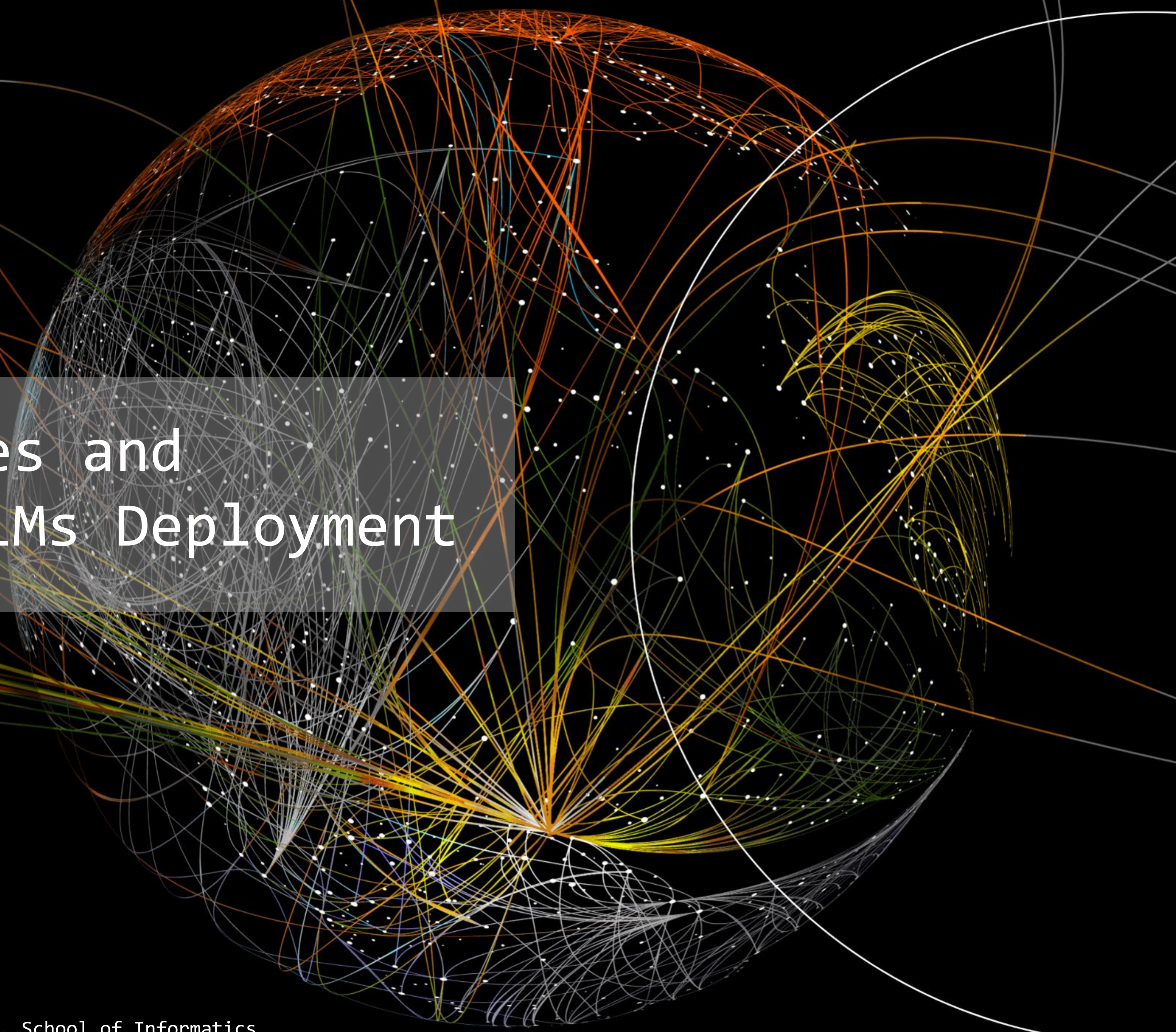


Navigating Challenges and Technical Debt in LLMs Deployment

Ahmed Menshawy, VP of AI Engineering, Mastercard
Zeeshan Nawaz, Principal AI Engineer
Mahmoud Fahmy, Senior AI Engineer
Pasquale Minervini, Researcher at the University of Edinburgh, School of Informatics

