AI helped write this infographic about generative AI training and inference

The role of memory and storage in training and inference

	Training	Inference
	In training, large datasets are processed through training models to teach generative AI to recognize patterns and generate new data.	When providing a prompt or making a request to generative AI, the model uses the training patterns to generate new data that is similar to the input.
Challenge	Training requires large amounts of fast memory and storage to hold the data and feed it to the compute as it is processed.	Inference requires fast memory and storage to store the model and process the data.
Solution	Micron's AI product offerings are optimized for these complex AI training workloads.	Micron's memory and storage solutions improve inference with faster data access and processing.

Micron solutions

Micron offers a range of solutions optimized for AI workloads, including DRAM, NAND flash and SSDs.

These solutions deliver high bandwidth, low latency, reliability and durability — all critical for AI workloads.

With Micron's memory and storage solutions, AI developers improve the performance and efficiency of their systems.

Micron's AI product offerings

HBM3E

The industry's fastest, highest-capacity[1] high-bandwidth memory is Micron HBM3E. Our memory supports AI training and acceleration in the most sophisticated compute platforms designed for cognitive technology.

DDR5

Performant AI server platforms require enormous amounts of memory and DDR5 is the fastest mainstream memory solution designed specifically for the needs of data center workloads. Micron's high-density modules provide the capacity to meet the extreme data needs of AI systems.

Data center SSD portfolio

A wide range of NVMe™ SSDs support the storage needs of vast data sets required for AI, from networked data lakes supported by the Micron 6500 ION to local SSD cache built with the Micron 9400 and 7450 NVMe SSDs.

LPDDR5X

For endpoint devices like mobile phones, striking a balance of power efficiency and performance is key for AI-driven user experiences. Micron LPDDR5X offers the speed and bandwidth you need to have powerful generative AI at hand.

[1] Micron HBM3E provides higher memory bandwidth that exceeds 1.2TB/s and 50% more capacity for same stack height. Data rate testing estimates based on shmoo plot of pin speed performed in manufacturing test environment.

Learn more at **micron.com/ai**