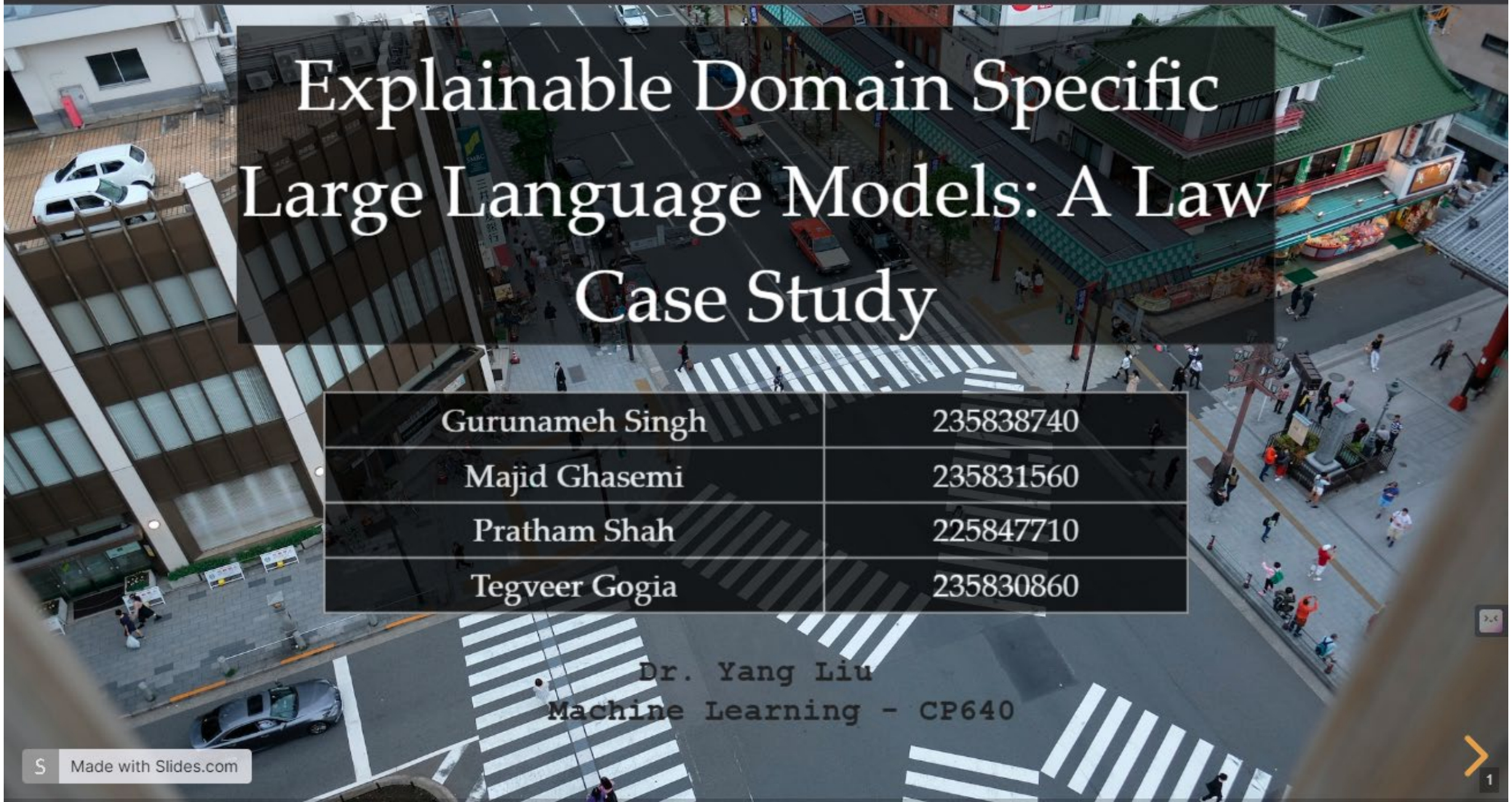


Explainable Domain Specific Large Language Models: A Law Case Study

Gurunameh Singh	235838740
Majid Ghasemi	235831560
Pratham Shah	225847710
Tegveer Gogia	235830860

