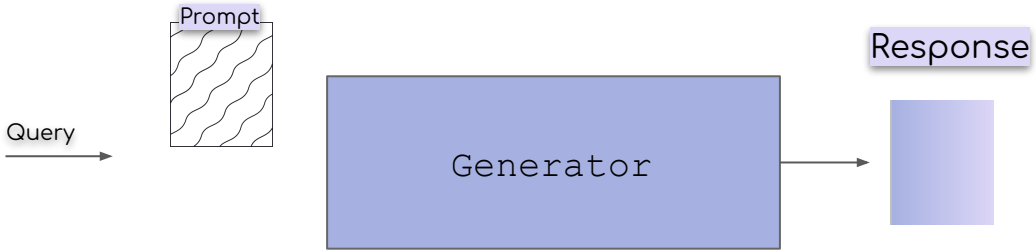
Retrieval-Augmented Generation to add knowledge

Gesuelli Pinto, Lautaro

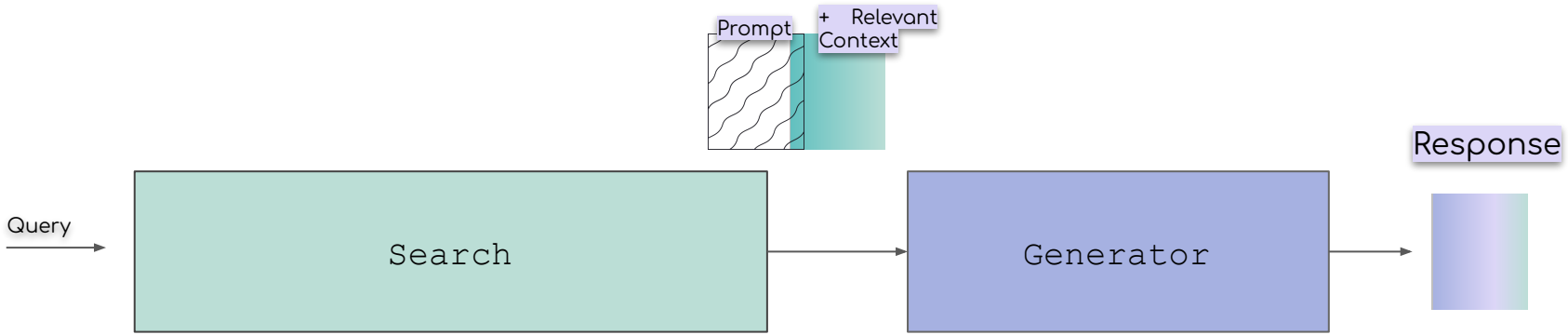
https://www.linkedin.com/in/lgesuellip/

Introduction

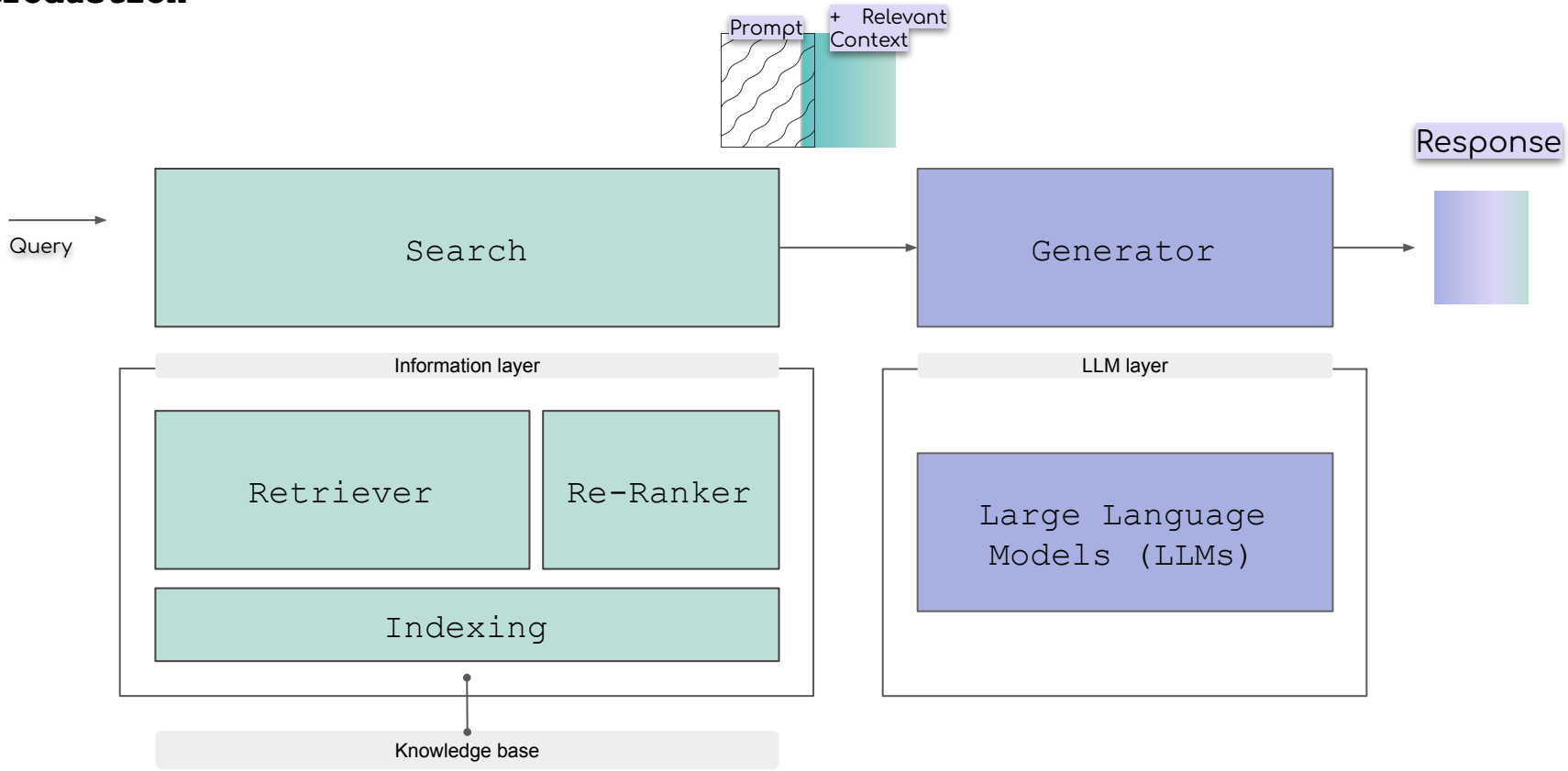
Just the Generative Model



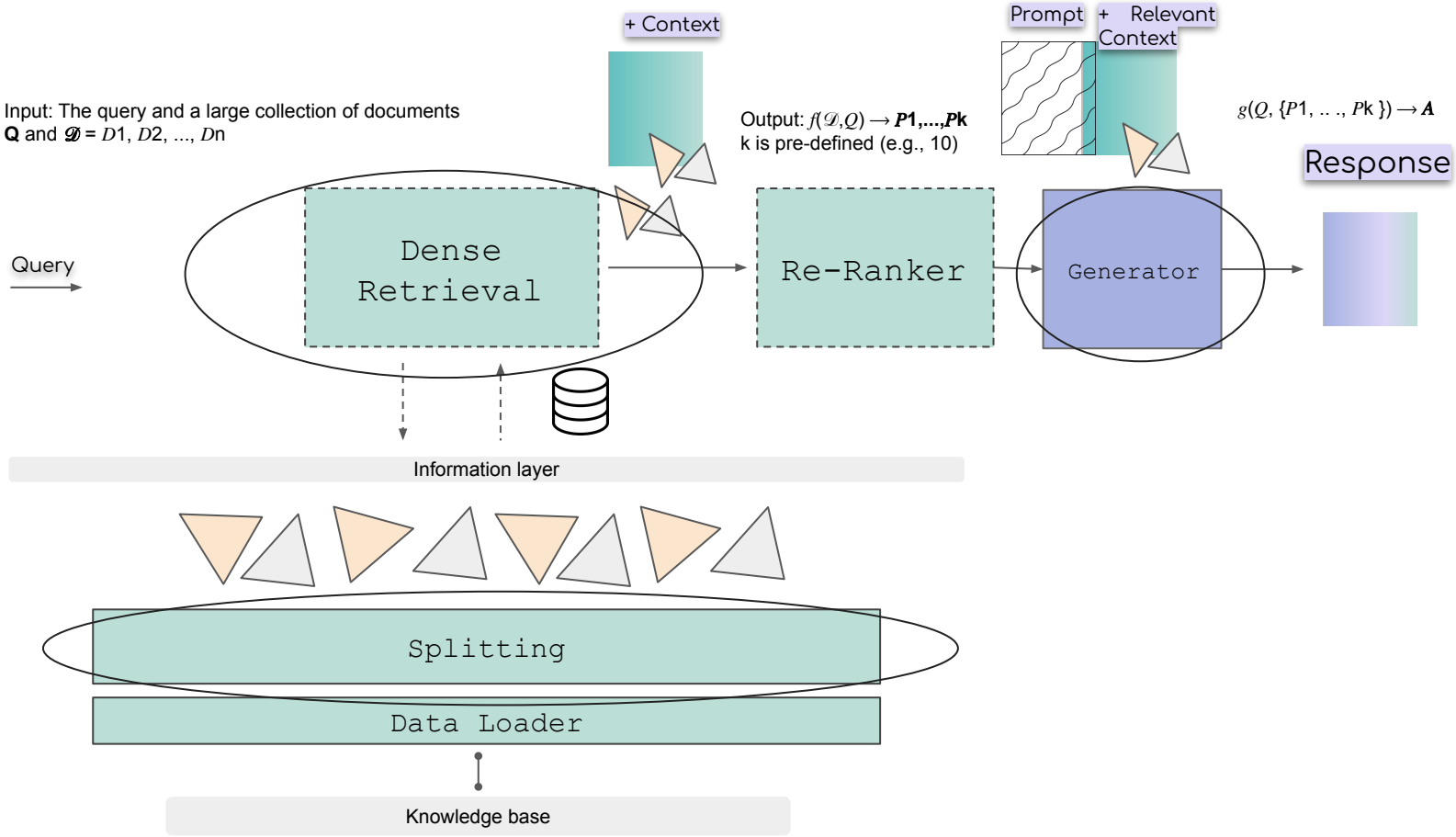
A Generative Model empowered by Search



Introduction



Introduction



References

Andrew NG and Cohere. “Large Language Models With Semantic Search.” Deeplearning.AI, 2023.

Ding, Y. Q. Y., Liu, J., Liu, K., Ren, R., Zhao, X., Dong, D., ... & Wang, H. (2020). RocketQA: an optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

Huyen, Chip. “Open Challenges in LLM Research.” Chip Huyen, 16 Aug. 2023.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Khattab, O., Potts, C., & Zaharia, M. (2021). Relevance-guided supervision for openqa with colbert. *Transactions of the association for computational linguistics*, *9*, 929-944.

Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., & Zaharia, M. (2022). Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Potts, C. (2023, April 17). *CS224u: Natural language understanding* [Lecture]. Stanford Linguistics.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Reimers, N. “Domain Adaptation for Dense Information Retrieval.” YouTube, 9 Feb. 2022.

Yan, Eugene. “Patterns for Building LLM-based Systems and Products.” eugeneyan.com, 30 July 2023.