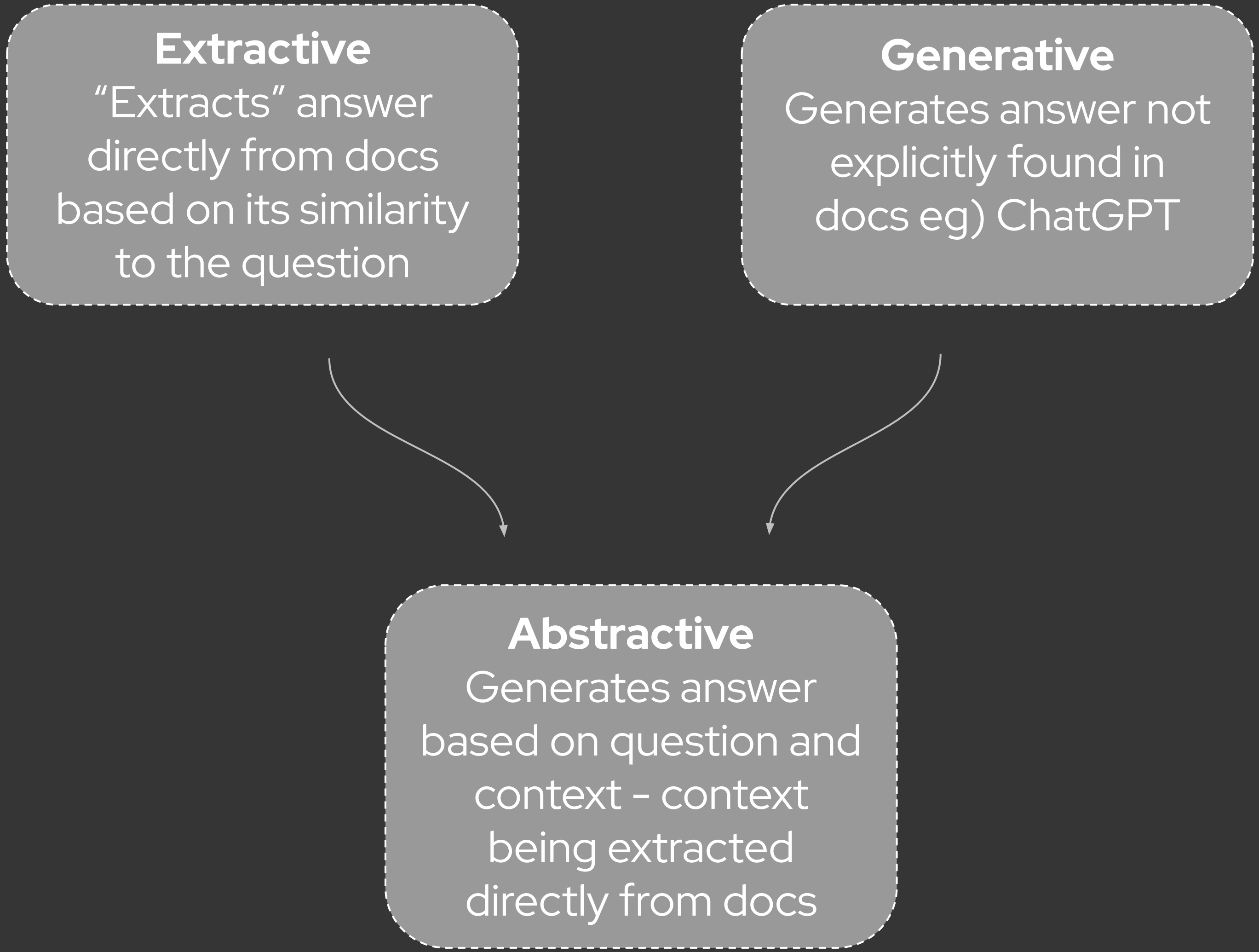
Ask Project Nexodus Docs/Project Aspen

Intern Final Presentation

Christina Xu

Ask Project Nexodus Docs: Leveraging LLMs for Documentation Q&A

Main strategies for QA tasks



Extractive
"Extracts" answer directly from docs based on its similarity to the question

Generative
Generates answer not explicitly found in docs eg) ChatGPT

Abstractive
Generates answer based on question and context - context being extracted directly from docs

What is Fine-tuning

- LLMs are pre-trained on specific domains and tasks such as text generation, question answering, etc. We might want to train the LLM to adapt to our data and task

Adapter Tuning

Add more layers to the pre-trained model and train weights only in those additional layers

https://dataman-ai.medium.com/fine-tune-a-gpt-lora-e9b72ad4ad3

What is Fine-tuning?

