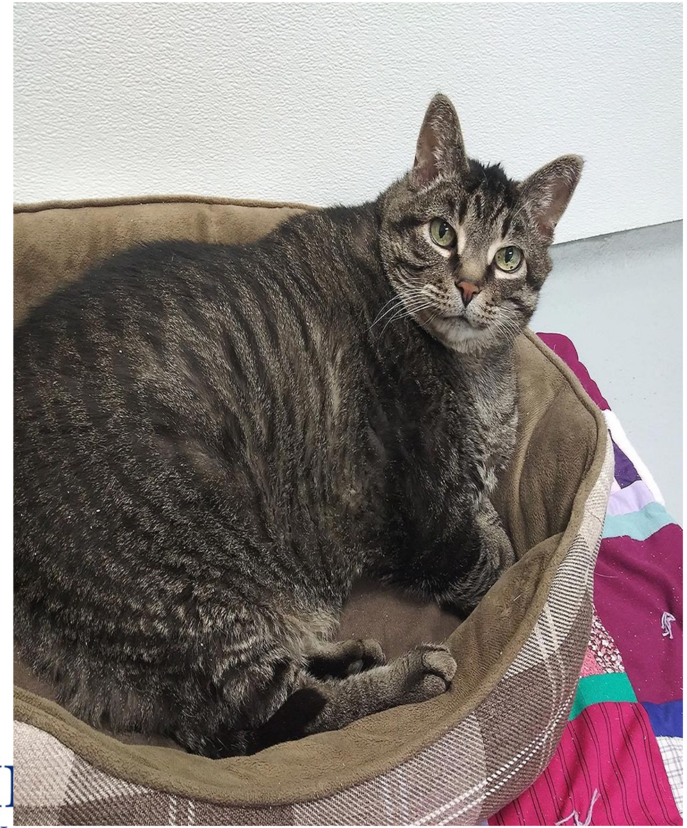


- Recent work shows that multitask fine-tuning improves zero-shot performance.
- Multitask Fine-tuning: fine-tuning the model on a dataset of prompted tasks.
    - Explicitly fine-tune the model to solve different tasks.

✍: Ayo Ajayi and Muhannad Muzammil Godil

Method

- All previously mentioned **architecture** was trained on **168 billion tokens**
- **Multitask fine tuning** is then considered and evaluation is done on **zero shot performance**
- The training budget was similar across all models **15 petaflops-days over 830,000 TPUv4 hours**



JOHNS HOPKI
UNIVERSITY

Results:

- We use 2 **zero-shot** benchmarks to evaluate our training **T0-Eval** & **EAI-Eval**

- T0-Eval provides **multiple prompts** per task vs EAI-Eval which provides **1 prompt** per task

✍: Ayo Ajayi and Muhannad Muzammil Godil

To And EAI

- To was just a model that was evaluated for a set of tasks in 2021 (Sanh ,2021)

- We use the same set of tasks which is why it is called **To-Eval**

- **EAI - Eval** uses just 1 prompt per task it has 200 + tasks available to test AR models on

- The main difference is that To provides **multiple prompts** per task while EAI gives **one**

