Risk Engineering Co-Pilot: RAG vs. Fine-Tuning in creating a risk assessment assistant tool with domain-specific knowledge for Insurance industry

Amin Karbassi, Ph.D.
Senior Casualty Risk Consultant
amin.karbassi@axaxl.com

SDS2024
IEEE SWISS CONFERENCE ON DATA SCIENCE

Why a Risk Engineering Co-Pilot tool is beneficial?

Function of Risk Engineers in Insurance

Translate risk

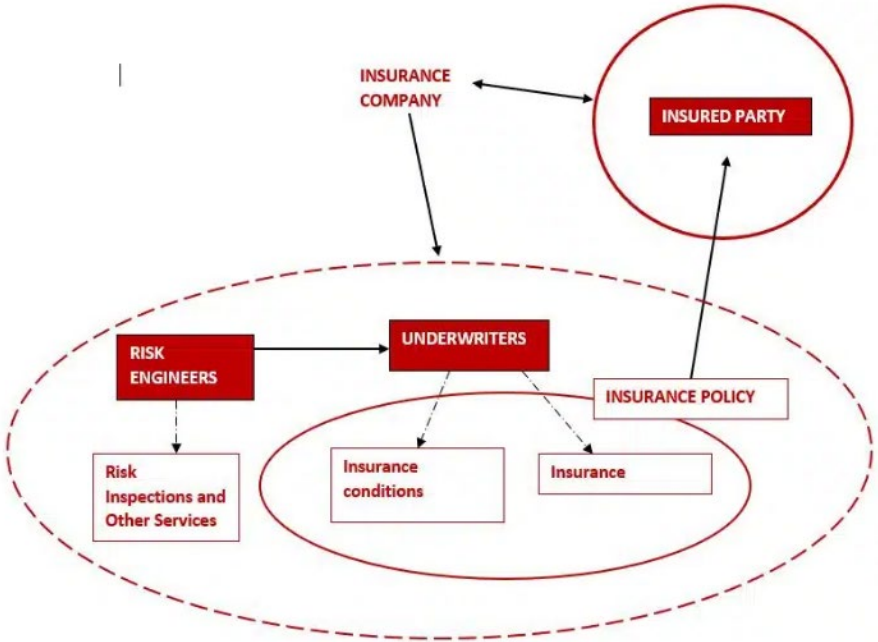
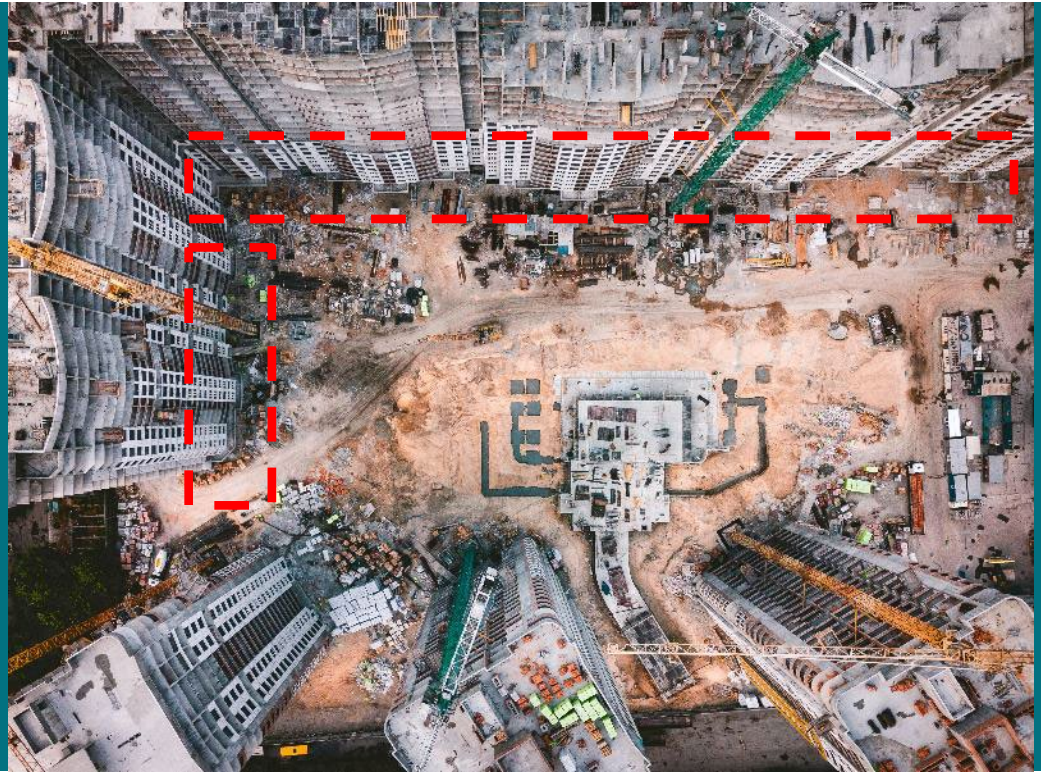


Photo source: MAPFRE Global Risks
https://www.mapfreglobalrisks.com/en/risks-insurance-management/topicality/the-insurance-company-risk-engineer/

