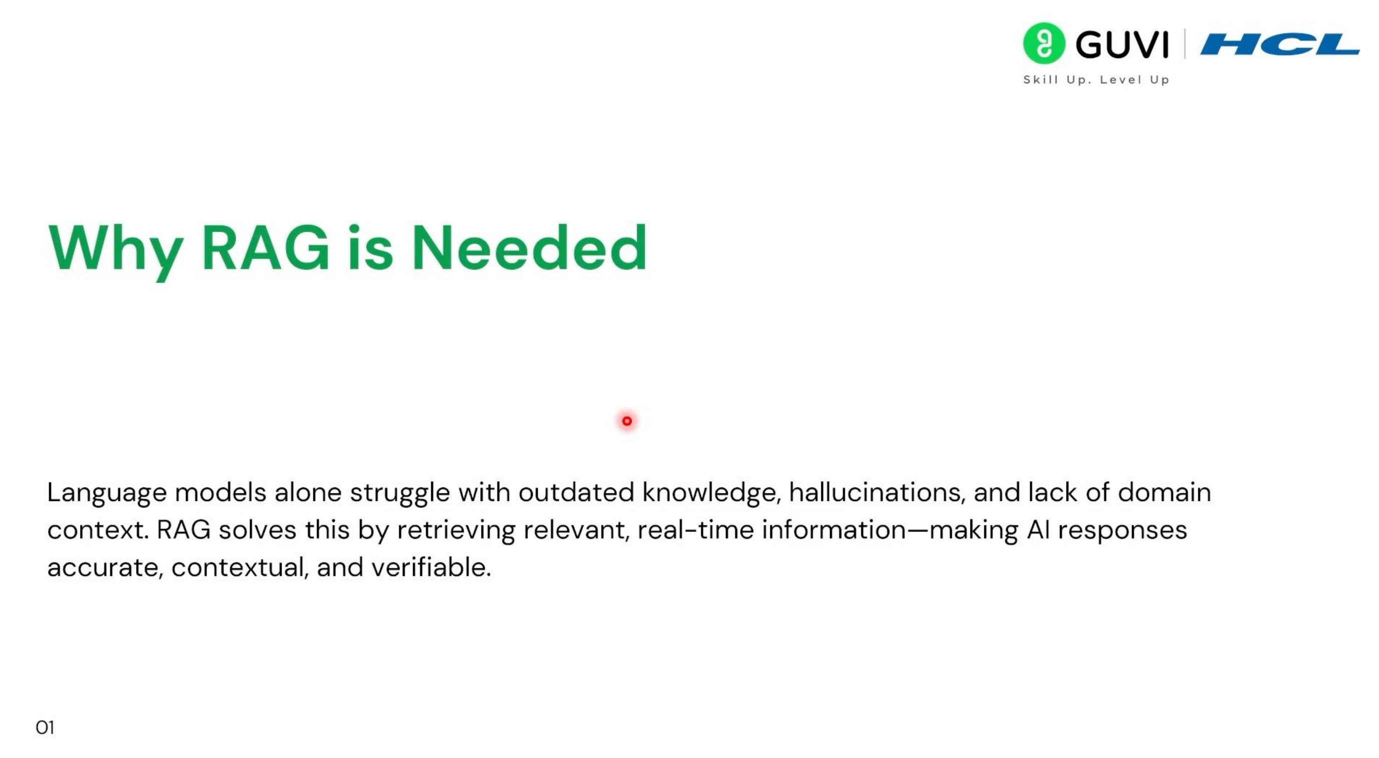
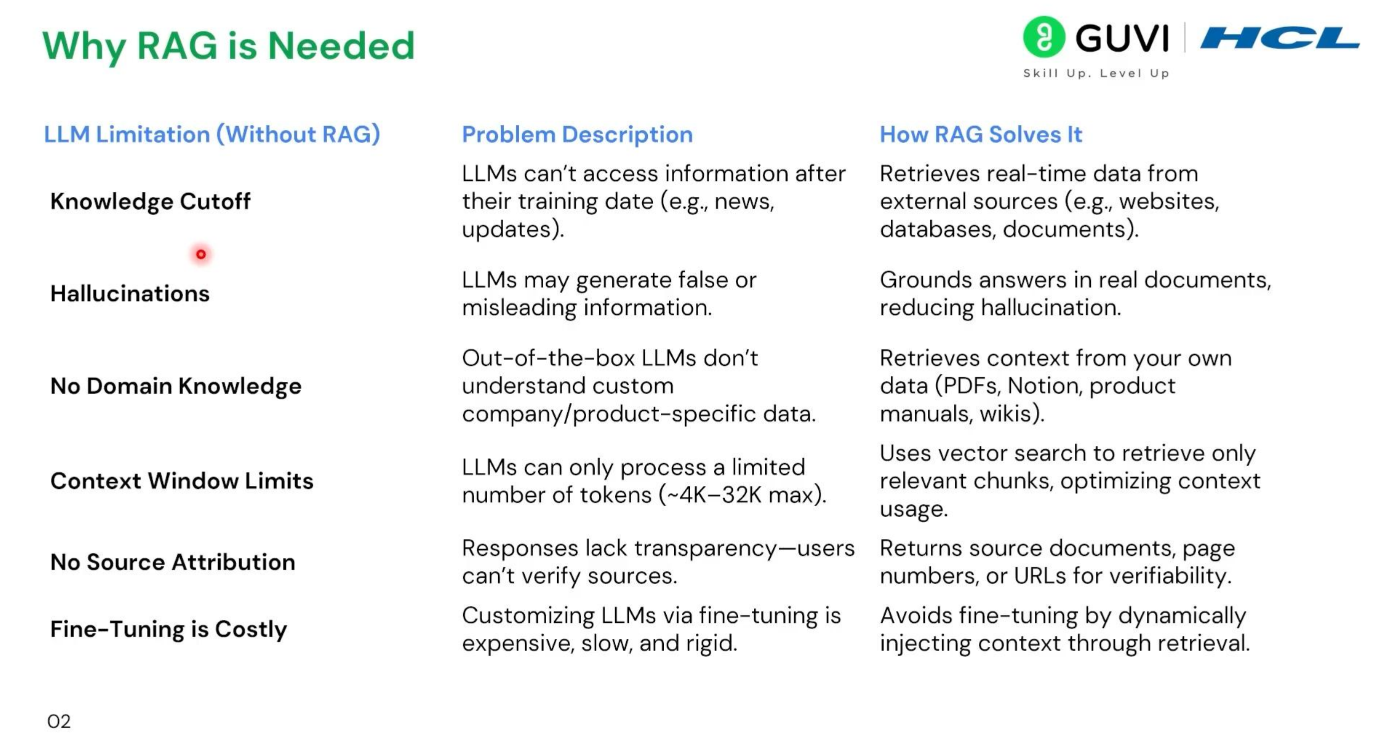
Why RAG is Needed:



Why RAG is Needed

Language models alone struggle with outdated knowledge, hallucinations, and lack of domain context. RAG solves this by retrieving relevant, real-time information—making AI responses accurate, contextual, and verifiable.



Why RAG is Needed

LLM Limitation (Without RAG)	Problem Description	How RAG Solves It
Knowledge Cutoff	LLMs can't access information after their training date (e.g., news, updates).	Retrieves real-time data from external sources (e.g., websites, databases, documents).
Hallucinations	LLMs may generate false or misleading information.	Grounds answers in real documents, reducing hallucination.
No Domain Knowledge	Out-of-the-box LLMs don't understand custom company/product-specific data.	Retrieves context from your own data (PDFs, Notion, product manuals, wikis).
Context Window Limits	LLMs can only process a limited number of tokens (~4K–32K max).	Uses vector search to retrieve only relevant chunks, optimizing context usage.
No Source Attribution	Responses lack transparency—users can't verify sources.	Returns source documents, page numbers, or URLs for verifiability.
Fine-Tuning is Costly	Customizing LLMs via fine-tuning is expensive, slow, and rigid.	Avoids fine-tuning by dynamically injecting context through retrieval.