Dr. Yang Liu
Machine Learning - CP640

AGENDA

1. Large Language Models (LLMs)
2. Explainable AI
3. Developed Model
4. Results
5. Conclusion

Structured data

Text

Voice

3D signals

Images

Large Language Model

Training

Adaptation

Information extraction

Instruction following

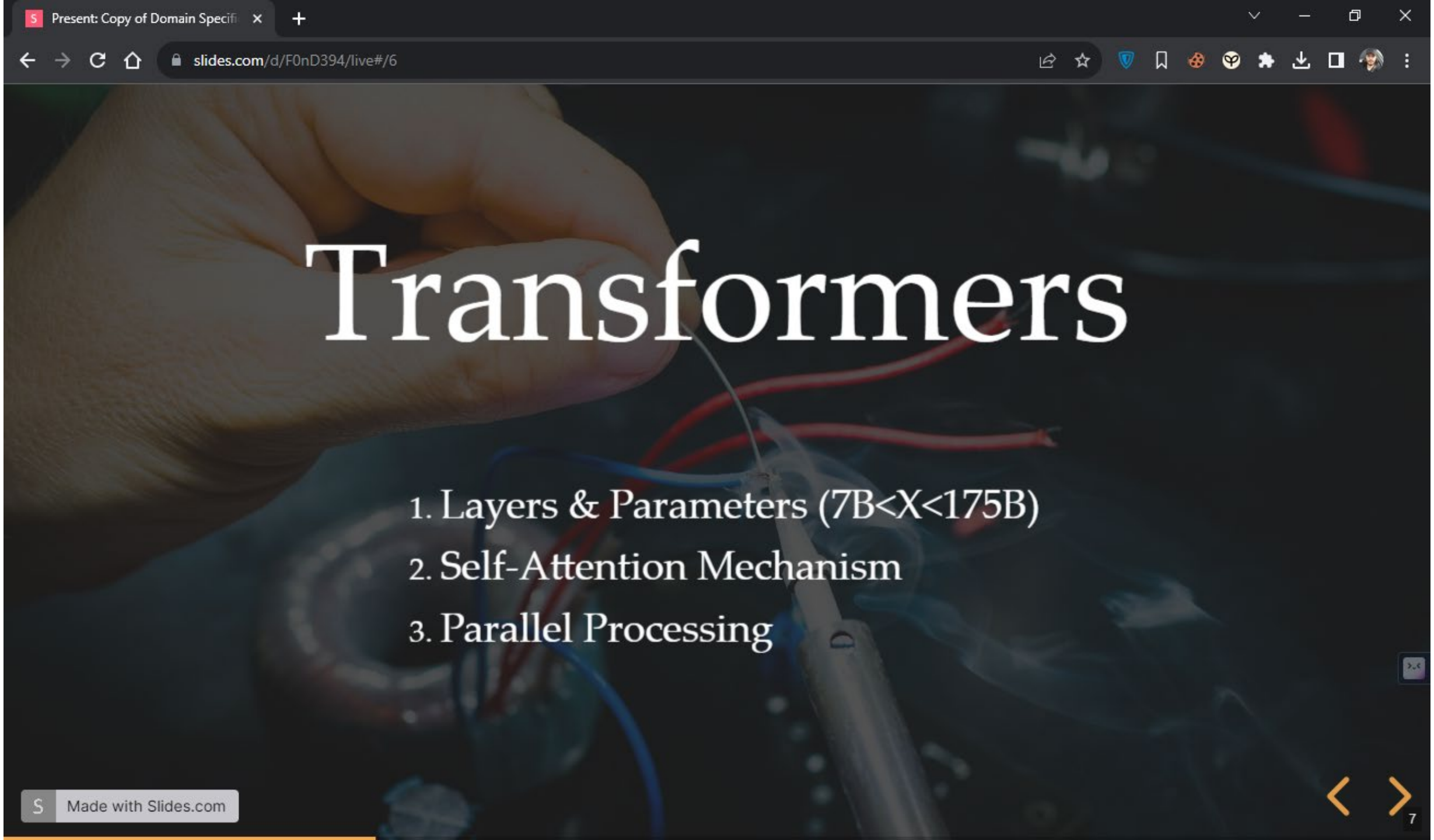
Object recognition

Image captioning

Q&A

Sentiment analysis

Architecture of LLMs


Present: Copy of Domain Specifi
slides.com/d/F0nD394/live#/6
Transformers
1. Layers & Parameters (7B<X<175B)
2. Self-Attention Mechanism
3. Parallel Processing
Made with Slides.com
7

Pre-Training

The model is trained on a large, diverse dataset of text to learn a general understanding of language. This involves unsupervised learning, often using tasks like masked language modeling.