✍️: Ayo Ajayi and Muhannad Muzammil Godil

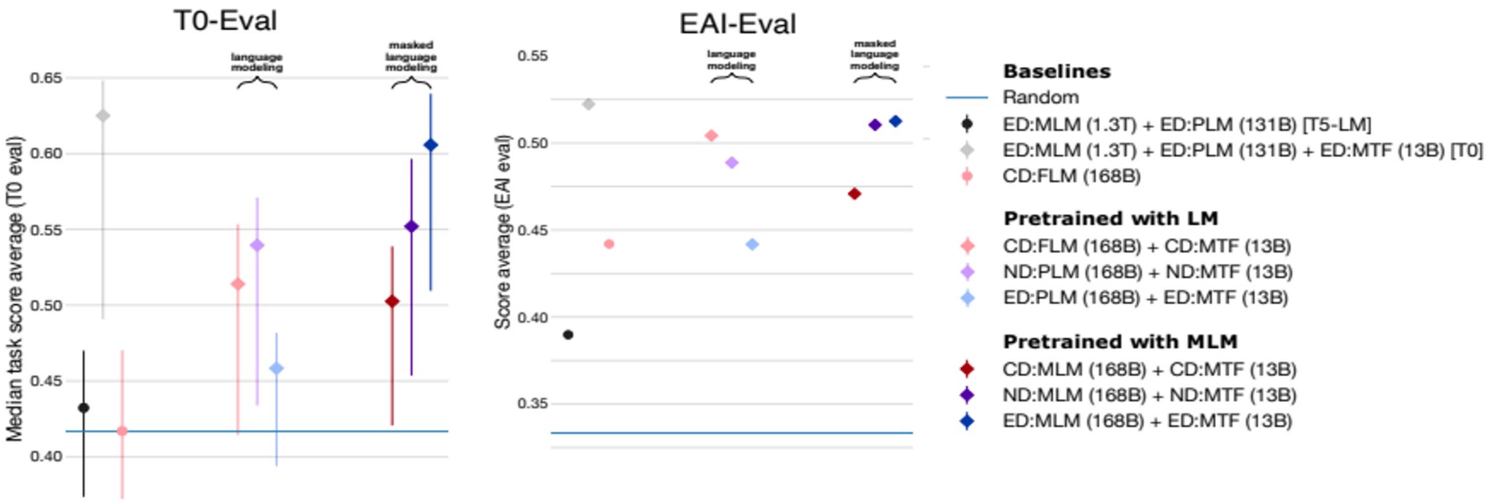
Self supervised pre-training

	EAI-EVAL	T0-EVAL
Causal decoder	**44.2**	**42.4**
Non-causal decoder	43.5	41.8
Encoder-decoder	39.9	41.7
Random baseline	32.9	41.7

Causal decoder-only models pretrained with a **full language modeling objective** achieve best **zero-shot generalization** when evaluated immediately after **self-supervised pre-training**.

✍: Ayo Ajayi and Muhannad Muzammil Godil

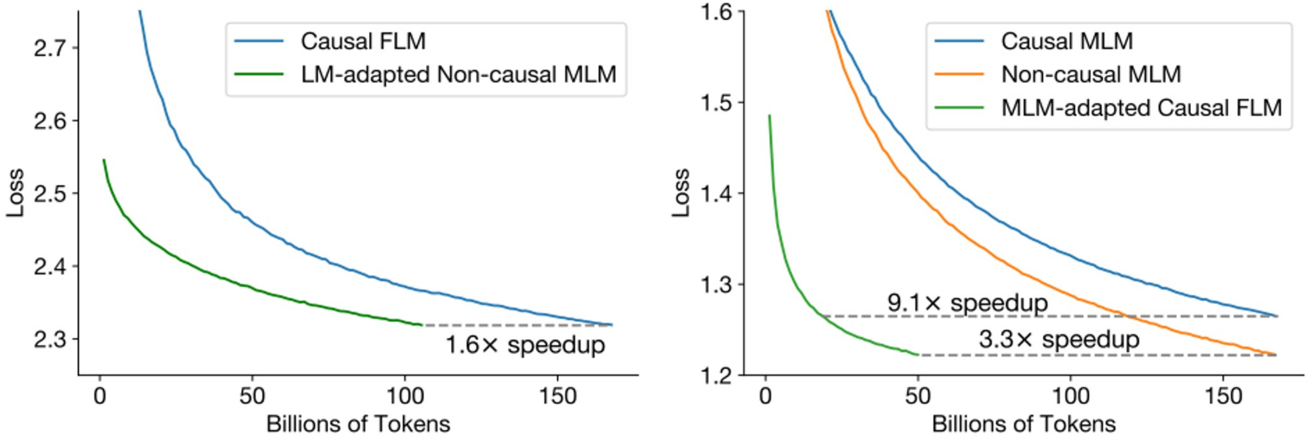
After Multi Task fine tuning



Encoder-decoder models pretrained with **masked language modeling** achieve the best zero-shot performance after **multitask finetuning** .

✍: Ayo Ajayi and Muhannad Muzammil Godil

Adaptation from architectures



Decoder-only models can be efficiently adapted from one architecture/objective prior to the other. Starting with a **causal decoder-only model**, pretraining it with a full language modeling objective, and then using non-causal masked language modeling adaptation before taking it through **multitask fine tuning .**

Positive

🟢 Systematic empirical investigation of Language Modelling choices

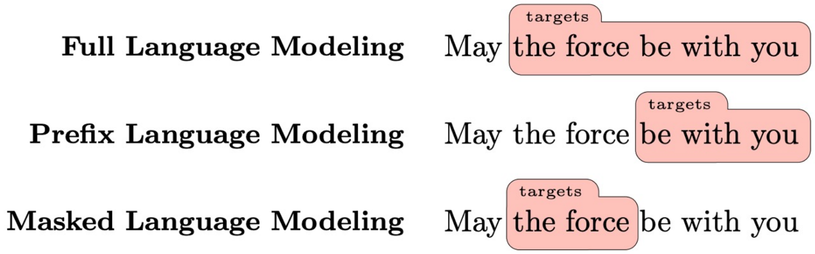
🟢 Actionable takeaways that anyone using these models can benefit from

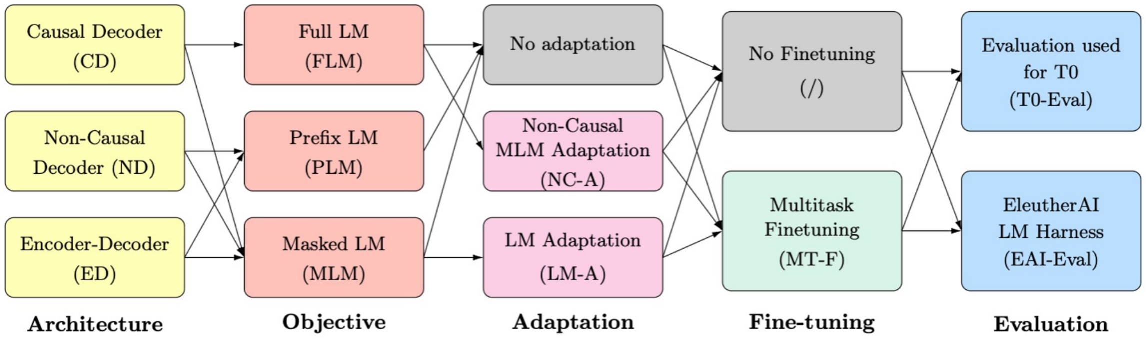
🟢 Interesting experiments in adaptation between architecture and objectives