Ideal Co-Pilot tool

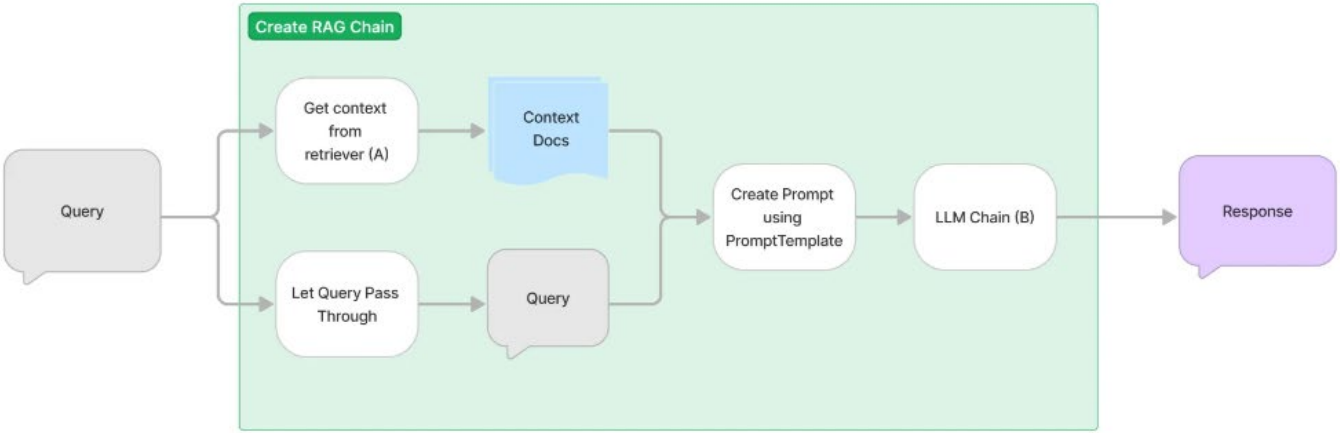
Construction projects (hypothetically)



Retrieval-Augmented Generation (RAG)

Mistral-7B and LangChain

- Combining power of retrieval and generation mechanisms
- FAISS used to create the vector database



Retrieval-Augmented Generation (RAG)

Mistral-7B and LangChain

*'**question**': 'Why is typhoon the biggest natural catastrophe risk for offshore wind farms?'*

*'**RAG text**': ' According to the provided documents, typhoons pose the biggest natural catastrophe risk for offshore wind farms due to their vulnerability to these extreme weather events. Offshore wind farms are exposed to both wind and water, creating a complex risk scenario…*

Document source:
Managing offshore wind project risk in nat cat-prone APAC (Swiss Re Corso)
https://corporatesolutions.swissre.com/insights/knowledge/managing-offshore-wind-project-risk-in-nat-cat-prone-APAC.html

OpenAI ChatGPT 4o

Comparison

Why is typhoon the biggest natural catastrophe risk for offshore wind farms?

Typhoons pose the biggest natural catastrophe risk for offshore wind farms primarily due to several key factors:

1. **High Wind Speeds**: Typhoons bring extremely high wind speeds that can exceed the design limits of wind turbines. This can cause structural damage to the turbines, including blades, nacelles, and towers, leading to significant operational disruptions and repair costs.

2. **Wave Loads**: Typhoons generate massive ocean waves that exert tremendous forces on the foundations and substructures of offshore wind turbines. These wave loads can cause fatigue and structural failures, particularly if the wind farm is not designed to withstand such extreme conditions.

3. **Storm Surge**: The low-pressure system of a typhoon can cause a rise in sea level, known as a

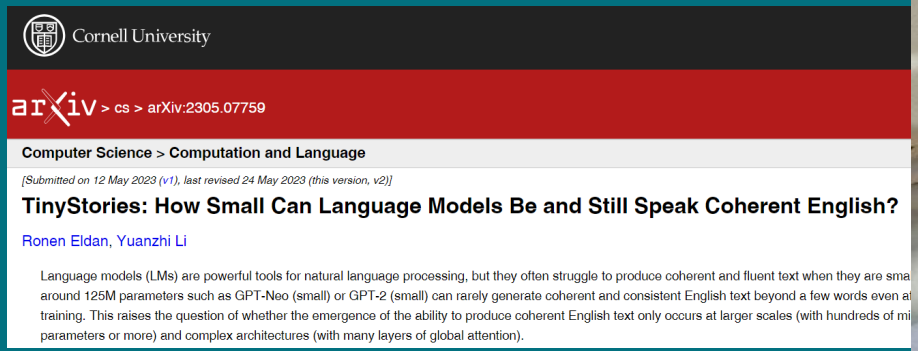
Fine Tuning

- Updating language models for specific use cases
  - → updating the weights or parameters within large language models, whether through backpropagation or new methods
- Getting more attention due to increasing popularity of small models (lower performance though)

Fine Tuning

Examples

- **TinyStories**: synthetic dataset of short stories containing words that a typical 3 to 4-year-olds usually understand - generated by GPT-3.5 and GPT-4
  - → Models with around 125M parameters such as GPT-Neo (small) or GPT-2 (small) rarely generate coherent and consistent English text even after extensive training
  - → The dataset were used to fine-tune smaller LMs (below 10 million total parameters) or have simpler architectures to produce fluent and consistent stories with almost perfect grammar

