The background features a large, semi-transparent sphere on a black field. A dense, intricate network of thin, glowing lines in shades of orange, yellow, and white crisscrosses the sphere and extends into the surrounding space, resembling a complex data visualization or a neural network structure.

Navigating Challenges and Technical Debt in LLMs Deployment

Ahmed Menshawy
Vice President of AI Engineering, Mastercard



What we are going to cover today

- 1 From Excellence in Structured Data to LLMs
- 2 Intelligence Augmentation
- 3 Challenges and Technical Debt



Artificial Intelligence

The unstructured data problem

> 80%

of Enterprise data
is unstructured.

(Gartner 2017)

71%

of Enterprise are
struggling with how to
manage and protect
Unstructured data

(SailPoint 2017 Market Pulse
Survey)





Intelligence Augmentation

Stop talking about tomorrow's AI doomsday when AI poses risks today - Nature 618, 885-886 (2023)

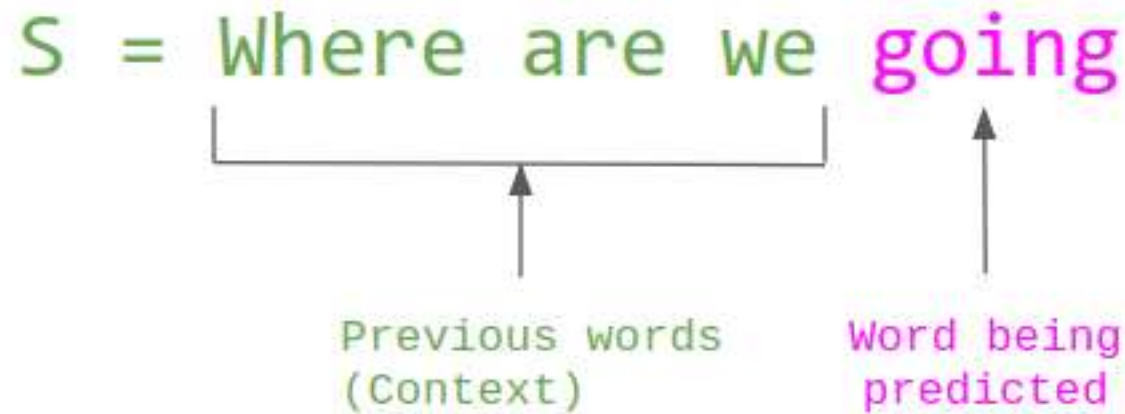
“The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis, but it has no power of anticipating any analytical revelations or truths. Its province is to assist us in making available what we are already acquainted with.”

The Analytical Engine, 1843 - Ada Lovelace: The World's First Computer Programmer



The Age of Language Models

R. Miikkulainen and M.G. Dyer. Natural language processing with modular neural networks and distributed lexicon. Cognitive Science, 15:343- 399, 1991.