Traditional approaches are not practical

- LLMs are trained on specific domains and tasks such as text generation, question answering, etc. We might want to train the LLM to adapt to our data and task

Adapter Tuning

Ac
the r
traini nly
in
lay

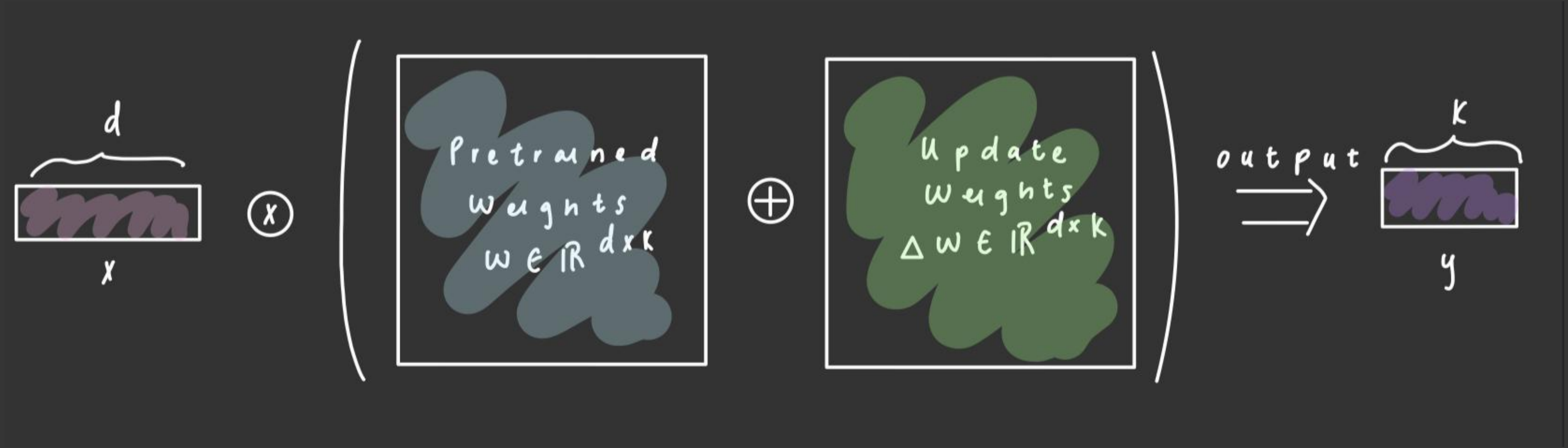
Drawbacks

- Inference latency - the more layers, the longer it takes for the model to generate an answer

https://dataman-ai.medium.com/fine-tune-a-gpt-lora-e9b72ad4ad3

Fine-tuning with LoRA (Low Rank Approximation)

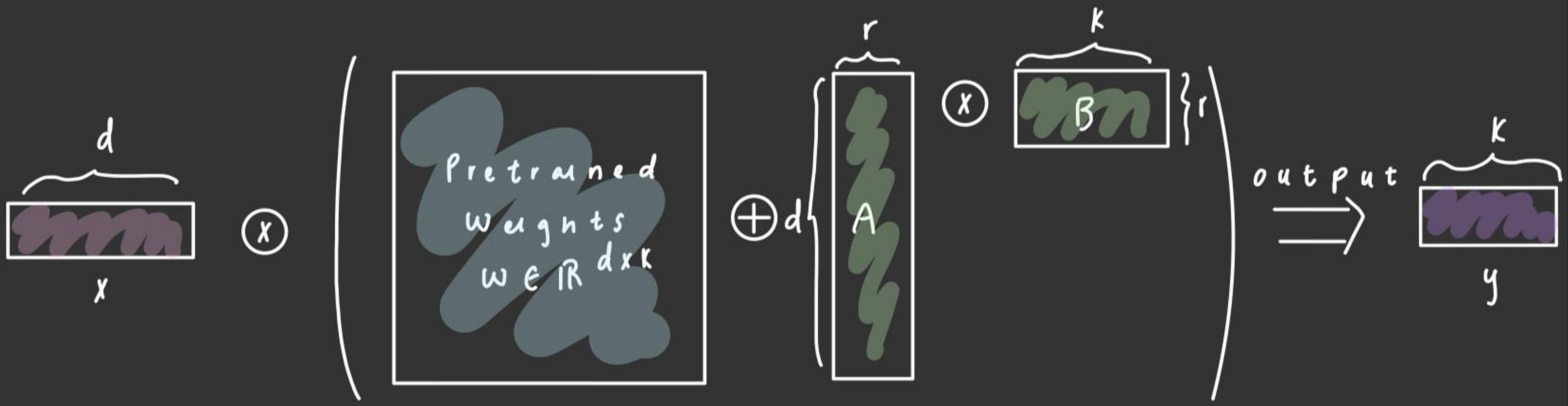
Update pretrained weights in the model


d
x
Pretrained
Weights
$W \in \mathbb{R}^{d \times k}$
Update
Weights
$\Delta W \in \mathbb{R}^{d \times k}$
output
k
y

