**DataSci 346 Project Rough Script:**

**Slide 1: Intro (Arlo) Quick**

Introduce self, CEO & Company “ClosedAI”. People have been waiting anxiously for our next product.

Introduce CEO(chief ethics officer) Stephan Delport and CSO (chief scarping officer).

**Slide 2: Intro 2 (Arlo) Quick**

Show market share and company stonks of ClosedAI, OpenAI, Tesla, Apple and Samsung. Closed AI is taking the world by storm.

Quick summary of ClosedAI, we want to have a global monopoly of wealth and our end goal is to replace humans with AI. Coming in about years.

And our newest invention, what you all been patiently waiting for:

**Slide 3: Intro of Model (Arlo) Quick**

‘CHAT ‘n bietjie bol’. Stands for CHAT Now By Expecting Text, Just …TODO

This is a NS LLM, a brand new type of LLM model. It stands for ‘Not So Large Language Model’. And this is an innovative science that allows you to run an LLM to answer all your data science, ML or AI related questions…

But before we get into this, the HR tannie has been on my ass. Something about “Data Ethics”, have no idea what that means, and, “Legal Obligation to be ethical” or something. I hand you over to my chief ethics officer Stephan delport.

**Slide 4: Scraping (Andre) Medium**

Explanation of where and how data was scraped. Joke about all of the effort and restrictions of scraping public domains and how much effort it is to get around it.

What type of data is scraped

How its done, Python libraries used etc.

Show how many rows were scraped and in what format

**Slide 5: NLP (Andre) Quick**

NLP done for Scraping etc.

**Slide 6: Ethics (Stephan) Long**

Where data is scraped, what data is scraped…why data is scraped (to cater for our specific model)

Wthical principles we used…

All other ethical bullshit.

Sarcastic Jokes about ethics and that we ‘we forced to implement these ’ as AI cant take over humans yet

Handing over to the Chief scraping officer (Andre) Medium

**Slide 7: Llama Model (Arlo): Long**

Speak about Llama model. Training time, amount of parameters. Complexity. Runtime. Model architecture and brief overview of how it works.

3 billion params etc.

**Slide 8: Tokenization, Quantization (Stephan) Medium**

What it is, how its implemented, how it works etc..

**Slide 9: UnSloth Optimization (Andre) Medium**

What it is, how it works, memory distribution to allow laptops to run it, optimization

**Slide 10 Kaggle Cloud to Train (Arlo)**

Info on Kaggle Cloud

**Slide 11: The model in Practice (Arlo, Stephan, Andre) Medium**

Can un on and be trained on a laptop. Can answer all question needs etc.

Show practical inference example. Then ask class for their own prompts to use on ur own model.

Model 1: Arlo  
Model 2: Andre  
Comparison

Stephan prompt demo. Still in beta..no gui. (forget definition…ask model)…also ask class for prompt

**Slide 12: Conclusion 1: We are hiring (Arlo, Stephan and Andre) Quick**

Arlo says we are hiring, we recently had a new position open up, what was the guys name that left again?? It was Evan right?  
Andre: No his name was Elon Musk

Arlo: Ah yes, why did he leave again?

Stephan: I believe he jumped off a building

Arlo: Ag ja shame he’ll be missed,he was a good ou, he clearly couldn’t handle our incredibly intensive working hours.

Slide 4:

I’m quite confused why I was tasked with scraping our data after I received 0% for the webscraper project in Computer Science 343, but here we are.

What is cross validated and that it is part of the Stack Exchange network.

Ever-expanding field of data science and not at all easy to keep up with all of the new ideas and implementations in the machine learning and statistics fields. Cross validated can be used as a platform to ask questions and find:

Bullets

Slide 5:

Objective was to gather insights of common challenges and solutions in data science.

We scraped the 39668 highest voted questions and answer pairs, where each answer was each respective question’s most voted answer.

Only the content body of each question and answer was scraped.

Questions or answers that contained code blocks were ignored, meaning that our model can’t generate blocks of code. This was intentional however, because we had to think of other AI models such as ChatGPT and Gemini as they probably would have had to shut down if our model could also produce code.

Used Beautifulsoup to scrape the data and used an exponential backoff function for rate limiting

Slide 6:

Now time for everyone’s favourite part… Wrangling. Spending hours upon hours of sifting through almost 40000 questions and answer pairs to determine what was important and what should be left out.

We removed post notices contained in the body of questions that were either locked or closed, retained important HTML tags which contained relevant text and then stripped these remaining HTML tags to extract plain text.

Some of the characters were also encoded as Unicode characters, which then has to be converted to either HTML or ASCII characters. Each question and answer pair was then structured into a list of dictonaries, which categorised a role of users for questions and a role of assistant for answers. This list was then exported a new JSON file.

Slide 12:

We had to finetune our model.

Finetuning is the process of updating the actual "brains" of the language model through back-propagation. But, finetuning can get very slow and very resource intensive.

We used UnSloth optimisation to finetune our model

What is Unsloth?

A finetuning framework designed to enhance productivity by streamlining the process of adapting large language models like Llama-3.2. UnSloth:

* Optimises memory usage
* Accelerates training time
* Automates hyperparameter tuning

**How UnSloth Optimization Works:**

* **Layer-Freezing:** Only fine-tune specific layers, reducing computation.
* **Gradient Checkpointing:** Lowers memory usage by recomputing activations.
* **Dynamic Hyperparameter Tuning:** Automatically adjusts settings during training.
* **Efficient Data Loading:** Speeds up training with optimized data handling.

**Benefits of UnSloth in Our Llama-3.2 Model:**

* Fine-tuning is **2 times faster** compared to traditional methods.
* Utilizes **70% less memory**, optimizing resource efficiency.
* Maintains **model accuracy**, ensuring high performance without compromise.