Essentials for building a **generative AI application**

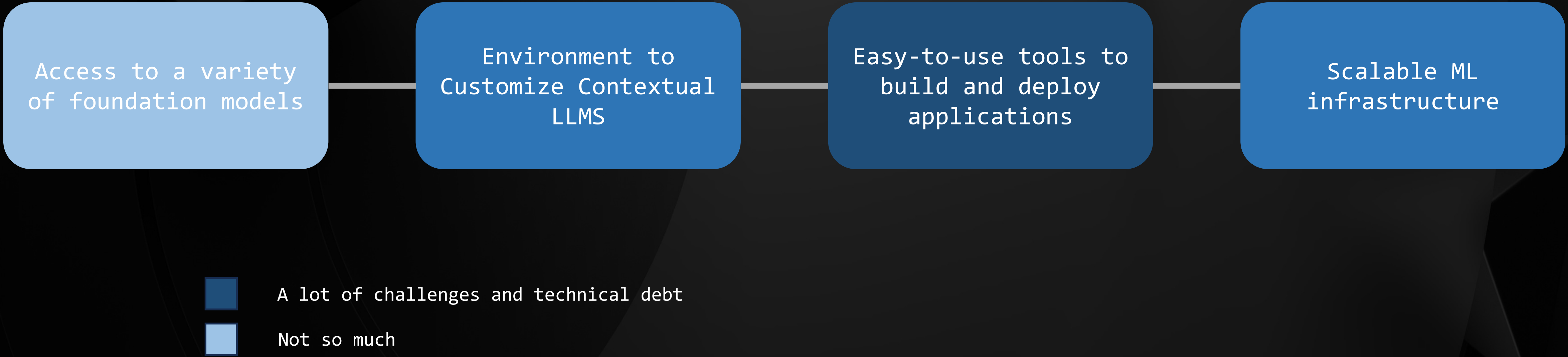
Access to a variety of foundation models

Environment to Customize Contextual LLMS

Easy-to-use tools to build and deploy applications

Scalable ML infrastructure

Essentials for building a **generative AI application**



Access to a variety of foundation models

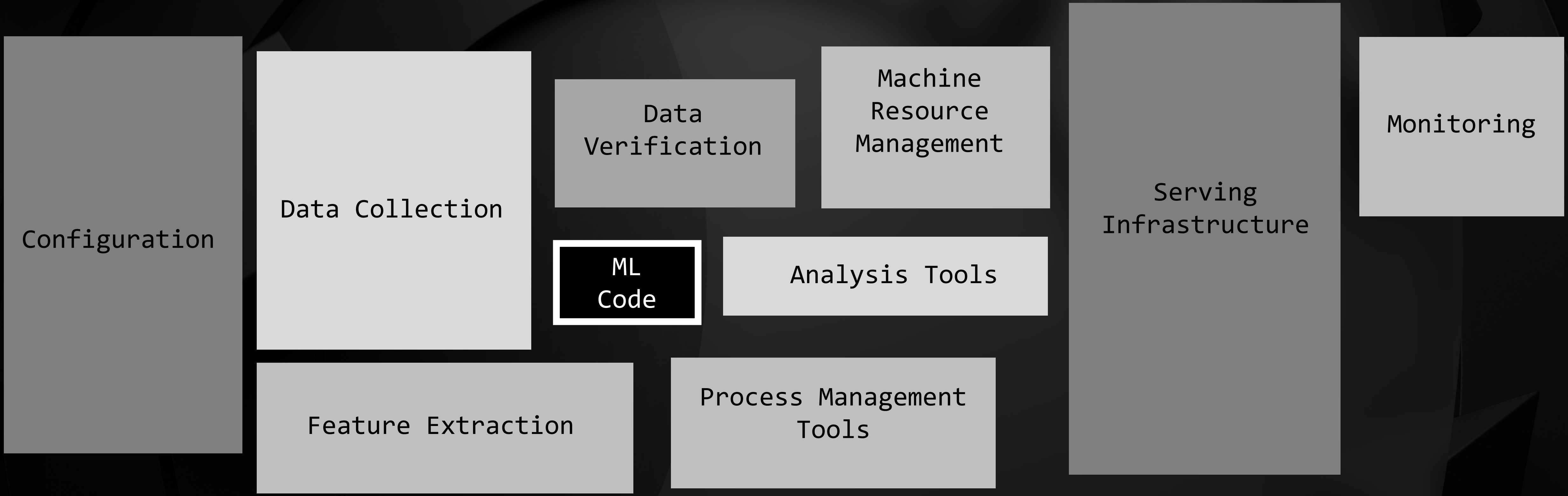
Environment to Customize Contextual LLMS