https://dataman-ai.medium.com/fine-tune-a-gpt-lora-e9b72ad4ad3

During training...

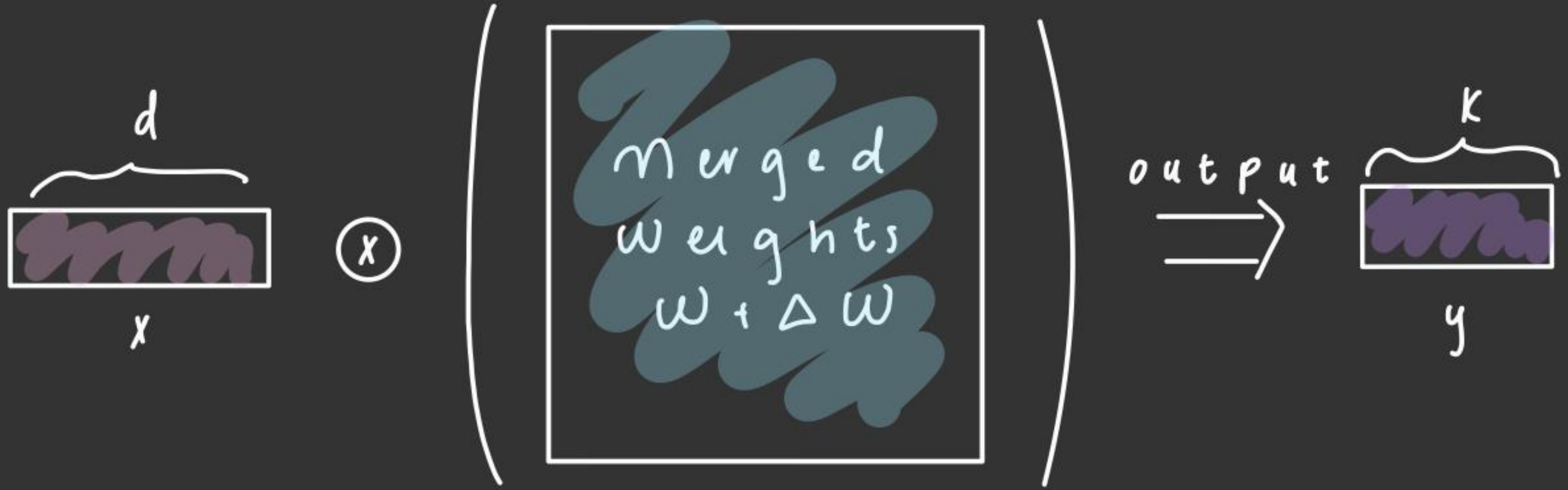
Decompose ΔW into A and B



Consider a 100 x 100 matrix ΔW. That would mean we would have to train 10,000 parameters. If we decompose it into matrices A and B, which are 100 x 1 and 1 x 100, respectively, we only have 100 parameters to train in each or 200 in total

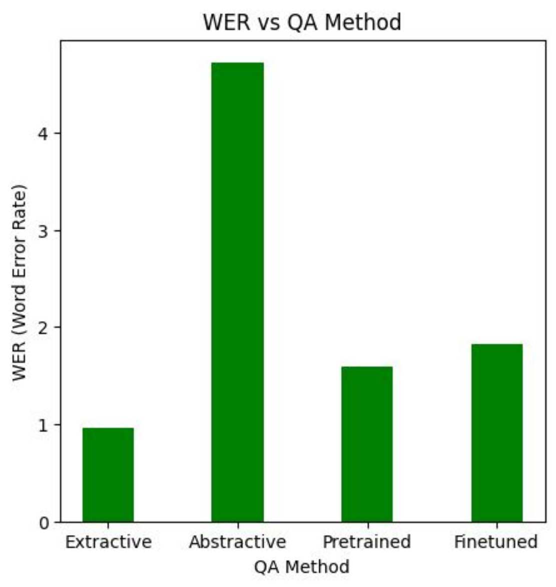
After training...