🔍: Isabel, Karan

Positive

- 🟢 Release of code and model benchmarks
- 🟢 Clear writing and visualizations





🔍: Isabel, Karan

Potential Caveats

- 🟡 Can only do so many ablations with a limited budget
  - E.g. no discussion of effects of hyperparameter choices, non-linearity functions, etc

- 🟡 Unclear if these experiments generalize to larger models

- 🟡 No breakdown of each individual task within the benchmarks, only averages across tasks
  - No statistical analysis or hypothesis testing

🔍: Isabel, Karan

Potential Caveats

- 🟡 Not clear why the compute budget was chosen to remain constant as opposed to overall model size or decoder size
- 🟡 Not including encoder models for adaptation
    - encoder-decoder -> causal decoder (mentioned)
    - encoder-decoder -> encoder (not mentioned)
    - Mentioning it would answer “How important is Encoder vs Decoder information in context of Zero Shot learning?”

Table 2. **Shared architecture for all models trained.** Encoder-decoder architectures are doubled in size to obtain a pretraining compute budget similar to the decoder-only architecture.

	MODELS ARCHITECTURE	
	Decoder-only	Encoder-decoder
Parameters	4.8B	11.0B
Vocabulary	32,128	
Positional embed.	T5 relative	
Embedding dim.	4,096	
Attention heads	64	
Feedforward dim.	10,240	
Activation	GEGLU (Shazeer, 2020)	
Layers	24	48
Tied embeddings	True	
Precision	`bfloat16`	

🔍: Isabel, Karan

Empiricists

CD(GPT) vs. ED(BART)

Yongrui Qi & Fadil Isamotu

pretrained Only

👩🏽‍🔬: Yongrui, Fadil

Mask Filling: Prompt → “The sun is <mask>. ”

GPT3

```python
import openai
openai.api_key = ("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")
my_prompt = '''The sun is [MASK].

    Replace [MASK] with the most probable 5 words to replace, and give me their probabilities.'''
# Here set parameters as you like
response = openai.Completion.create(
  engine="text-davinci-002",
  prompt=my_prompt,
  temperature=0,
  max_tokens=100,
)

print(response['choices'][0]['text'])
```

Result:

```
1. shining: 0.348
2. bright: 0.298
3. beautiful: 0.183
4. glorious: 0.091
5. lovely: 0.081
```

BART

```
from transformers import BartTokenizer, BartForConditionalGeneration

tokenizer = BartTokenizer.from_pretrained("facebook/bart-large")
model = BartForConditionalGeneration.from_pretrained("facebook/bart-large")

TXT = "The sun is <mask> ."
input_ids = tokenizer([TXT], return_tensors="pt")["input_ids"]
logits = model(input_ids).logits

masked_index = (input_ids[0] == tokenizer.mask_token_id).nonzero().item()
probs = logits[0, masked_index].softmax(dim=0)
values, predictions = probs.topk(5)

tokenizer.decode(predictions).split()
```

Result:

```
['located', 'at', 'approximately', 'also', 'about']
```

Text Generation: Prompt→ I enjoy walking with my cute dog, ...

BART:

```
[{'generated_text': 'I enjoy walking with my cute dog, should should just should should shouldto shouldBar justERAJusttoBarBar justtoBar justBarBar'}]
```

GPT2:

```
[{'generated_text': 'I enjoy walking with my cute dog, which has a little tendency to bite my paws!"\n\n(Photos by David Evans Photography)'}]
```

GPT3:

```
[{'generated_text': 'I enjoy walking with my cute dog, because it relaxes me.'}]
```

After multitask finetuning

Summarization: prompt’s word count = 327

The early decades of the 21st century saw expansion across the university's institutions in both physical and population sizes. Notably, a planned 88-acre expansion to the medical campus began in 2013. Completed construction on the Homewood campus has included a new biomedical engineering building in the Johns Hopkins University Department of Biomedical Engineering, a new library, a new biology wing, an extensive renovation of the flagship Gilman Hall, and the reconstruction of the main university entrance. These years also brought about the rapid development of the university's professional schools of education and business. From 1999 until 2007, these disciplines had been joined within the School of Professional Studies in Business and Education (SPSBE), itself a reshuffling of several earlier ventures. The 2007 split, combined with new funding and leadership initiatives, has led to the simultaneous emergence of the Johns Hopkins School of Education and the Carey Business School. On November 18, 2018, it was announced that Michael Bloomberg would make a donation to his alma mater of $1.8 billion, marking the largest private donation in modern history to an institution of higher education and bringing Bloomberg's total contribution to the school in excess of $3.3 billion. Bloomberg's $1.8 billion gift allows the school to practice need-blind admission and meet the full financial need of admitted students. In January 2019, the university announced an agreement to purchase the Newseum, located at 555 Pennsylvania Ave. NW, in the heart of Washington, D.C., with plans to locate all of its D.C.-based graduate programs there. In an interview with The Atlantic, the president of Johns Hopkins stated that “the purchase is an opportunity to position the university, literally, to better contribute its expertise to national- and international-policy discussions.” In late 2019, the university's Coronavirus Research Center began tracking worldwide cases of the COVID-19 pandemic by compiling data from hundreds of sources around the world. This led to the university becoming one of the most cited sources for data about the pandemic.