Easy-to-use tools to build and deploy applications

Scalable ML infrastructure

■ A lot of challenges and technical debt

□ Not so much



Configuration

Data Collection

Data Verification

Machine Resource Management

ML Code

Analysis Tools

Feature Extraction

Process Management Tools

Serving Infrastructure

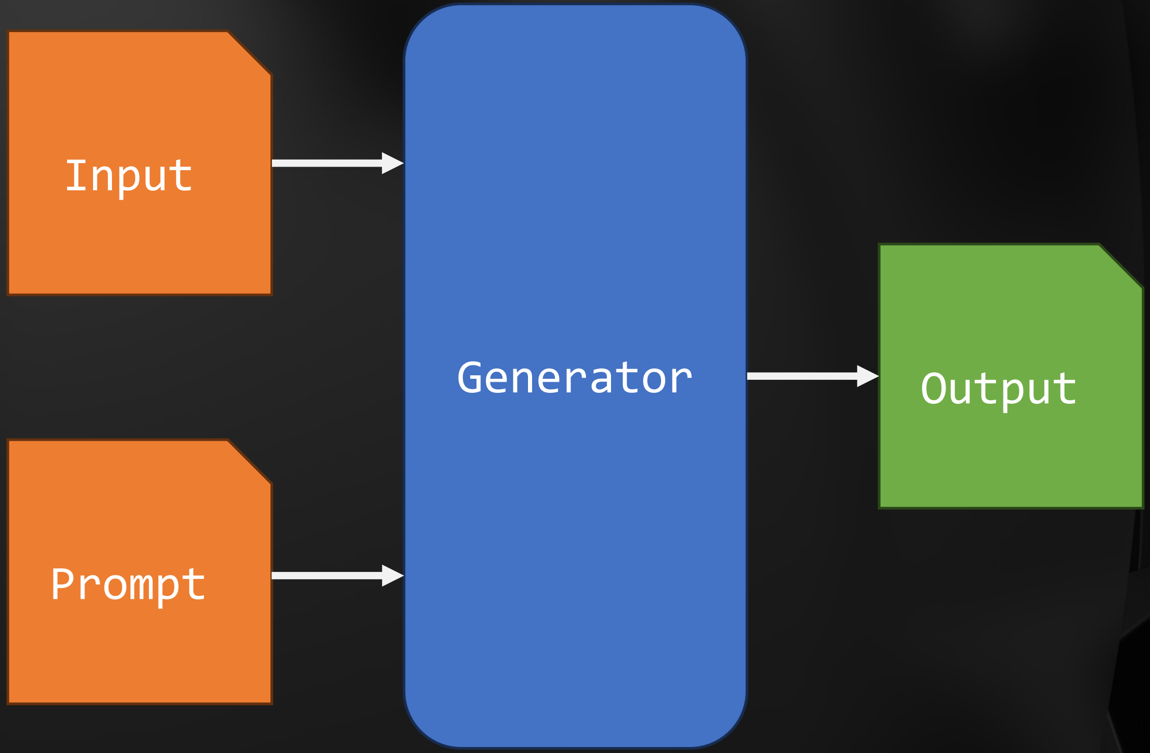
Monitoring

Vast and complex Surrounding challenges and technical debt around ML deployment*

* D. Sculley et al. 2015. Hidden technical debt in Machine learning systems. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15). MIT Press, Cambridge, MA, USA, 2503–2511.