$$P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})$$

User Interface

Multi-task, instruction fine-tuning

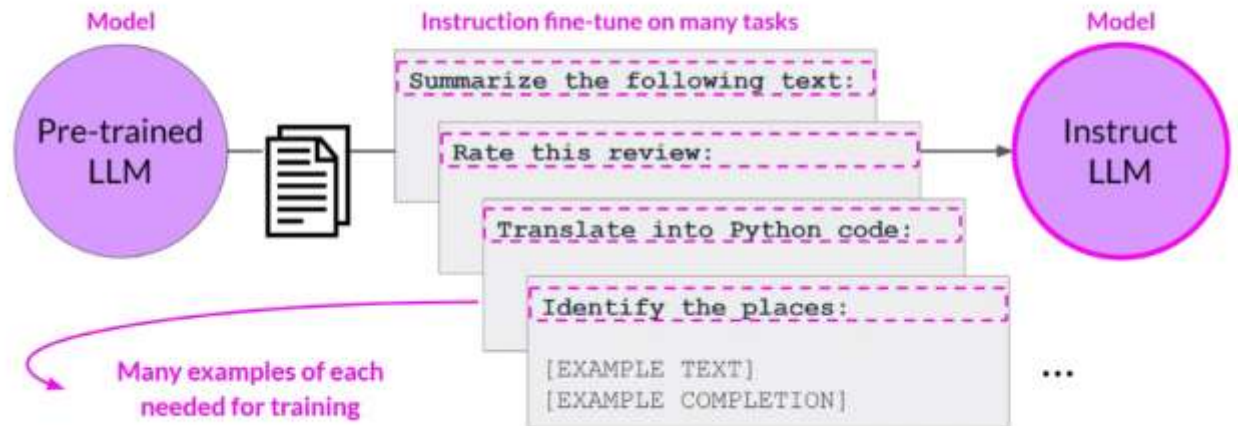
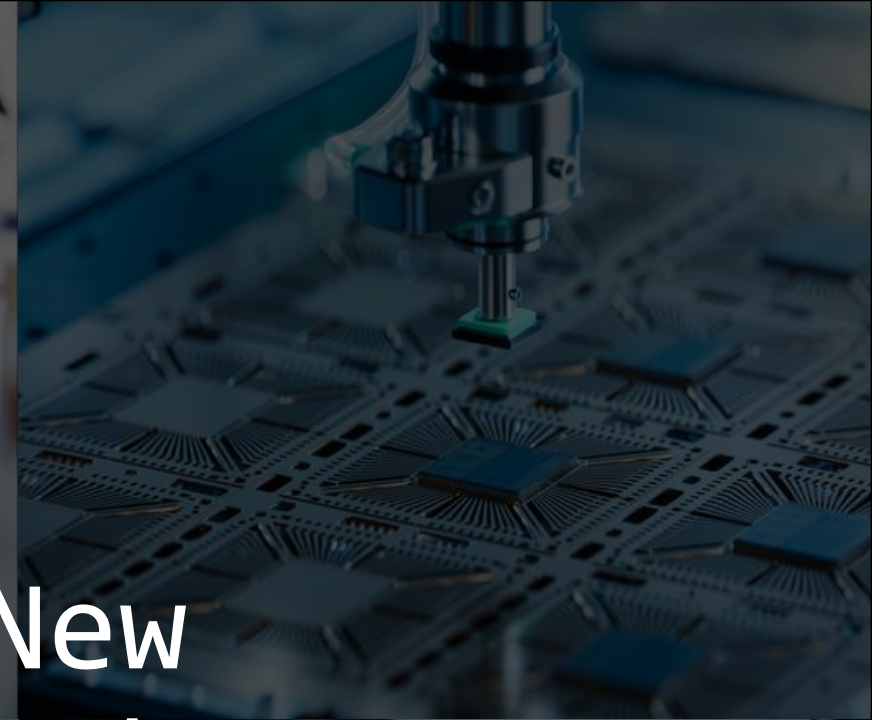



Image credit: [Medium](#)



Accelerating New Innovations Everywhere



“Mastercard jumps into generative AI race with model it says can boost fraud detection by up to 300%”