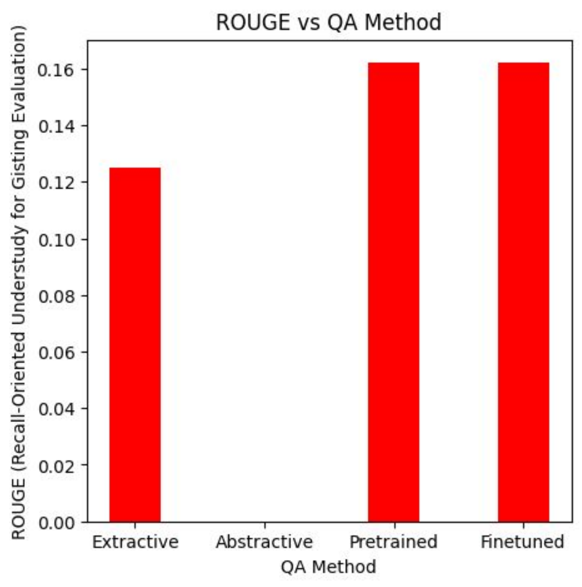
Merge W with ∆W

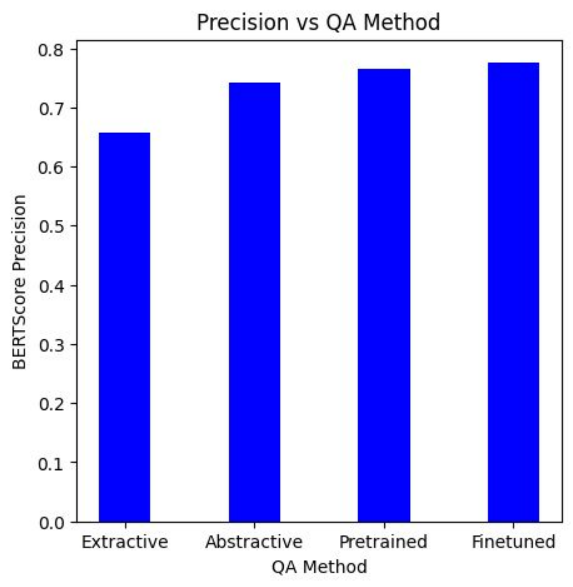


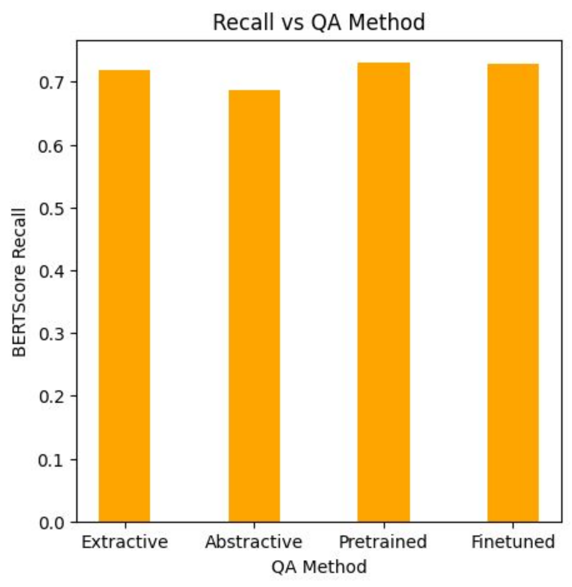
LLM Strategy Evaluation

Key Idea: human language is difficult to quantitatively evaluate

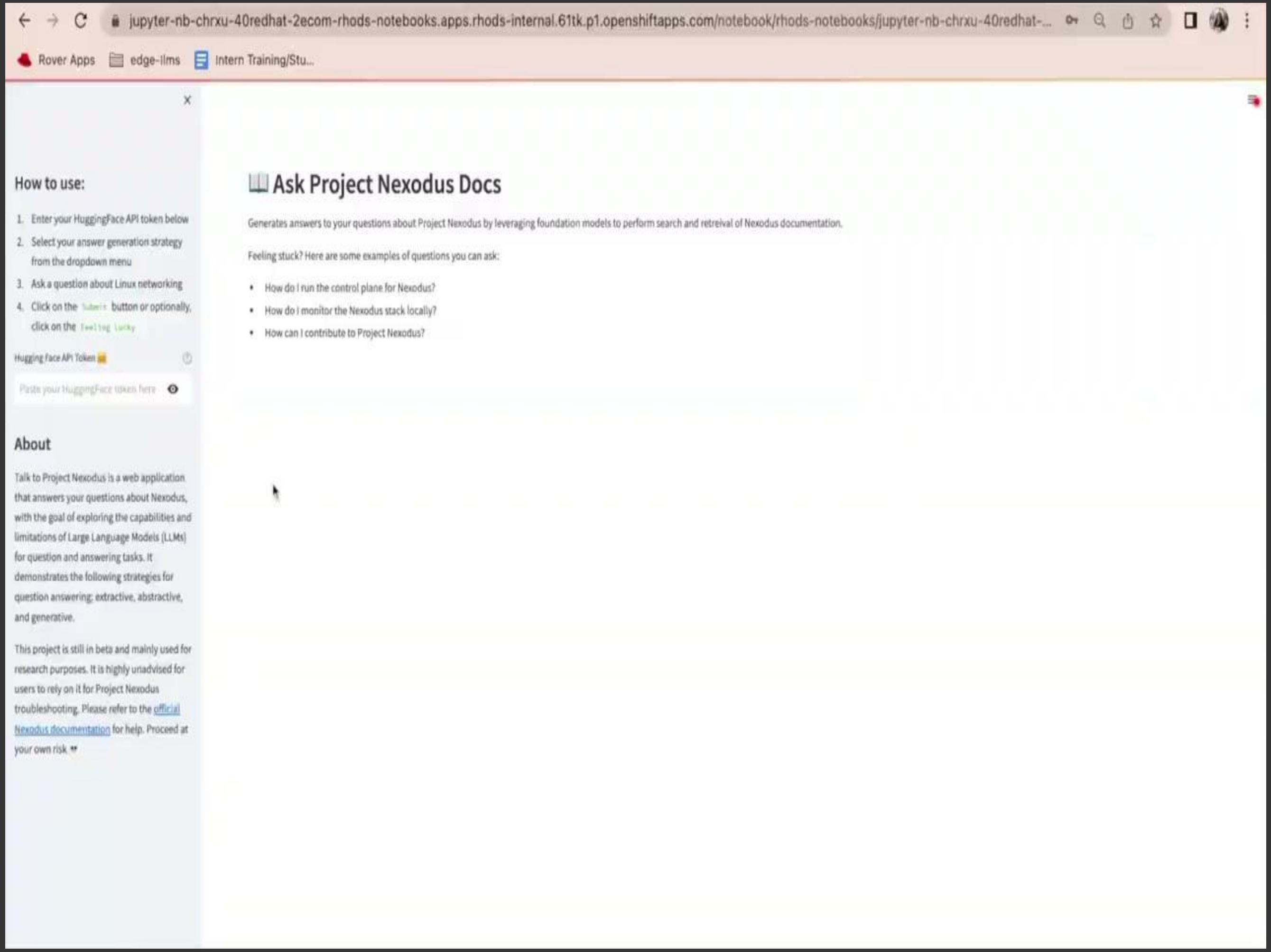








https://huggingface.co/spaces/evaluate-metric/wer
https://huggingface.co/spaces/evaluate-metric/rouge
https://huggingface.co/spaces/evaluate-metric/bertscore



jupyter-nb-chrxu-40redhat-2ecom-rhods-notebooks.apps.rhods-internal.61tk.p1.openshiftapps.com/notebook/rhods-notebooks/jupyter-nb-chrxu-40redhat-...

How to use:

1. Enter your HuggingFace API token below
2. Select your answer generation strategy from the dropdown menu
3. Ask a question about Linux networking
4. Click on the `Submit` button or optionally, click on the `Feeling Lucky`

Hugging Face API Token 🔒

Paste your HuggingFace token here

About

Talk to Project Nexodus is a web application that answers your questions about Nexodus, with the goal of exploring the capabilities and limitations of Large Language Models (LLMs) for question and answering tasks. It demonstrates the following strategies for question answering: extractive, abstractive, and generative.

This project is still in beta and mainly used for research purposes. It is highly unadvised for users to rely on it for Project Nexodus troubleshooting. Please refer to the official Nexodus documentation for help. Proceed at your own risk 👀

📖 Ask Project Nexodus Docs

Generates answers to your questions about Project Nexodus by leveraging foundation models to perform search and retreival of Nexodus documentation.

Feeling stuck? Here are some examples of questions you can ask:

- How do I run the control plane for Nexodus?
- How do I monitor the Nexodus stack locally?
- How can I contribute to Project Nexodus?