Challenges with Closed-book(Non-parametric) Approach

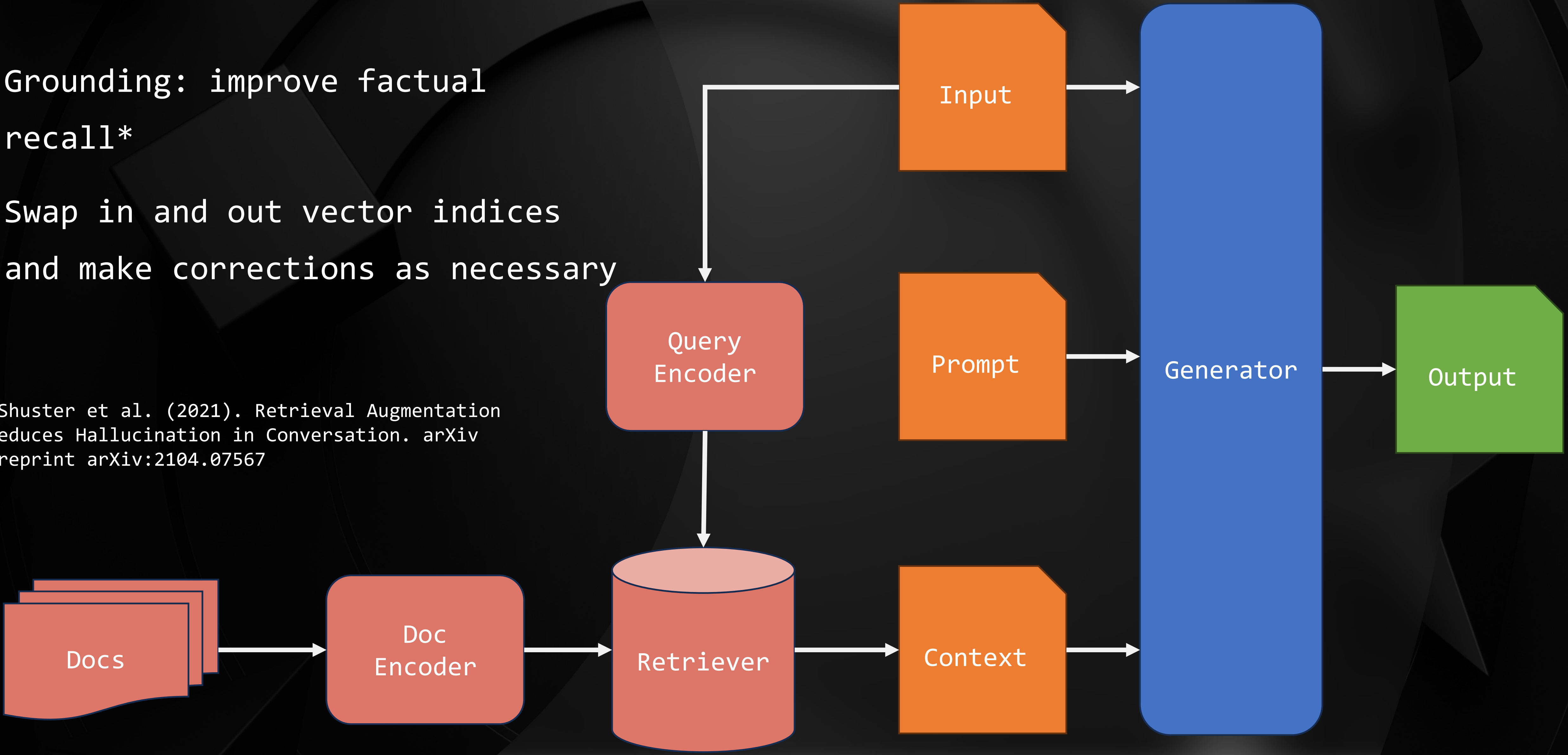
- Problems:
  - Hallucination
  - Attribution
  - Staleness
  - Revisions
  - Customization
- Solution:
  - Couple to external memory