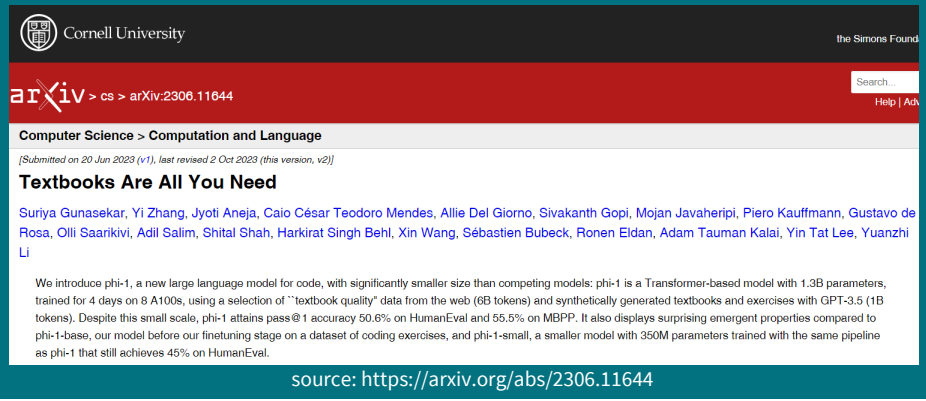


source: https://arxiv.org/abs/2305.07759

Fine Tuning

Examples of usefulness

- **phi-1-small**: fine-tuned a 1.3B parameter model on textbook-quality coding exercises to produce a 350M parameter model with very good results



Cornell University

the Simons Found

arXiv > cs > arXiv:2306.11644

Search...

Help | Adv

Computer Science > Computation and Language

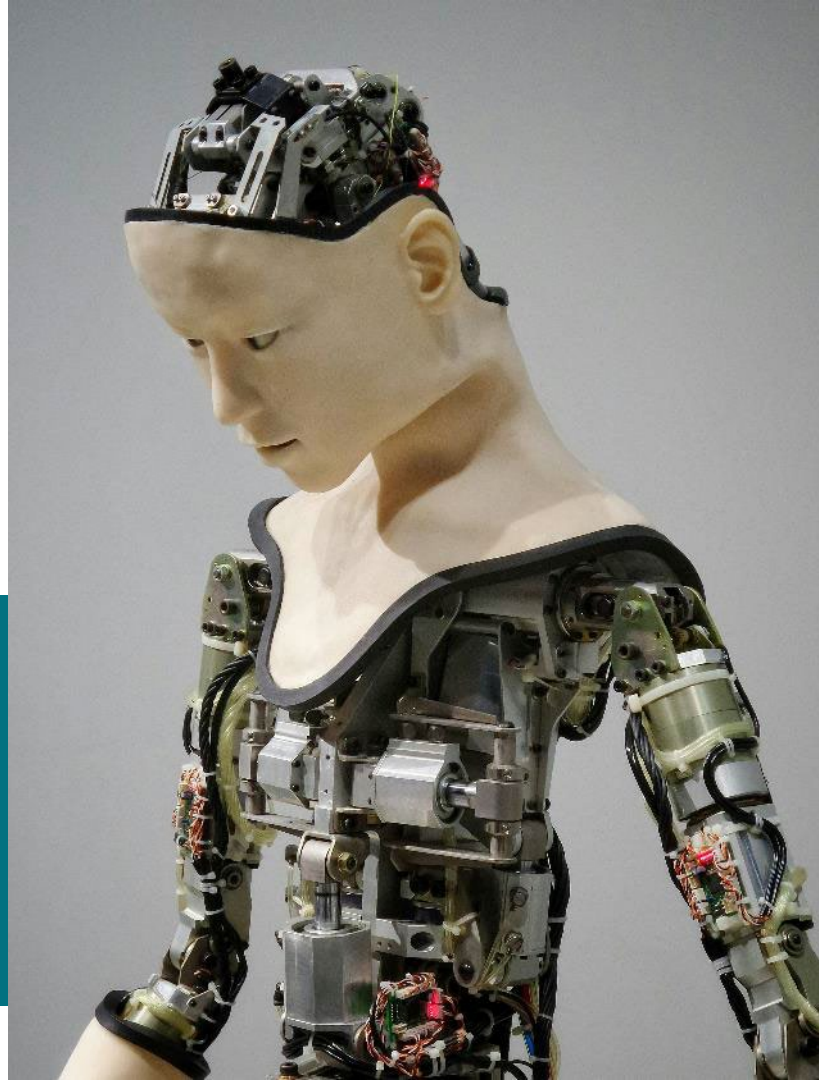
[Submitted on 20 Jun 2023 (v1), last revised 2 Oct 2023 (this version, v2)]

Textbooks Are All You Need

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, Yuanzhi Li

We introduce phi-1, a new large language model for code, with significantly smaller size than competing models: phi-1 is a Transformer-based model with 1.3B parameters, trained for 4 days on 8 A100s, using a selection of ``textbook quality" data from the web (6B tokens) and synthetically generated textbooks and exercises with GPT-3.5 (1B tokens). Despite this small scale, phi-1 attains pass@1 accuracy 50.6% on HumanEval and 55.5% on MBPP. It also displays surprising emergent properties compared to phi-1-base, our model before our finetuning stage on a dataset of coding exercises, and phi-1-small, a smaller model with 350M parameters trained with the same pipeline as phi-1 that still achieves 45% on HumanEval.

source: https://arxiv.org/abs/2306.11644



Fine Tuning

Main obstacle(s)

- Ongoing trend of larger models:
  - → need for more efficient ways to deal with their weights and parameters
- Problem of performance gains:
  - → LLMs are large and difficult to manage due to their billions of parameters
  - → Fine-tuning involves adjusting some weights in a pre-trained network rather than changing all of them
  - → This presents challenges in terms of hardware capabilities and efficient deployment
- **Possible solution**: leverage best parts of model without overwhelming computing resources