Feb 1, 2024 - CNBC Press Release

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

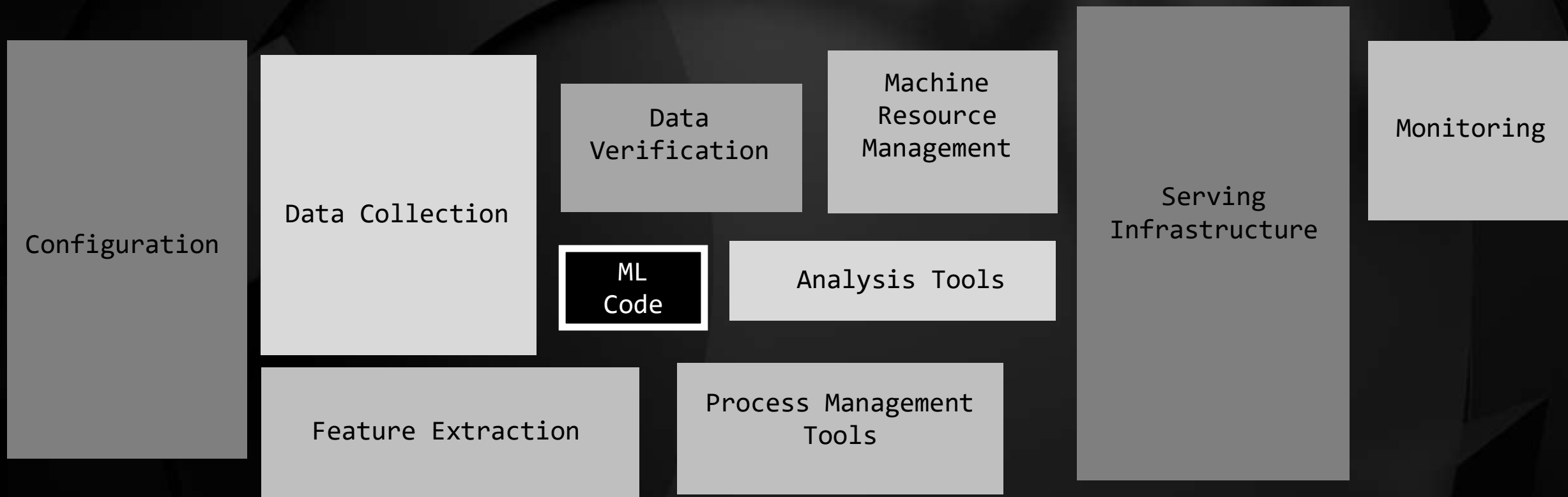
Easy-to-use tools to
build and deploy
applications

Scalable ML
infrastructure



A lot of challenges and technical debt

Not so much



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.

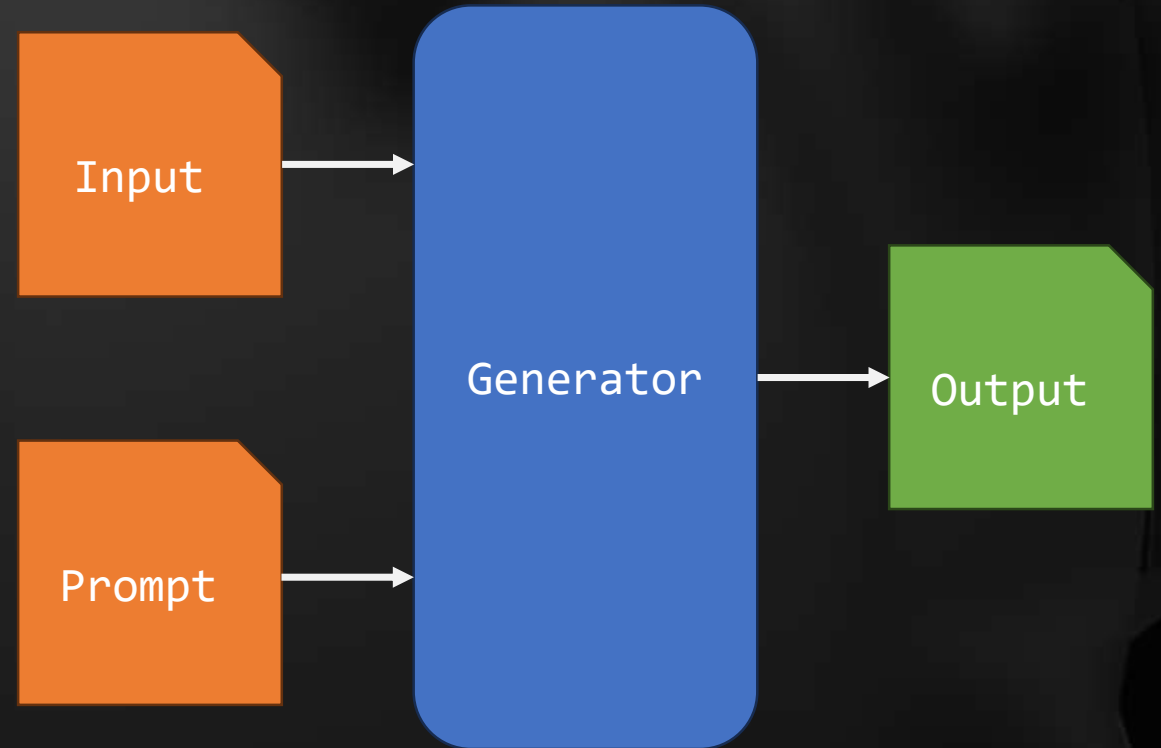
Challenges with Closed-book(Parametric) Approach