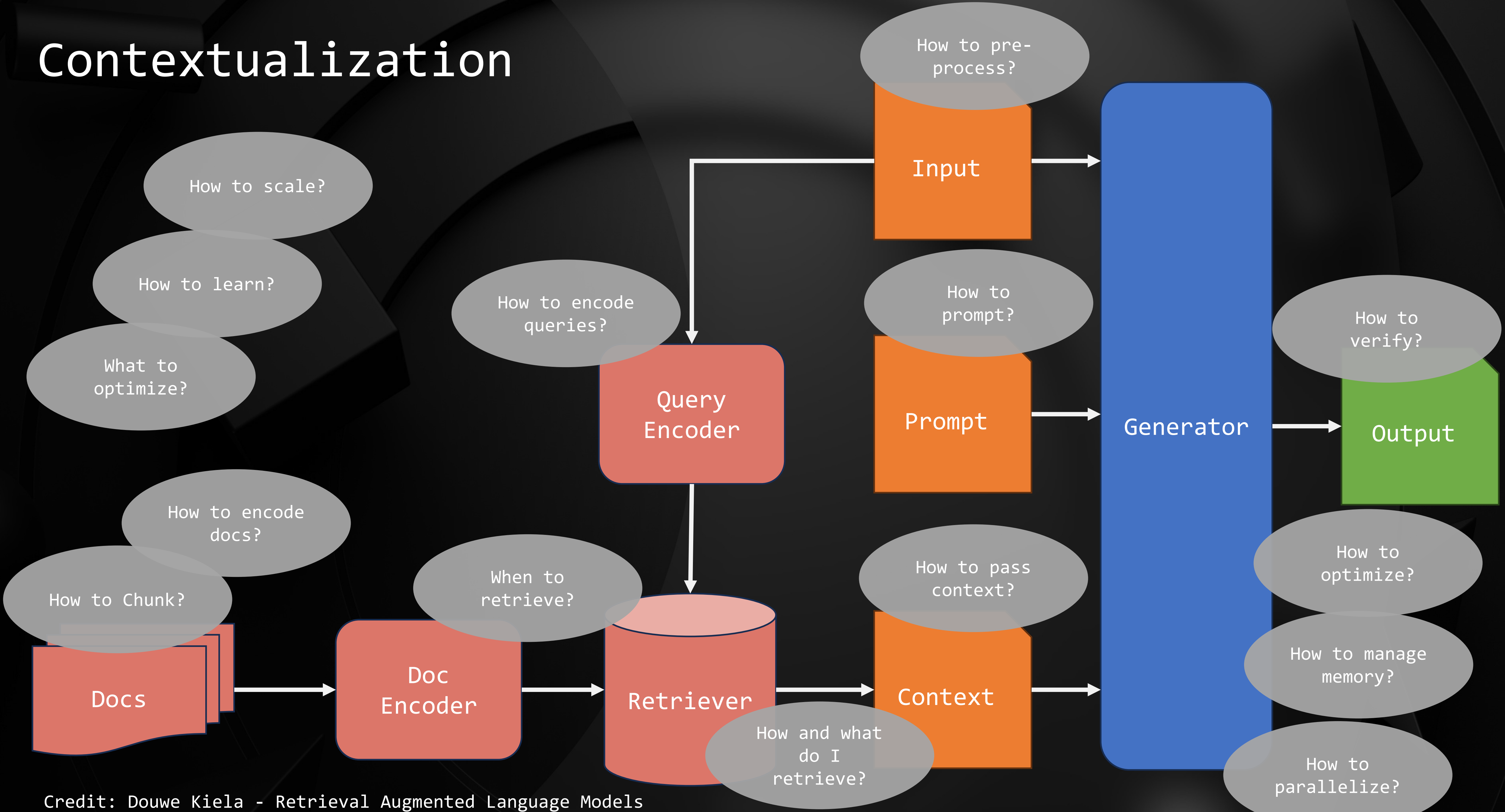
Contextualization: Open-Book (Non-parametric approach)

- Grounding: improve factual recall*
- Swap in and out vector indices and make corrections as necessary