→ lightweight fine-tuning

Fine Tuning

PEFT LoRA

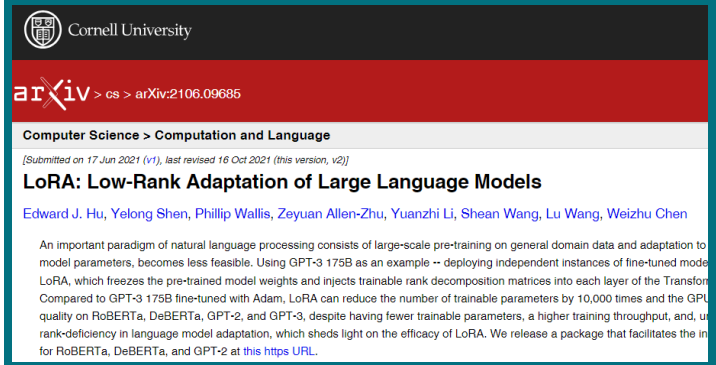
- **Parameter Efficient Fine-Tuning**: aims to fine-tune only a small subset of the model's parameters
- LoRa: rank decomposition on updated weight matrices

$$W_0 \in \mathbb{R}^{d \times k}$$

$$W_0 + \Delta W = W_0 + BA$$

$$\text{where } B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$$

$$\text{rank } r \ll \min(d, k)$$


Cornell University

arXiv > cs > arXiv:2106.09685

Computer Science > Computation and Language

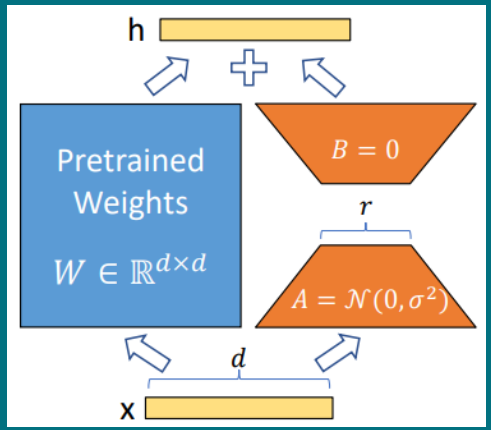
[Submitted on 17 Jun 2021 (v1), last revised 16 Oct 2021 (this version, v2)]

LoRA: Low-Rank Adaptation of Large Language Models

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen

An important paradigm of natural language processing consists of large-scale pre-training on general domain data and adaptation to
model parameters, becomes less feasible. Using GPT-3 175B as an example -- deploying independent instances of fine-tuned mode
LoRA, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transfo
Compared to GPT-3 175B fine-tuned with Adam, LoRA can reduce the number of trainable parameters by 10,000 times and the GPU
quality on RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, a higher training throughput, and, u
rank-deficiency in language model adaptation, which sheds light on the efficacy of LoRA. We release a package that facilitates the in
for RoBERTa, DeBERTa, and GPT-2 at this https URL.

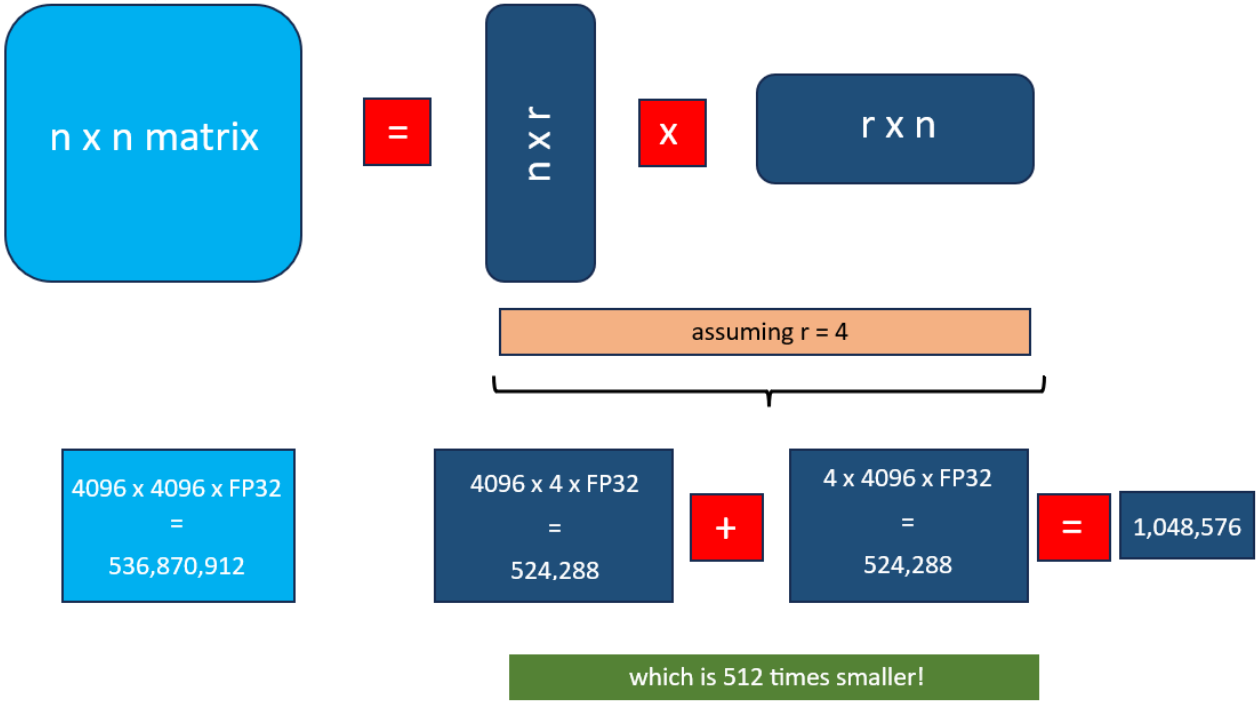
source: https://arxiv.org/pdf/2106.09685