- Problems:

- Hallucination
- Attribution
- Staleness
- Revisions
- Customization

- Solution:

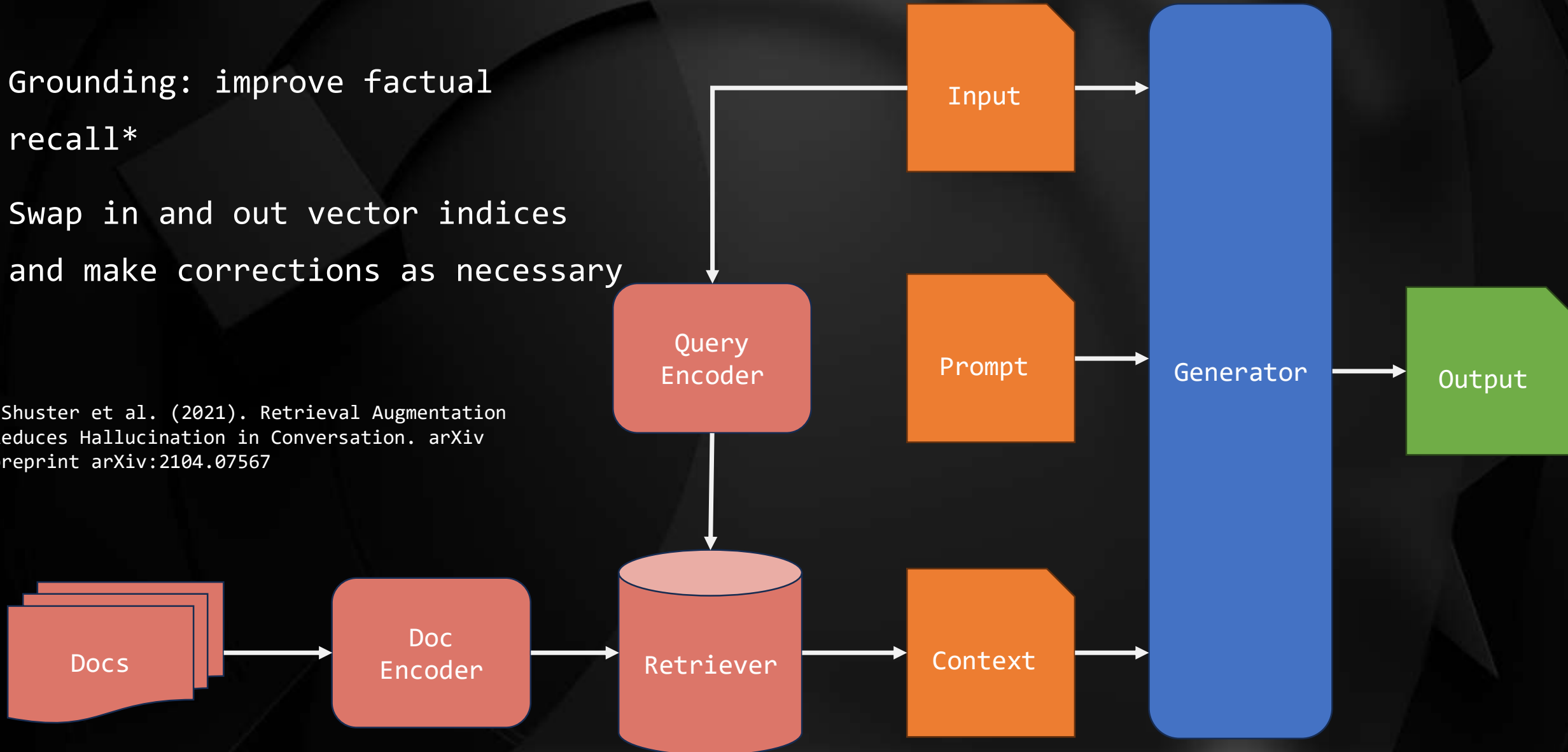