BART

Result: Word count = 42

```
The early decades of the 21st century saw expansion across the
university's institutions. A planned 88-acre expansion to the medical
campus began in 2013. In January 2019, the university announced an
agreement to purchase the Newseum, located at 555 Pennsylvania Ave. NW.
```

GPT2

Result: Word count = 148

The early decades of the 21st century saw expansion across the university's institutions in both physical and population sizes. Notably, a planned 88-acre expansion to the medical campus began in 2013. Completed construction on the Homewood campus has included a new biomedical engineering building in the Johns Hopkins University Department of Biomedical Engineering, a new library, a new biology wing, an extensive renovation of the flagship Gilman Hall, and the reconstruction of the main university entrance. These years also brought about the rapid development of the university's professional schools of education and business. From 1999 until 2007, these disciplines had been joined within the School of Professional Studies in Business and Education (SPSBE), itself a reshuffling of several earlier ventures. The 2007 split, combined with new funding and leadership initiatives, has led to the simultaneous emergence of the Johns Hopkins School of Education and the Carey Business School….

GPT3

Result: Word count = 43

```
Johns Hopkins University is a private research university in Baltimore, Maryland.
Founded in 1876, the university was named for its first benefactor, the American
entrepreneur, abolitionist, and philanthropist Johns Hopkins. His $7 million bequest
(approximately $144.5 million in today's dollars)—of which half
```

Archaeologist

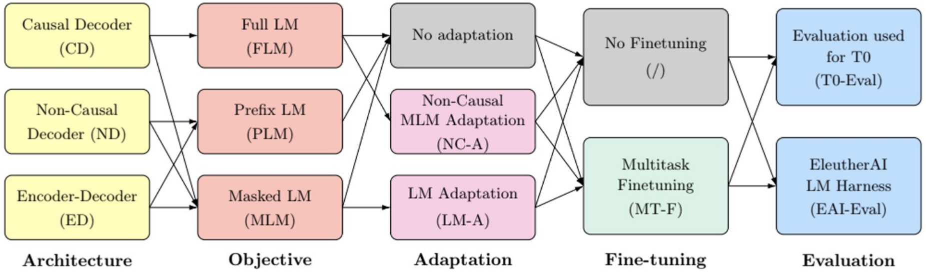
Putting things in the context of the literature …

🏺: Haoyue, Ha

Large Language Models in zero-shot generalization overview

Overview components of Large Language Models (LLM) in zero-shot generalization:

- Pretraining Objective: Self-supervised training technique for LLM
- Architecture: Backbone corresponds to objective
- Adaptation: Add pretraining data after casting/converting architecture
- Fine-tuning: Update model parameter
- Evaluation: Evaluate zero-shot generalization
- => **Relationship** between **components matters** in generalization **evaluation**!



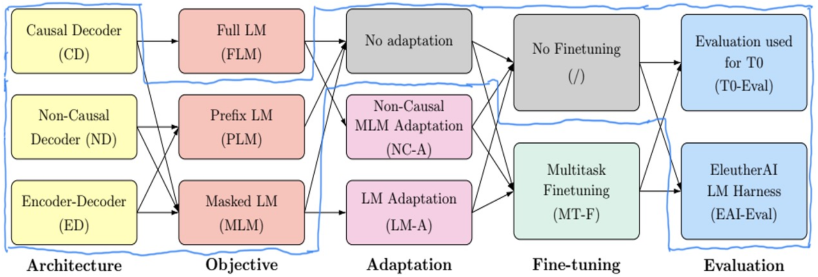
: Ha, Haoyue

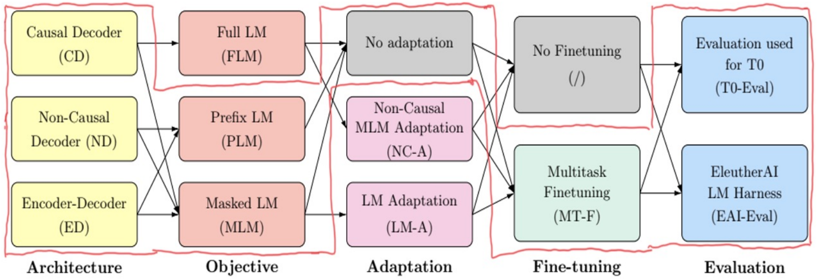
Wang et. al., 2022: What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?