Fine Tuning

PEFT LoRA

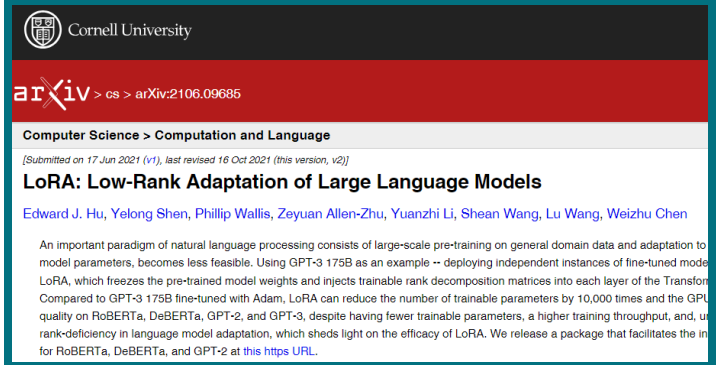
$$W_0 + \Delta W = W_0 + BA$$



Fine Tuning

LoRA (Benefits and Limitations)

- Reduction in memory and storage usage
  - → on GPT-3 175B, with r = 4 and only the weights of **query** and **value** projection matrices being adapted (ref. LoRa paper)
    - *checkpoint size is reduced by roughly 10,000× (from 350GB to 35MB)*
  - → The 350GB model still needed during deployment
    - storing 100 adapted models only requires 350GB + 35MB * 100 ≈ **354GB** as opposed to 100 * 350GB ≈ **35TB**
- Switch between tasks at a much lower cost by only swapping the LoRA weights


Cornell University

arXiv > cs > arXiv:2106.09685

Computer Science > Computation and Language

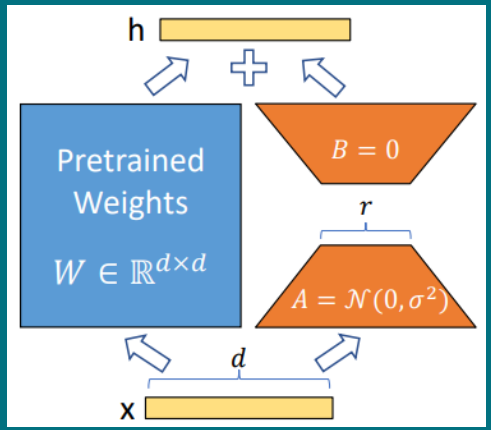
[Submitted on 17 Jun 2021 (v1), last revised 16 Oct 2021 (this version, v2)]

LoRA: Low-Rank Adaptation of Large Language Models

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen

An important paradigm of natural language processing consists of large-scale pre-training on general domain data and adaptation to
model parameters, becomes less feasible. Using GPT-3 175B as an example -- deploying independent instances of fine-tuned mode
LoRA, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transfor
Compared to GPT-3 175B fine-tuned with Adam, LoRA can reduce the number of trainable parameters by 10,000 times and the GPU
quality on RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, a higher training throughput, and, u
rank-deficiency in language model adaptation, which sheds light on the efficacy of LoRA. We release a package that facilitates the in
for RoBERTa, DeBERTa, and GPT-2 at this https URL.

source: https://arxiv.org/pdf/2106.09685



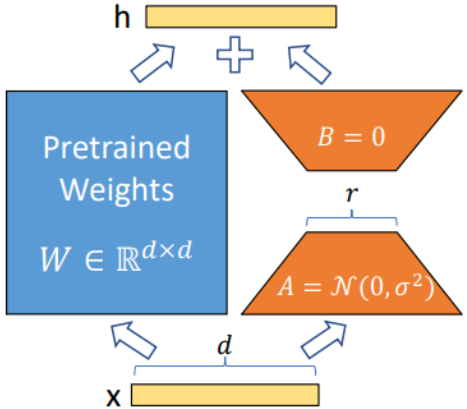
Fine Tuning

LoRA on HuggingFace

- PEFT Library

```
from peft import LoraConfig

config = LoraConfig(init_lora_weights="gaussian", ...)
```



source: https://arxiv.org/pdf/2106.09685

LoraConfig

class peft.LoraConfig <source>

```
( peft_type: Union = None, auto_mapping: Optional = None, base_model_name_or_path: Optional = None,
revision: Optional = None, task_type: Union = None, inference_mode: bool = False, r: int = 8,
target_modules: Optional[Union[list[str], str]] = None, lora_alpha: int = 8, lora_dropout: float = 0.0,
fan_in_fan_out: bool = False, bias: Literal['none', 'all', 'lora_only'] = 'none', use_rslora: bool =
False, modules_to_save: Optional[list[str]] = None, init_lora_weights: bool | Literal['gaussian',
'loftq'] = True, layers_to_transform: Optional[Union[list[int], int]] = None, layers_pattern:
Optional[Union[list[str], str]] = None, rank_pattern: Optional[dict] = <factory>, alpha_pattern:
Optional[dict] = <factory>, megatron_config: Optional[dict] = None, megatron_core: Optional[str] =
'megatron.core', loftq_config: Union[LoftQConfig, dict] = <factory>, use_dora: bool = False,
layer_replication: Optional[list[tuple[int, int]]] = None )
```