- Couple to external memory



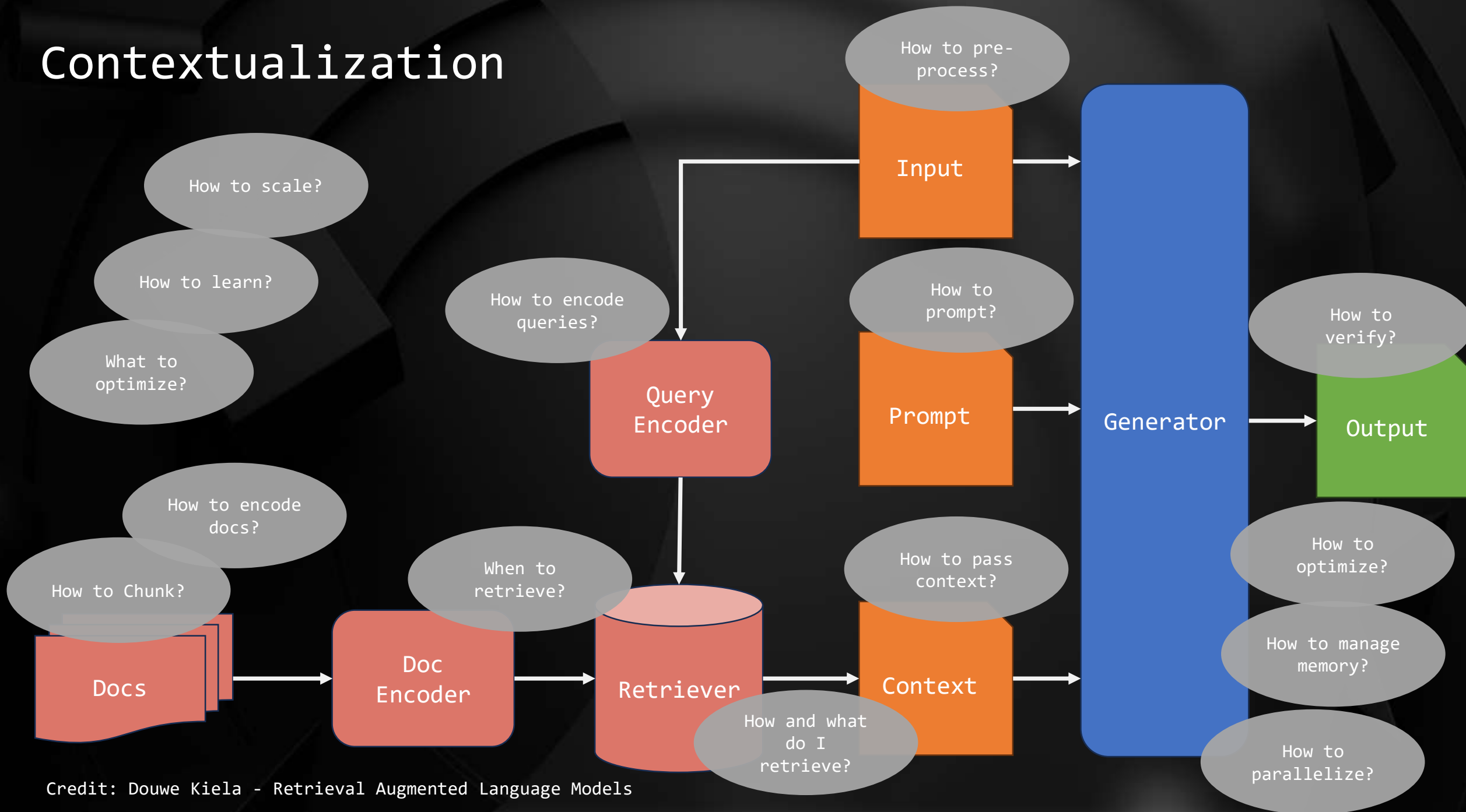
Contextualization: Open-Book (Non-parametric approach)

