Captured from G Craig Vachon

There is a debate raging in the nascent AI world.

Long context LLMs vs RAG for fine tuning.

One certainty, context is critical in fine tuning, and both Gemini and RAG are innovative approaches to handling it.

Gemini is a #LLM that uses a context-based architecture to generate more accurate and relevant responses. It achieves this by breaking down a conversation into a series of "moves," each of which is a unit of meaning that depends on the previous moves. By keeping track of the conversation's context in this way, #Gemini is able to generate responses that are more natural and engaging.

#RAG, on the other hand, is a technique for grounding LLMs in reality by dynamically retrieving knowledge from external sources during inference. This allows the LLM to draw on a vast body of knowledge that it hasn't been explicitly trained on, which can improve its accuracy and ability to handle complex queries.

AI Redefined (AIR) has created a best-of-both-worlds solution that can generate accurate and engaging responses while also staying grounded in real-world knowledge. We do this by creating a context-rich digital twin from which we retrieve data dynamically.