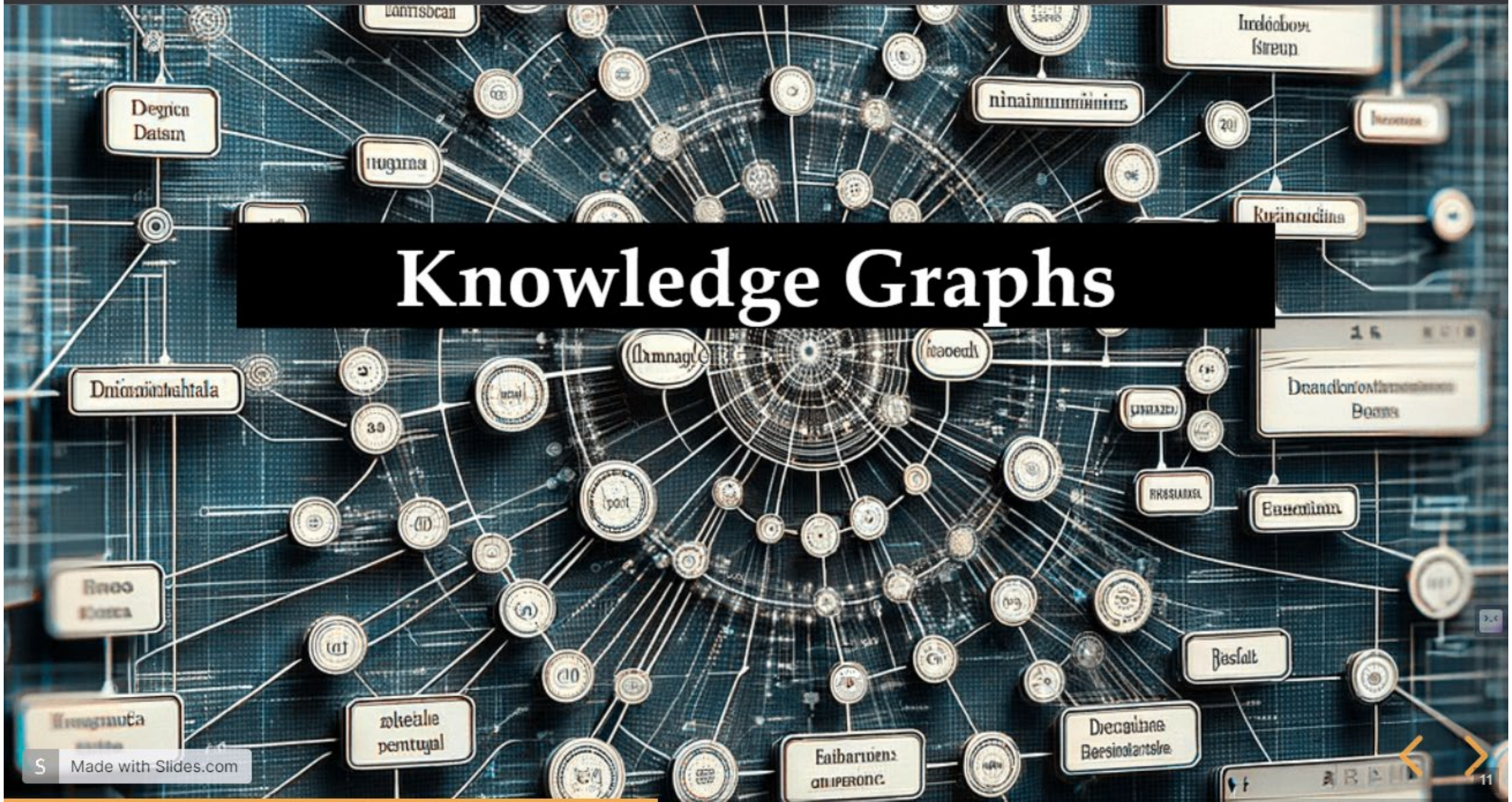
Fine-Tuning

The pre-trained model is further trained on a smaller, task-specific dataset. This step adapts the model to specific applications, such as translation or sentiment analysis.