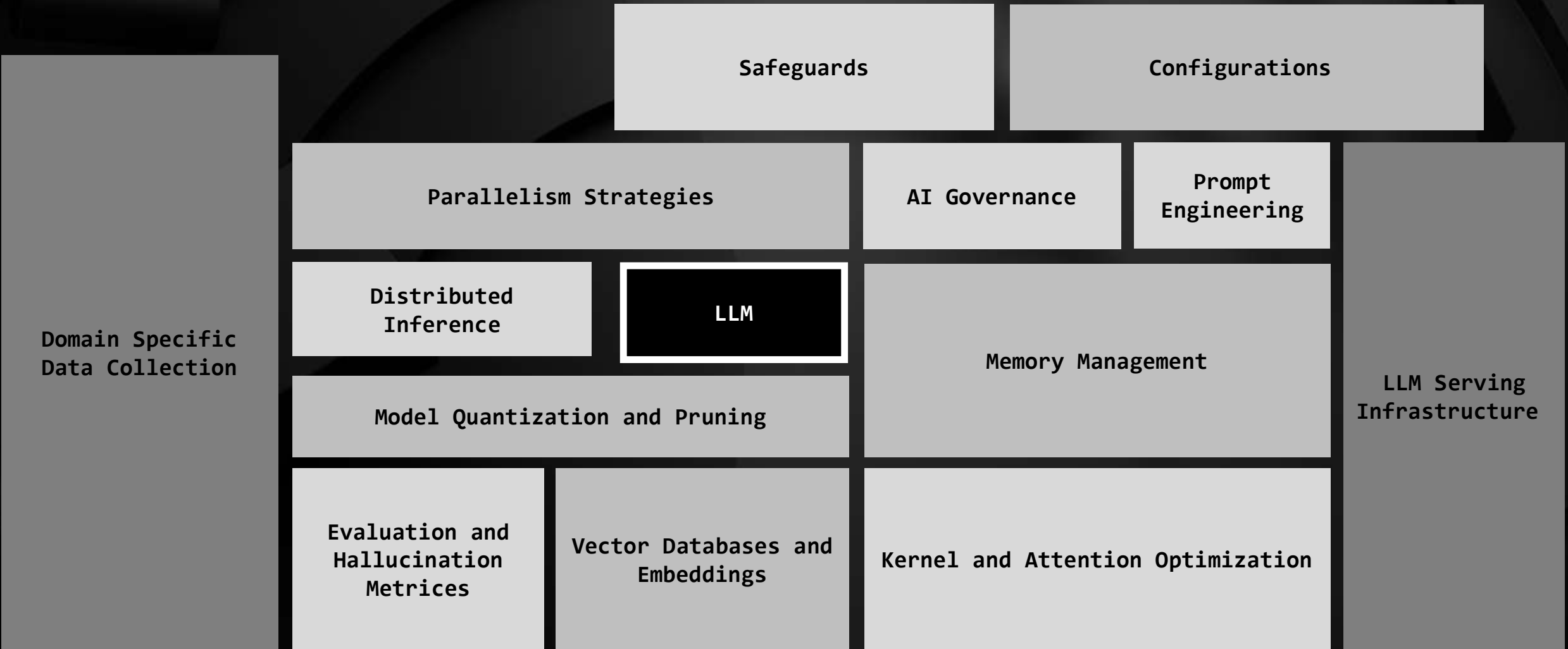
- Grounding: improve factual recall*
- Swap in and out vector indices and make corrections as necessary

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



Contextualization





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

Thank you!