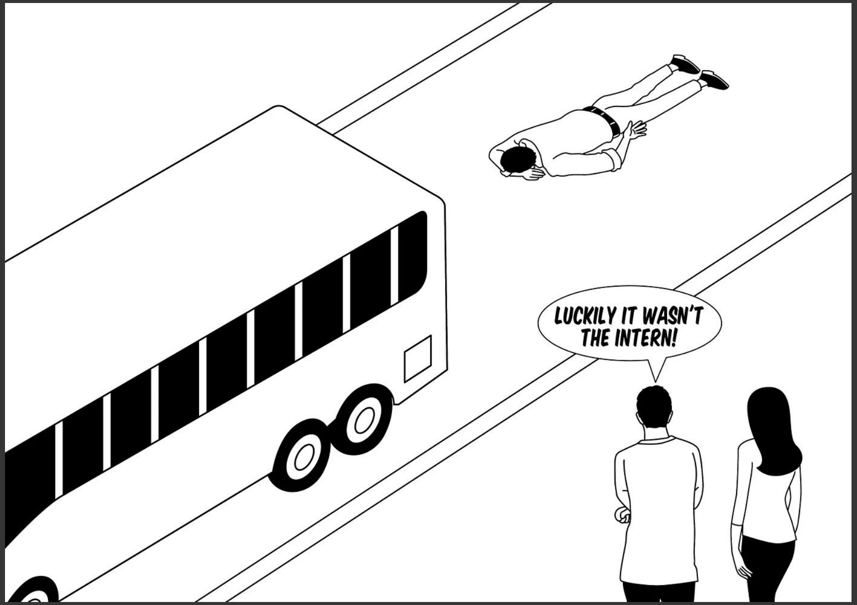
Project Aspen: Bus Factor

What is Project Aspen?

Analyzes data from open source projects to empower contributors and participants to make data driven decisions about open source communities and projects.

Bus Factor

How high the risk is to a project should the most active people leave?

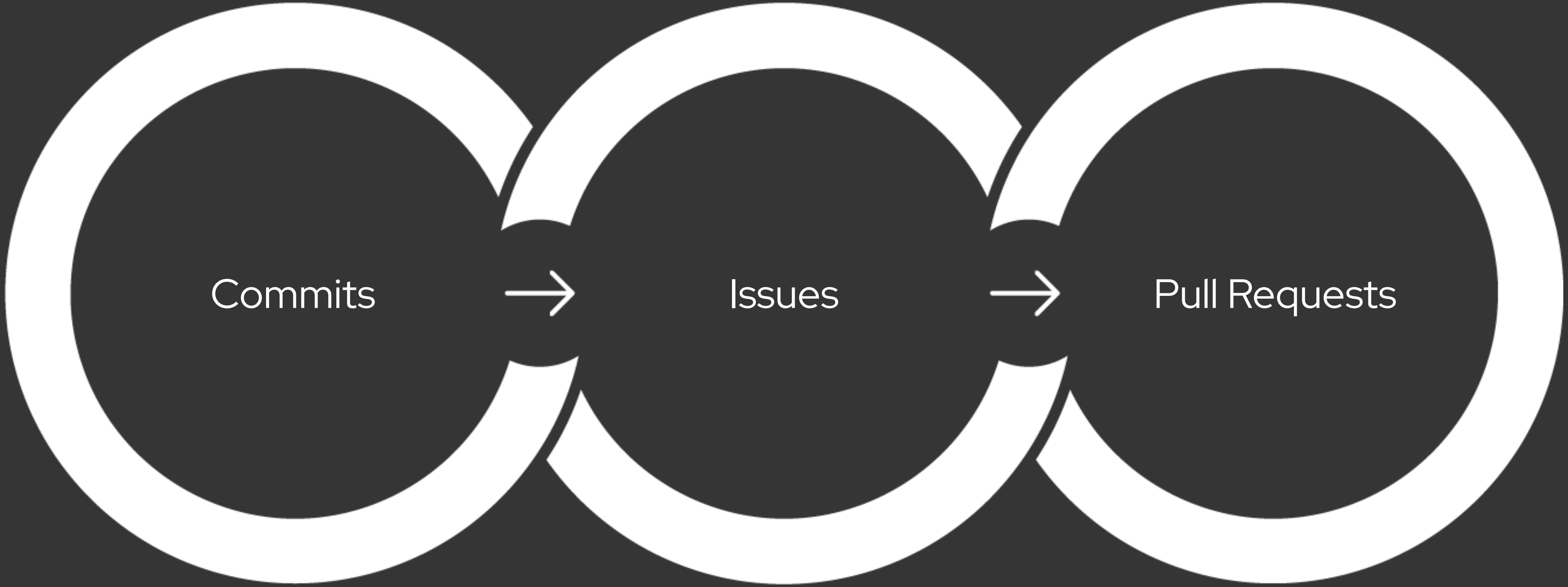


- Quantifies the amount of contributors a project can afford to lose before it stalls by hypothetically having these people get run over by a bus
- Typically, it is the smallest number of people that make up 50% of contributions