Prior works

- Raffel et. al., 2020: shows that encoder-decoder models outperform decoder-only LLMs for **transfer learning**

Raffel





Sanh

- Sanh et. al., 2021: concludes **multitask finetuned** encoder-decoder LLM outperforms decoder-only models on zero-shot generalization
- etc.

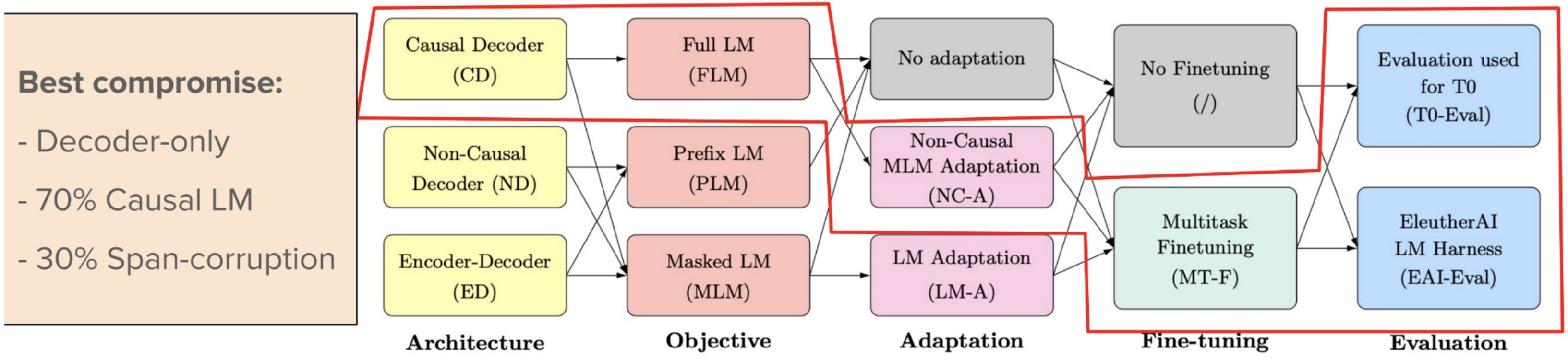
Problem: Prior works only consider a small part of the picture => Unfair comparison

Motivation: What we need is **results** on **all possible paths** of this **directed graph!**

: Ha, Haoyue

Wang's finding

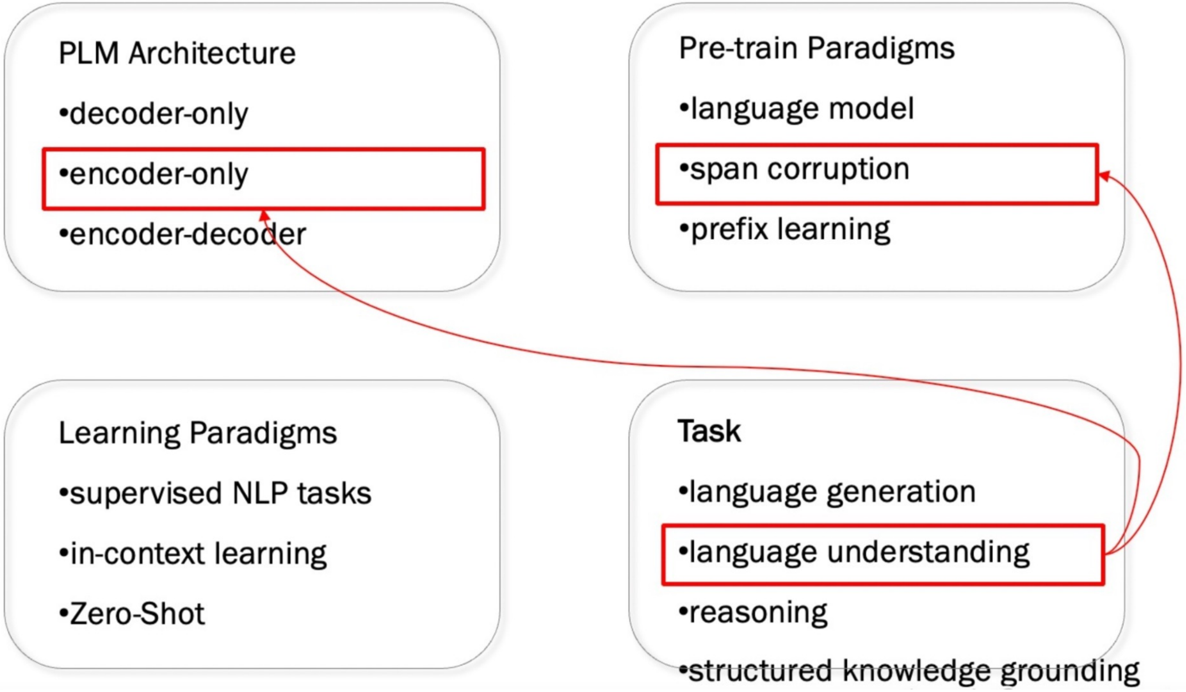
Wang et. al., 2022: What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?



