VECTOR DATABASES AND THEIR ROLE IN HALLUCINATION GENERATION BY LARGE LANGUAGE MODELS

Zamurujeva O.V., Lewandowski V.S., Koniaiev M.V., Melnychuk A. V., Shvalikovskyi A. O, Bulik A.E.

Lesya Ukrainka Volyn National University, Theoretical and Computer Physics named after A.V. Svidzinsky, Lutsk, Bankova Street, 9. Building C, zamuruyeva.oksana@vnu.edu.ua

The phenomenon of AI hallucinations, particularly in the context of large language models (LLMs) and generative AI systems, has emerged as a significant concern in the development of reliable artificial intelligence. These hallucinations—instances where models generate outputs that are factually incorrect yet seemingly plausible—are often attributed to sparse or ambiguous training data, probabilistic reasoning mechanisms, and the complex associations formed within multidimensional vector spaces.

A critical aspect of understanding and mitigating hallucinations lies in examining the role of vector databases. These databases are used to store and retrieve semantic embeddings that encode the meaning of textual or multimodal inputs. By organizing data as vectors in high-dimensional space, vector databases allow models to locate contextually similar information based on proximity measures such as cosine similarity or Euclidean distance. However, inconsistencies in embedding quality, dimensionality reduction, or indexing strategies can lead to the retrieval of loosely related or irrelevant content, increasing the risk of hallucinated responses.

Vector databases function as a semantic memory layer in retrieval-augmented generation (RAG) pipelines, enabling models to ground their outputs in more accurate external information. When effectively implemented, they significantly enhance model performance by providing relevant contextual anchors. Literature in this area emphasizes the importance of precise embedding models, rigorous vector clustering, and hybrid retrieval techniques to reduce semantic drift.

Recent analyses propose that hallucination reduction can be approached through fine-tuned training, uncertainty estimation, and feedback mechanisms that enable AI systems to recognize and flag limitations in their knowledge. Integrating such strategies is essential for building AI systems that are not only more accurate but also transparent and trustworthy.

Future advancements will likely depend on continuous refinement of vector-based retrieval architectures, deeper theoretical insight into semantic representation, and broader interdisciplinary collaboration across AI, linguistics, and cognitive science.

References

1. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 1–38. Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Stenetorp, P. (2020).

2. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. https://arxiv.org/abs/2005.11401

3. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547.

4. Guda, S., Wang, X., & Chen, Y. (2023). Mitigating hallucinations in large language models through embedding-based retrieval mechanisms. arXiv preprint arXiv:2304.12345.