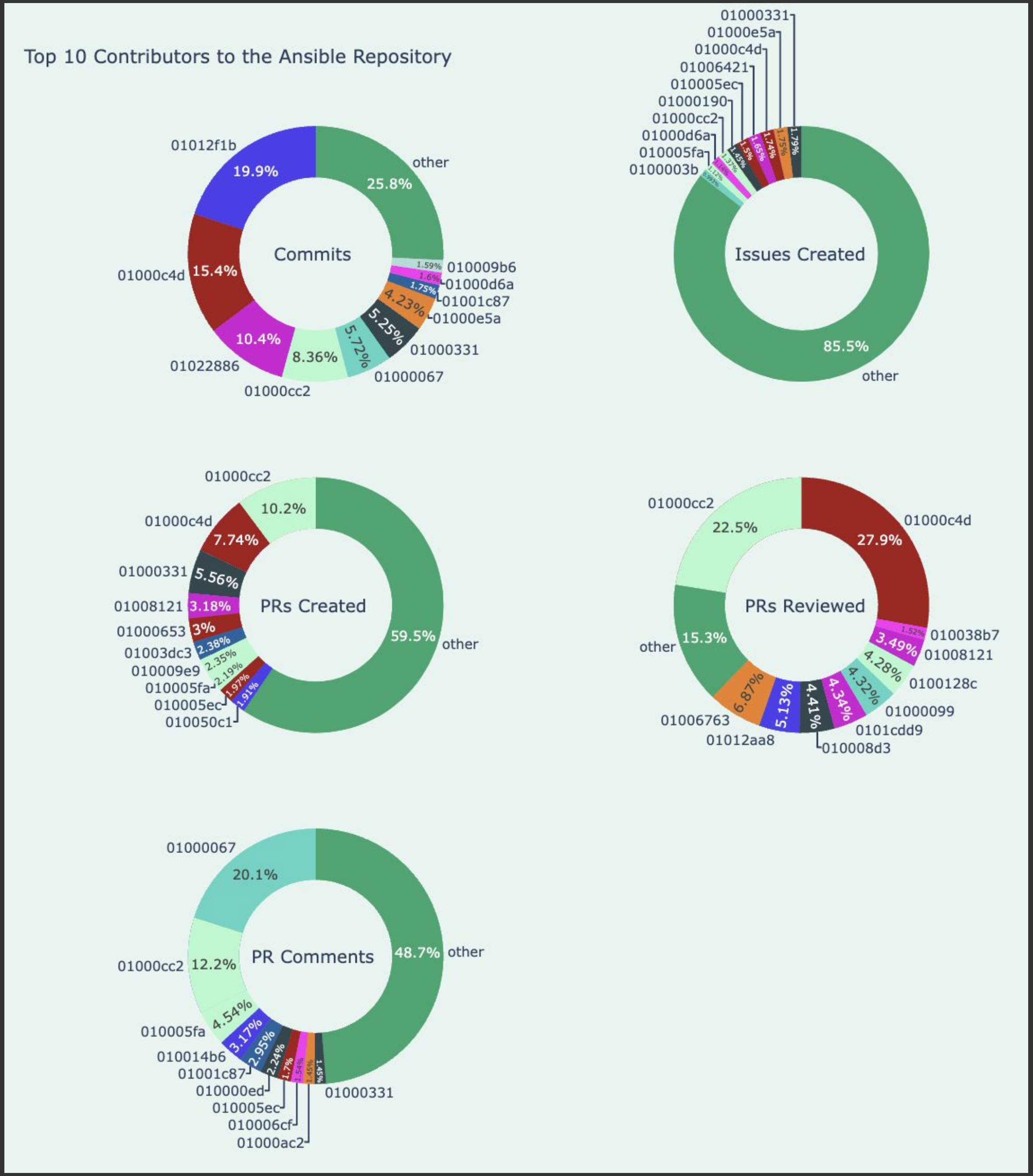
https://chaoss.community/kb/metric-bus-factor/

How do we define “contributions”?