: Ha, Haoyue

Rethinking the current fine-tuning method

- We have a variety of pre-training paradigms:



: Ha, Haoyue

Rethinking the current fine-tuning method

- Different paradigms model different contextual relationships.

- Because of the method above, different pre-training paradigms are **adapted to different types of downstream tasks.**
  e.g. span corruption (T5) is more applicable to fact completion.
  PrefixLM/LM (GPT) is more suitable for open ended.

Problem:
Specific pre-training strategies need to be chosen for specific downstream task types!
Deploying specific models for different downstream tasks is very resource-intensive!

🏺: Ha, Haoyue

Subsequent Work

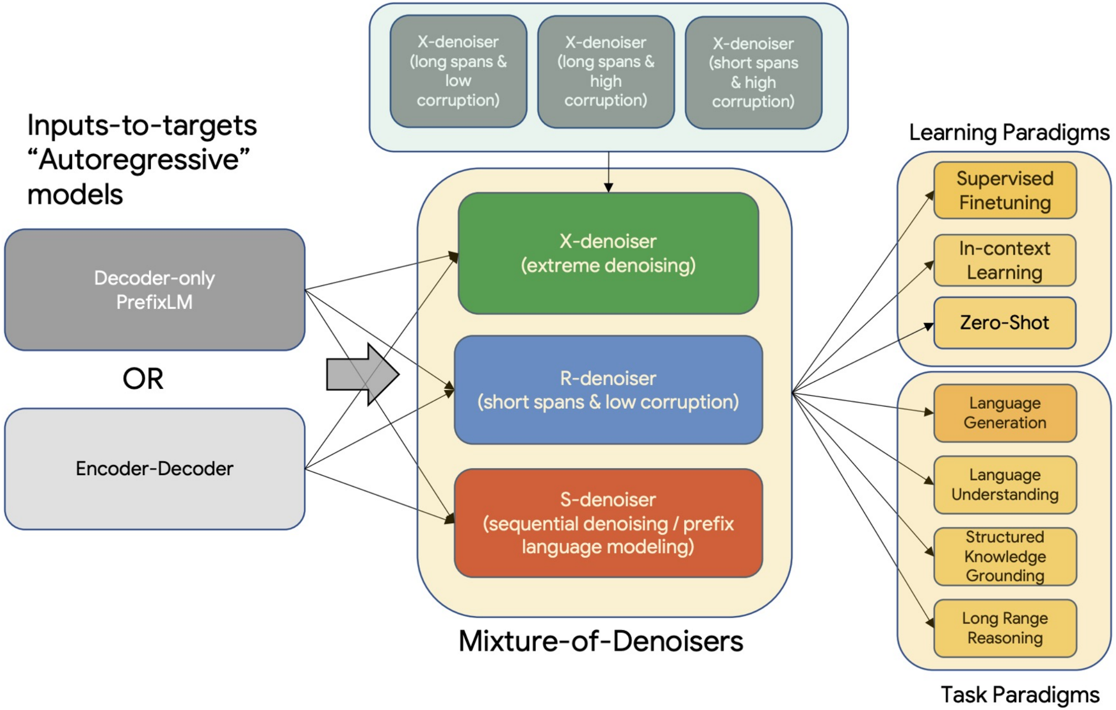
- Tay, Yi, et al. 2022 "Unifying Language Learning Paradigms." : What we need is a **unified large model**!

Motivation: Construct a pre-training strategy that is independent of the model architecture, and can be flexibly adapted to different types of downstream tasks.

Why?

- With a unified model, it is possible to focus on improving and extending individual models.
- It is desirable to have a pre-trained model that can perform well on multiple tasks. (Resource is limited, takes long time to train PLM)

Subsequent Work



Tay, Yi, et al. 2022: Unifying Language Learning Paradigms.

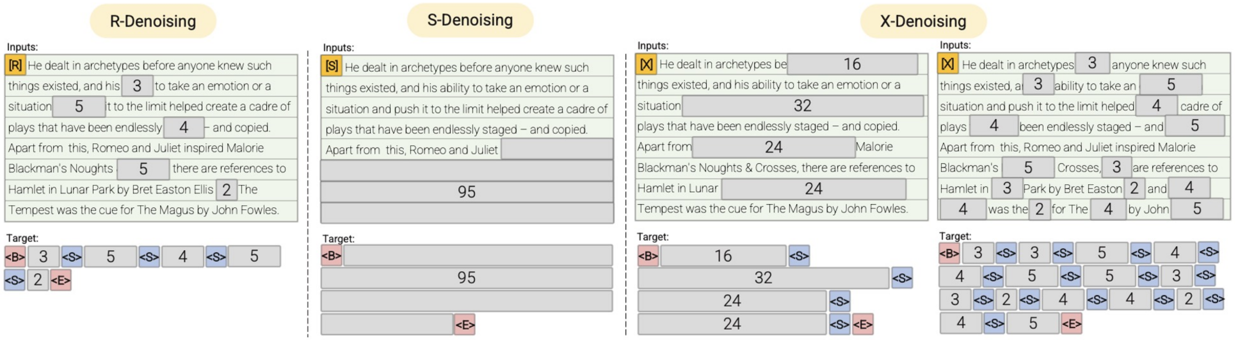
: Ha, Haoyue

Subsequent Work

R-Denoiser : regular denoising. Is the T5 span corruption objective.