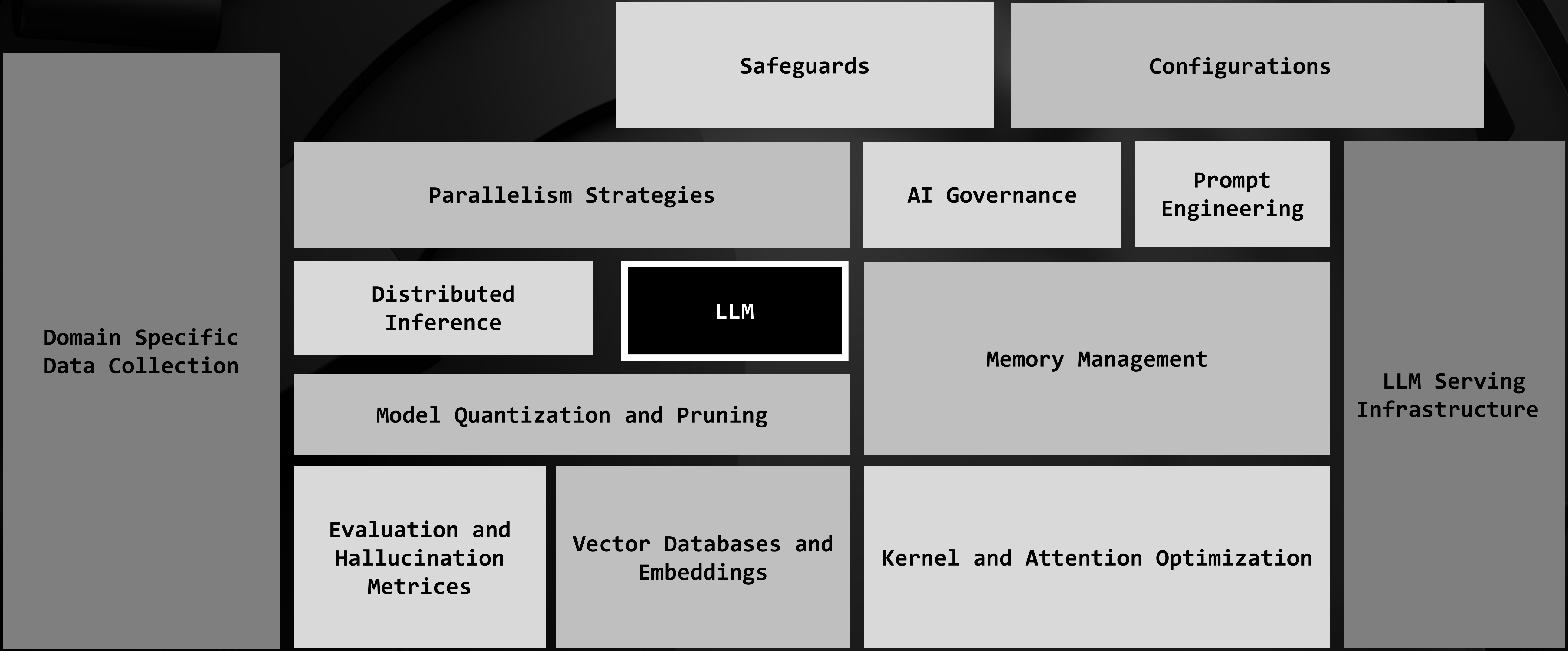
*Shuster et al. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. arXiv preprint arXiv:2104.07567



Contextualization



Credit: Douwe Kiela - Retrieval Augmented Language Models



Safeguards

Configurations

Domain Specific Data Collection

Parallelism Strategies

AI Governance

Prompt Engineering

LLM Serving Infrastructure

Distributed Inference

LLM

Memory Management

Model Quantization and Pruning

Evaluation and Hallucination Metrices

Vector Databases and Embeddings

Kernel and Attention Optimization

Vast and complex Surrounding challenges and technical debt around LLM deployment*

* Menshawy, A., Nawaz, Z., Fahmy, M., & Minervini, P. (2024). Navigating Challenges and Technical Debt in Large Language Models Deployment. Proceedings of the EuroMLSys '24. ACM Inc., New York, NY

> *“After reading, I began to wonder if LLMs are the correct way forward to achieve the tasks we are currently trying to solve using them. It seems we have to tackle large swaths of problems before we can maximize LLMs' efficiency.”*

Anonymous Reviewer