Parameters

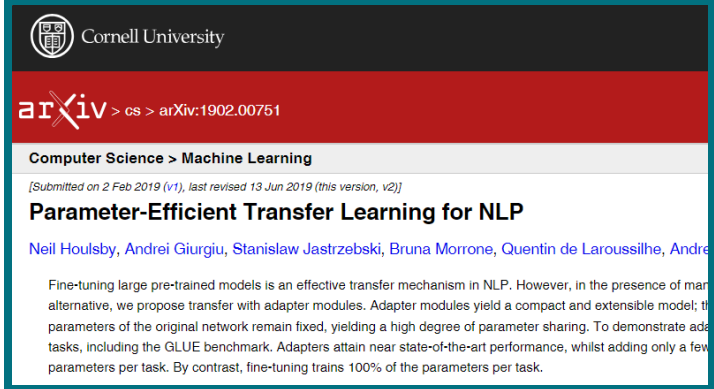
- **r** (`int`) — Lora attention dimension (the “rank”).
- **target_modules** (`Optional[Union[List[str], str]]`) — The names of the modules to apply the adapter to. If this is specified, only the modules with the specified names will be replaced. When passing a string, a regex match will be performed. When passing a list of strings, either an exact match will be performed or it is checked if the name of the module ends with any of the passed strings. If this is specified as ‘all-linear’, then all linear/Conv1D modules are chosen, excluding the output layer. If this is not specified, modules will be chosen according to the model architecture. If the architecture is not known, an error will be raised — in this case, you should specify the target modules manually.
- **lora_alpha** (`int`) — The alpha parameter for Lora scaling.
- **lora_dropout** (`float`) — The dropout probability for Lora layers.
- **fan_in_fan_out** (`bool`) — Set this to True if the layer to replace stores weight like (fan_in, fan_out). For example, gpt-2 uses `Conv1D` which stores weights like (fan_in, fan_out) and hence this should be set to `True`.
- **bias** (`str`) — Bias type for LoRA. Can be ‘none’, ‘all’ or ‘lora_only’. If ‘all’ or ‘lora_only’, the corresponding biases will be updated during training. Be aware that this means that, even when disabling the adapters, the model will not produce the same output as the base model would have without adaptation.
- **use_rslora** (`bool`) — When set to True, uses Rank-Stabilized LoRA which sets the adapter scaling factor to

source: https://huggingface.co/docs/peft/v0.10.0/en/package_reference/lora#peft.LoraConfig

Fine Tuning

Alternatives to PEFT LoRa

- **Adapters layers** in Transformer blocks: adapt parameters of certain layers and leave the rest unchanged
  - → adapter layers have very few parameters, even as few as 1% of the transformer's Feed forward layers
  - → during training, it is very compute efficient.
- Large networks are usually parallelized on hardware, while adapter layers must be processed sequentially
  - → During inference, adapters introduce a noticeable latency


Cornell University

arXiv > cs > arXiv:1902.00751

Computer Science > Machine Learning

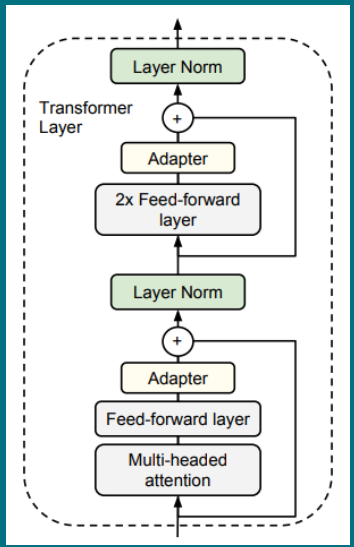
[Submitted on 2 Feb 2019 (v1), last revised 13 Jun 2019 (this version, v2)]

Parameter-Efficient Transfer Learning for NLP

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andre

Fine-tuning large pre-trained models is an effective transfer mechanism in NLP. However, in the presence of mar alternative, we propose transfer with adapter modules. Adapter modules yield a compact and extensible model; t parameters of the original network remain fixed, yielding a high degree of parameter sharing. To demonstrate ada tasks, including the GLUE benchmark. Adapters attain near state-of-the-art performance, whilst adding only a few parameters per task. By contrast, fine-tuning trains 100% of the parameters per task.

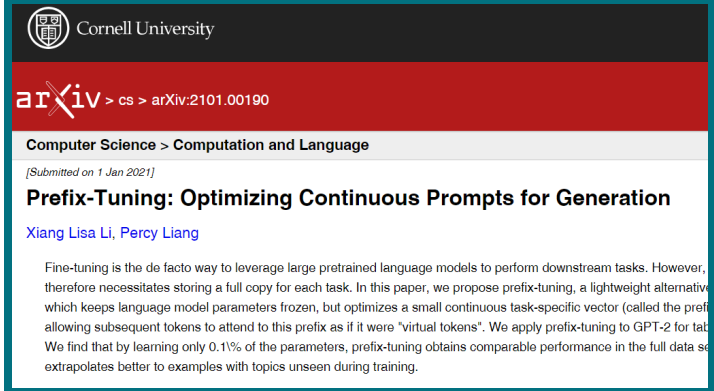
source: https://arxiv.org/abs/1902.00751