We can analyze bus factors from different perspectives



Commits → Issues → Pull Requests



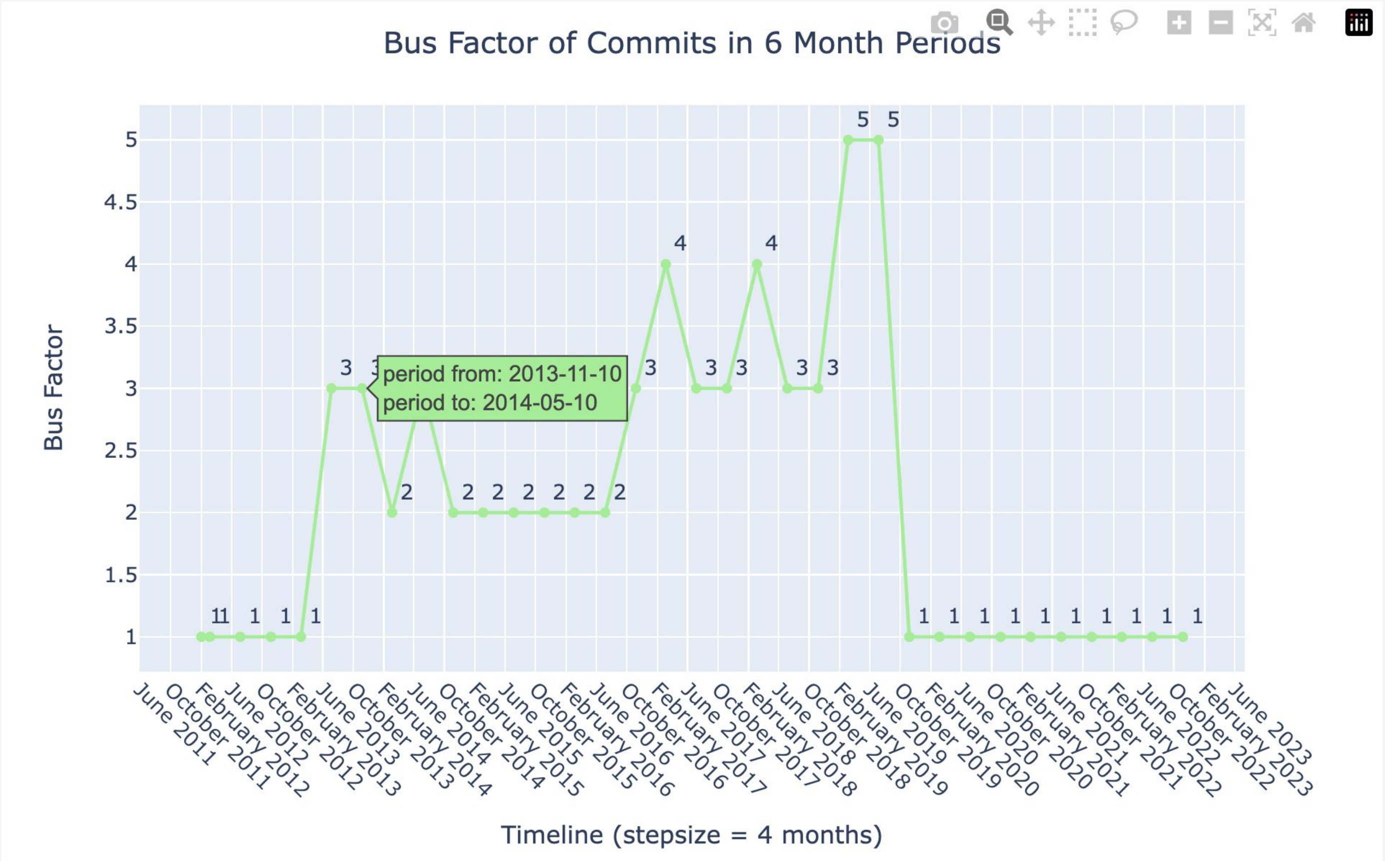
Key Insights

- There appears to be a trend in the top 10 contributors across all perspectives

  eg) 01012f1b, 01000c4d, 01000cc2

- The proportion between the top 10 and 'other' contributors for each perspective matches our intuition

Bus factor as a function of time



Bus Factor of Commits in 6 Month Periods

Thank you

Special thanks to Sanjay Arora, James Kunstle, Heidi Dempsey, Jen Stacy, and my fellow research interns

For questions or concerns regarding my projects, feel free to reach out to me via: chrxu@redhat.com

Let's connect!

https://github.com/oss-aspen/Rappel

https://github.com/christinaexyou/ask_project_nexodus_docs (WIP)

https://www.linkedin.com/in/christinaexyou/

https://medium.com/@christinaexyou