LLMs Training

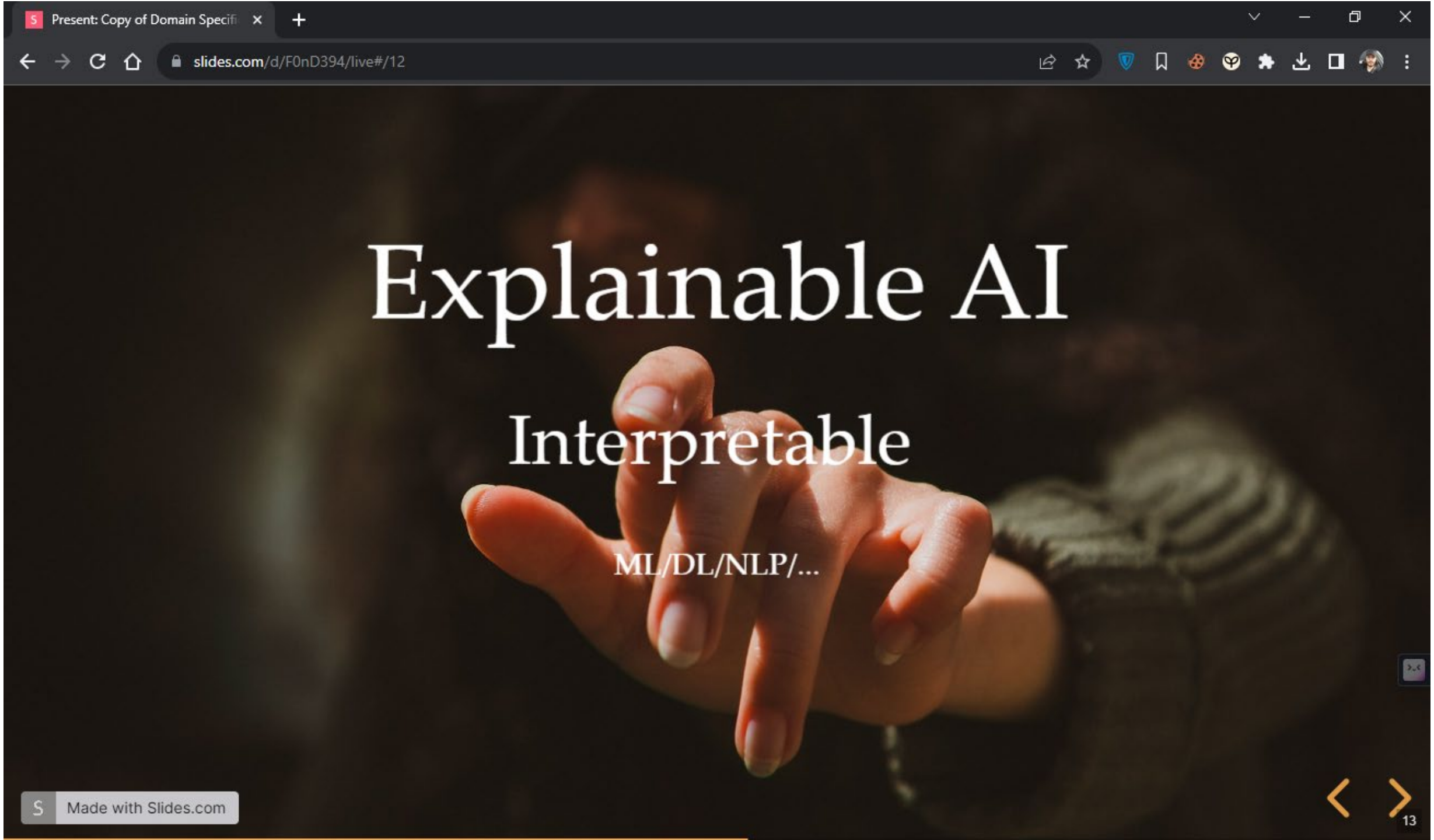
Downwards

- Computational and Environmental Cost
- Data and Bias Concerns
- Maintenance and Updating
- Scalability and Accessibility
- **Interpretability and Explainability**



LLMs vs Knowledge Graphs

- **Data Handling**
- **Output**
- **Flexibility vs. Accuracy**
- **Use Cases**


Present: Copy of Domain Specifi
slides.com/d/F0nD394/live#/12
Explainable AI
Interpretable
ML/DL/NLP/...
Made with Slides.com
13

XAI Methodologies in LLMs

- Local/Global Explanation
- Visualization
- Counterfactual Explanation
- **Natural Language Explanations**