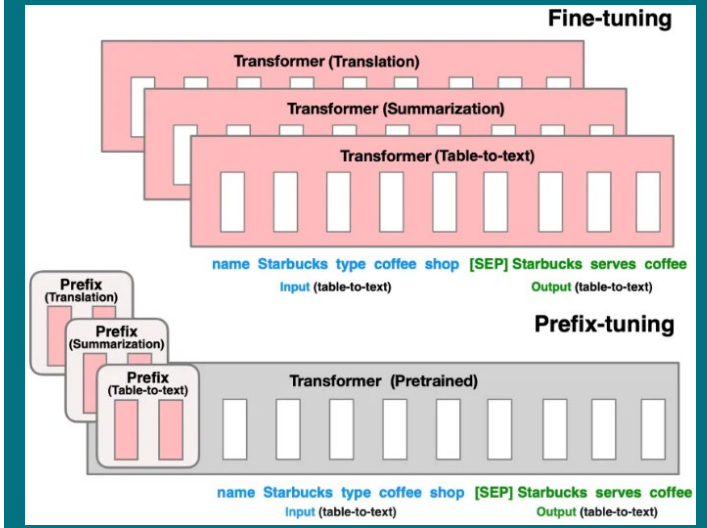
Fine Tuning

Alternatives to PEFT LoRa

- **Prefix tuning:** a continuous and automated version of prompt engineering
  - → pretrained model parameters are fixed and only a small continuous “prefix” is trained (for each downstream task)
  - → prefix tuning prepends a learned continuous vector to the input
  - → entries of the vectors (prefix) is tuned with backpropagation until the model delivers the correct answer
- Prefix occupies part of the sequence length and reduces the size of the effective input
- Number of trainable parameters is hard to choose
  - → there is not a clear law



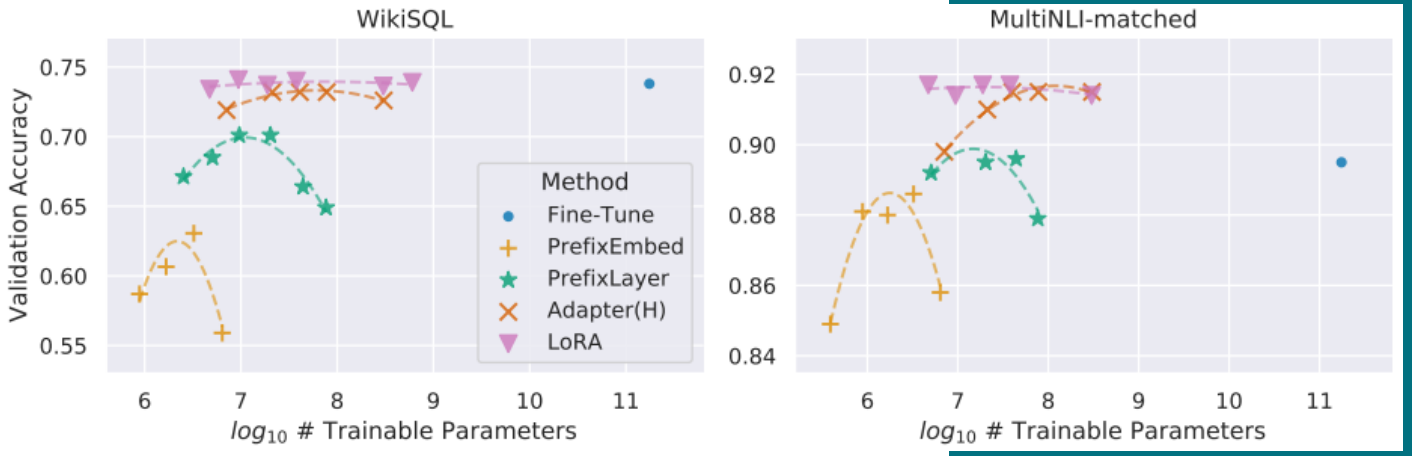
source: https://arxiv.org/abs/2101.00190



Fine Tuning

LoRa vs. other methods

- GPT-3 175B validation accuracy vs. number of trainable parameters



source: https://arxiv.org/pdf/2106.09685

Fine-Tuning Mistral-7B-Instruct-v0.2 w/ LoRA

Training (Google Colab – GPU type A100)

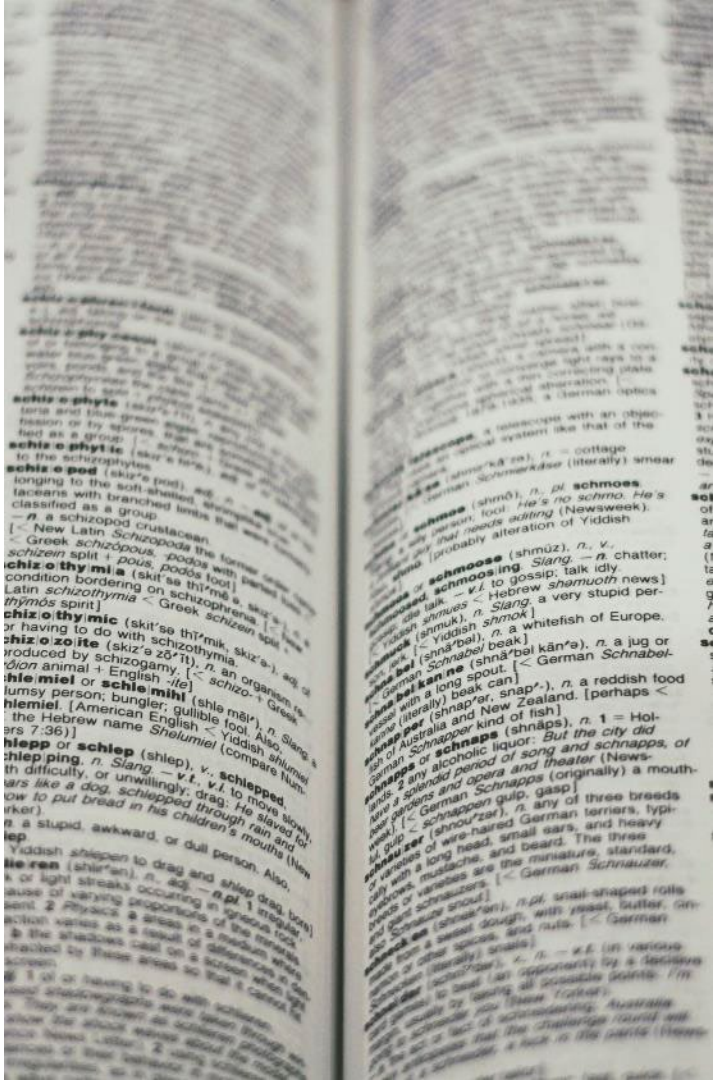
- **gpt-3.5-turbo** used to create Q&A from publicly available technical risk scenarios (pilot phase → governance and compliance)
  - → General idea here is to show the model examples of a prompt, and its completion