S-Denoiser : sequential denoising. Is connected to causal language models that are GPT-like.

X-Denoiser : extreme denoising. Can expose the model to a combination of objectives from T5 and Causal LMs.



Tay, Yi, et al. 2022: Unifying Language Learning Paradigms.

🏺 : Ha, Haoyue

Obj	Arch	Supervised SG	Supervised XS	Supervised SGD	Supervised TOT	One-shot SGL	One-shot XS	One-shot SGD	One-shot TOT	LM	All	Win
CLM	Dec	-13.6	-9.2	-0.7	-3.0	+1.8	-91.7	-2.2	-90.5	+208	-31.7	2/9
PLM	Dec	-13.3	-9.2	-0.5	-2.8	+10.5	-85.6	+158	+205	+185	-11.0	4/9
SC	Dec	-5.6	-6.2	-0.6	-1.3	+0.05	-84.5	+54	-23.8	+99	-20.6	3/9
SCLM	Dec	-6.0	-6.5	-0.2	-2.0	+5.9	-59.6	-11.3	-95	+204	-16.1	2/9
UniLM	Dec	-10.1	-8.2	-0.2	-2.3	-5.3	-69.1	+382	+110	+200	-16.1	3/9
UL2	Dec	-9.0	-6.9	0.0	-1.4	+9.8	+6.9	+340	+176	+209	**+14.1**	5/9
PLM	ED	-3.7	+2.9	-0.2	-0.6	-0.86	-13.3	+397	+86	+199	+16.7	5/9
SC*	ED	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
SCLM	ED	+0.7	+2.1	-0.2	-0.5	+3.2	-31.6	+508	+248	+201	+28.3	7/9
UniLM	ED	-1.2	-0.2	+0.1	-0.4	+3.5	-11.0	+355	+95	+173	+19.8	5/9
UL2	ED	+1.5	+2.6	+0.5	+0.4	+7.2	+53.6	+363	+210	+184	**+43.6**	**9/9**

Relative performance compared to standard encoder-decoder span corruption model (T5).

Obj	Arch	Supervised SG	Supervised XS	Supervised SGD	Supervised TOT	One-shot SG	One-shot XS	One-shot SGD	One-shot TOT	LM	All	Win
CLM*	Dec	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
PLM	Dec	+0.3	+0.1	+0.2	+0.2	+8.5	+74.3	+164	+3100	-8.0	+21.4	8/9
UniLM	Dec	+4.0	+1.1	+0.5	+0.7	-7.0	+274	+393	+2100	-2.5	+21.0	7/9
SC	Dec	+8.7	+3.4	+0.1	+1.8	-1.8	+87.0	+57.1	+700	-54.2	+13.9	7/9
SCLM	Dec	+1.8	+3.0	+0.5	+1.0	+4.0	+387	-9.3	-50	-1.3	+15.8	6/9
UL2	Dec	+5.2	+2.6	+0.6	+1.7	+7.9	+1190	+350	+2800	+0.3	**+45.7**	**9/9**
PLM	ED	+11.3	+13.4	+0.5	+2.5	-2.6	+946	+408	+1850	-2.9	+48.6	7/9
SC	ED	+16.5	+10.2	+0.6	+3.1	-1.8	+1107	+2.3	+950	-208	+31.7	7/9
SCLM	ED	+15.7	+12.5	+0.5	+2.6	+1.3	+726	+522	+3550	-2.2	+60.3	8/9
UniLM	ED	+14.2	+10.0	+0.7	+2.7	+1.6	+974	+365	+1950	-12.9	+52.6	8/9
UL2	ED	+17.4	+13.1	+1.2	+3.5	+5.3	+1754	+373	+3150	-8.3	**+76.1**	8/9

Relative performance compared to standard decoder causal language model (GPT-like)

🏺: Ha, Haoyue

Experiments

Table 5: Effect of different paradigm prompts on 1-shot evaluation, using a Encoder-Decoder architecture pre-trained using UL2 on 7B tokens.

Model/Prompt	1Shot XSum	1Shot SuperGLUE
Baseline T5	6.9/0.6/6.1	33.9
UL2 / None	13.2/1.4/10.8	38.3
UL2 / [R]	**13.5/1.5/11.1**	38.5
UL2 / [S]	11.6/1.2/10.0	38.5
UL2 / [X]	8.9/0.9/7.6	**38.7**

- In one-shot scenarios, it is almost always better to use paradigm prompts, but it is critical to pick the right one.