Utilized Dataset

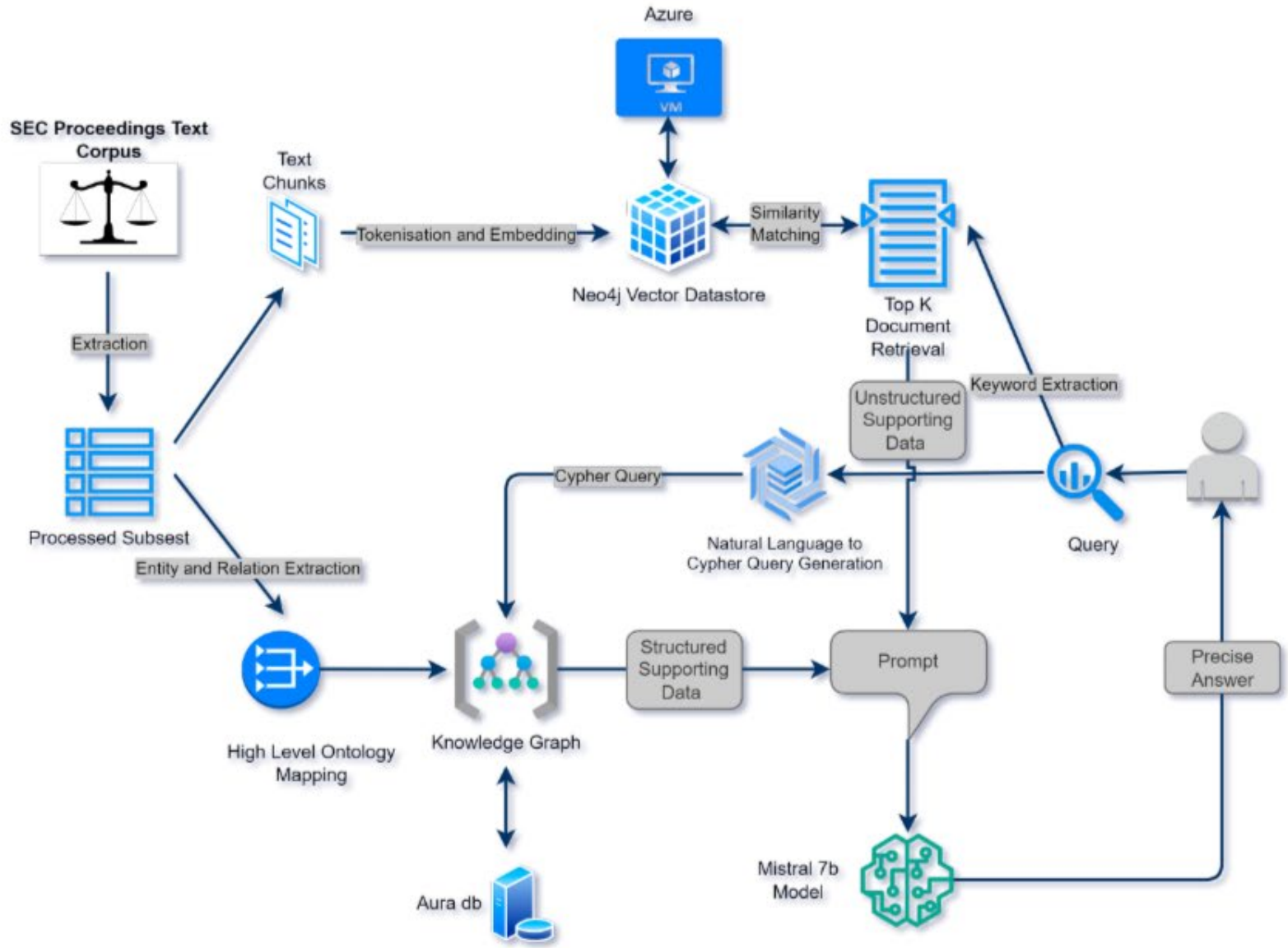
SEC Website

- Fraud
- Insider Trading
- Misappropriation of Funds
- Unregistered Brokers

Link to the dataset:

https://github.com/AnjaneyaTripathi/knowledge_graph

The Developed Model



Video’s Link

https://www.youtube.com/watch?v=l2R66CndT9I

Evaluation

Results

Semantic Similarity Score	Estimated Precision	Estimated Recall	Estimated F1-Score
0.73	0.63	0.73	0.68

Human Evaluation

70% of the responses generated by the model were accurate

17% of the responses generated by the model were moderately accurate

13% of the responses generated by the model were not entirely accurate

Conclusion