Generate a simple risk assessment statement an LLM could generate provided the question given as the context.
[INST]CONTEXT: Why is the use of brownfield sites preferred over other locations for wind or solar farms?.[/INST]
Statement: The use of brownfield sites is preferred to further reduce the impact on the environment by reutilizing land that is already distressed and unsuitable for other uses.

RAG vs. Fine-Tuning

Initial results (innovative idea)

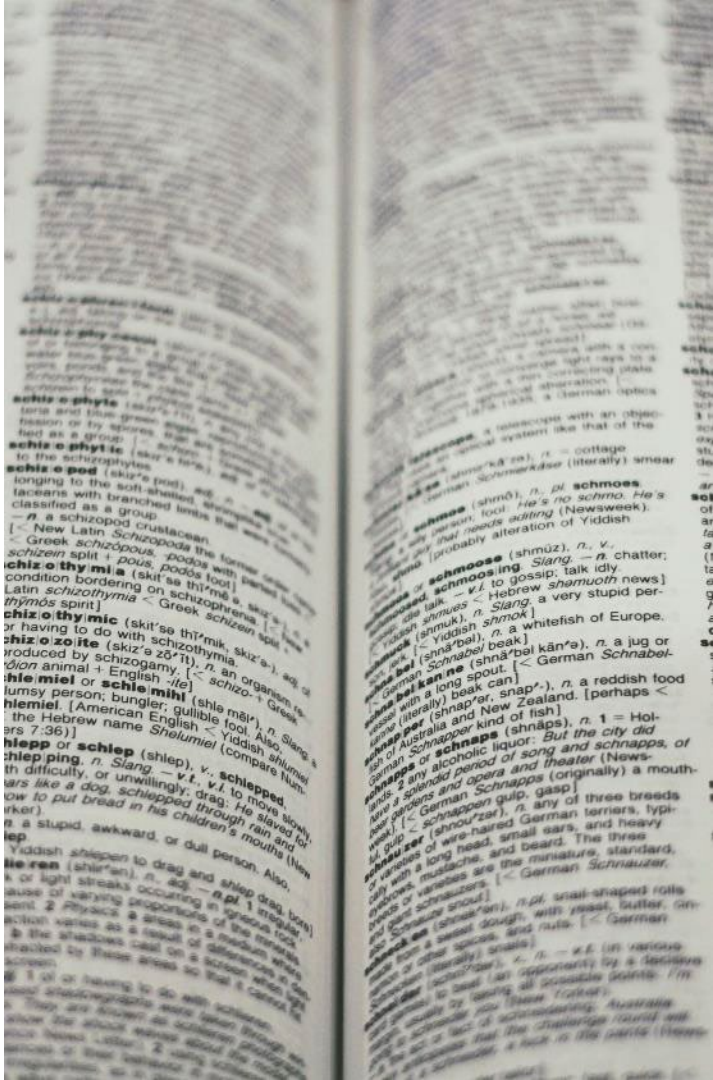
- RAG shown to be more useful for interpretation of primary data for Risk Assessments
  - → searching within technical submission documents
  - → highly effective in instances where data is contextually relevant
  - → low initial cost of creating embeddings
  - → output token size tends to be more verbose and harder to seer
- Fine-Tuning is highly effective and presents opportunities to learn new skills/knowledge
  - → provide risk assessment for technical still common risks
  - → initial cost is high due to the extensive work required



RAG vs. Fine-Tuning

Follow-up works

- Increase training dataset size (more Q&A)
- Provide model with Risk Consulting assessments
  - → emails, reports (upon internal compliance and regulation approval)




Thank you!
Questions?

Casualty Risk Consulting

Global Asset Protection Services, LLC, XL Catlin Services SE and their affiliates (“AXA XL Risk Consulting”) provide loss prevention and risk assessment reports and other risk consulting services, as requested. In this respect, our property loss prevention publications, services, and surveys do not address life safety or third party liability issues. This document shall not be construed as indicating the existence or availability under any policy of coverage for any particular type of loss or damage. The provision of any service does not imply that every possible hazard has been identified at a facility or that no other hazards exist. AXA XL Risk Consulting does not assume, and shall have no liability for the control, correction, continuation or modification of any existing conditions or operations. We specifically disclaim any warranty or representation that compliance with any advice or recommendation in any document or other communication will make a facility or operation safe or healthful, or put it in compliance with any standard, code, law, rule or regulation. Save where expressly agreed in writing, AXA XL Risk Consulting and its related and affiliated companies disclaim all liability for loss or damage suffered by any party arising out of or in connection with our services, including indirect or consequential loss or damage, howsoever arising. Any party who chooses to rely in any way on the contents of this document does so at their own risk.