🏺: Ha, Haoyue

Visionary Interpretation

🔭: Ammar, Elisee

Ammar
Elisée

Paper Approach

- The paper tested multitask finetuning with three model architectures and two different types of pretraining objectives.
- The authors suggest that other architectures can be tried out, like sparsely gated mixture of experts.
- We want to try a different solution approach with images, since this paper talks more regarding text

🔭: Ammar, Elisee

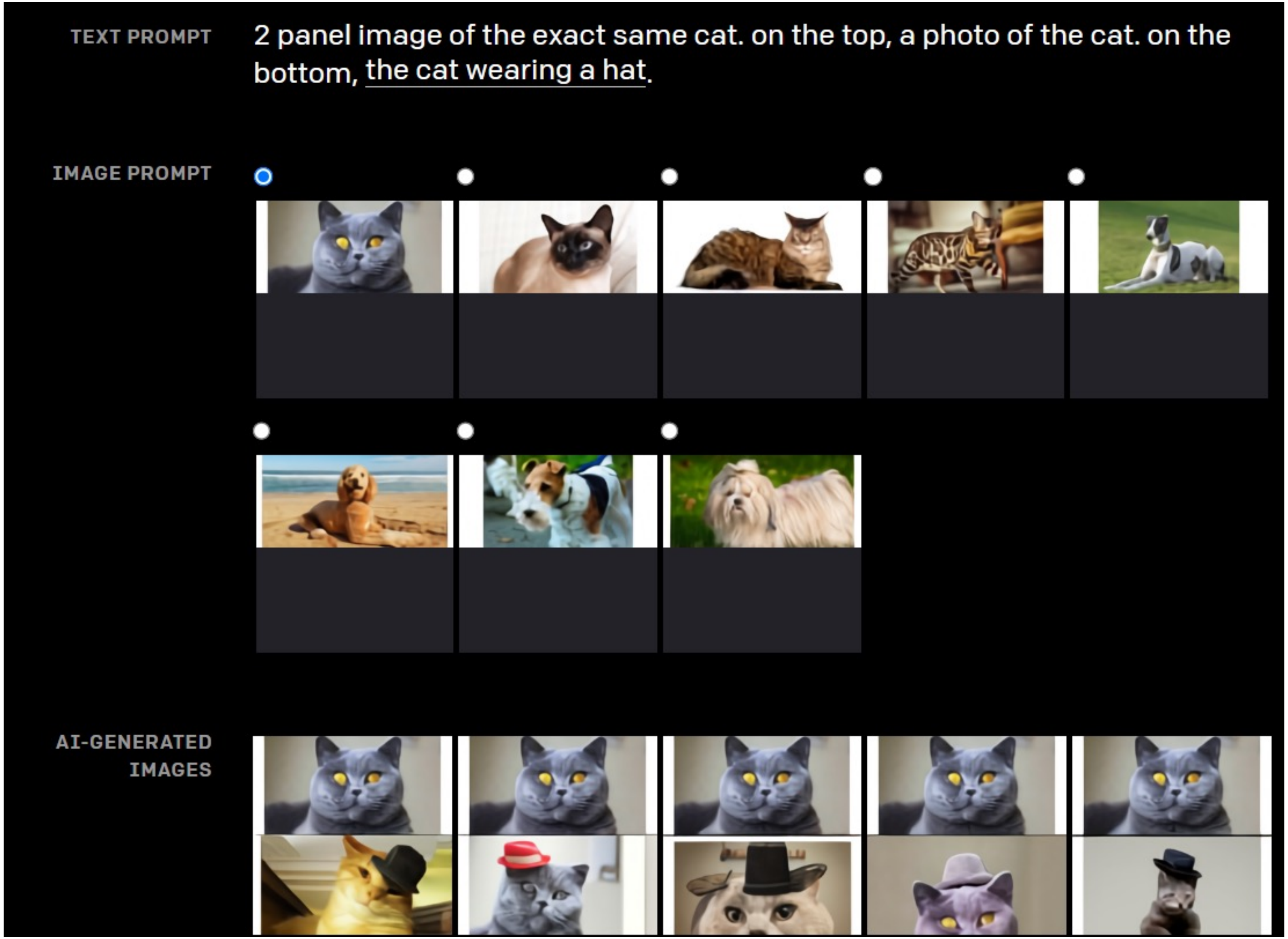
Application on Images

- Comparison of architectures and objectives for images
- VAEs and GANs can be considered for the generation block in the architecture
- Zero-Shot Text to Image Generation, 2021 ICML

  https://arxiv.org/pdf/2102.12092.pdf
- Uses VAE to generate images
- Limitation: Lack of proven options to compare in zero shot image generation, unlike text

🔭: Ammar, Elisee

DALL-E



🔭: Ammar, Elisee

Complimentary Work

- ***Using DeepSpeed and Megatron to Train Megatron-Turing NLG530B, A Large-Scale Generative Language Model***

- Microsoft DeepSpeed deep learning optimization library
- NVIDIA Megatron-LM large transformer model
- Training of the largest monolithic transformer-based language model, Megatron-Turing NLG 530B (MT-NLG),with 530 billion parameters.

🔭: Ammar, Elisee

Complimentary Work

- Will help further with development of large-scale training infrastructures, large-scale language models, and natural language generations.
- which achieves superior zero-, one-, and few-shot learning accuracies and new state-of-the-art results on NLP benchmarks.

🔭: Ammar, Elisee