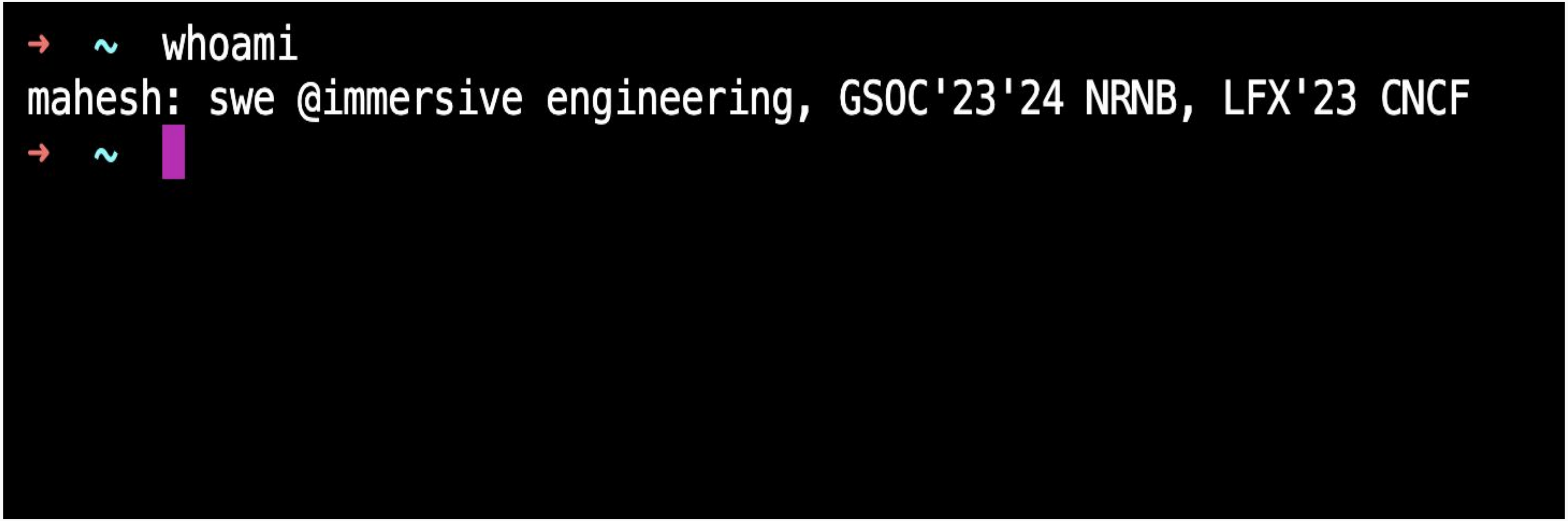
Cloud Native Meets GenAI: Unleashing the Future

Mahesh Kasbe

Whoami???



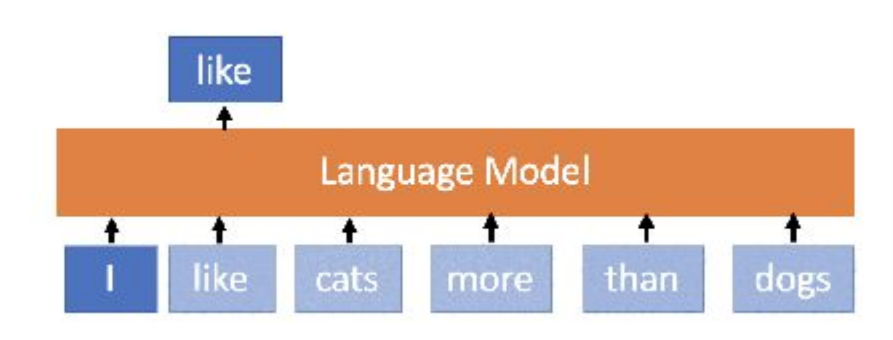
```
→  ~  whoami
mahesh: swe @immersive engineering, GSOC'23'24 NRNB, LFX'23 CNCF
→  ~  █
```

Agenda

- **Introduction to LLM’s**
- **Introduction to RAG**
- **What are agents**
  - Components of LLM Agents
- **Why GenAI in cloud Native**
- **What future holds**
- **Conclusion**

Introduction to LLMs

Large language models are **computational models that are capable of modeling and generating human language**. LLM's have the transform ability to predict the likelihood of word sequence or generate new text based on a given input.



What are LLMs good at???

- Text generation/code generation
- Chatbots and conversational ai
- Information retrieval
- Sentiment analysis

Input: 5 "tokens" -> i, like, cats, more, than

Output: 1 "token" -> dogs

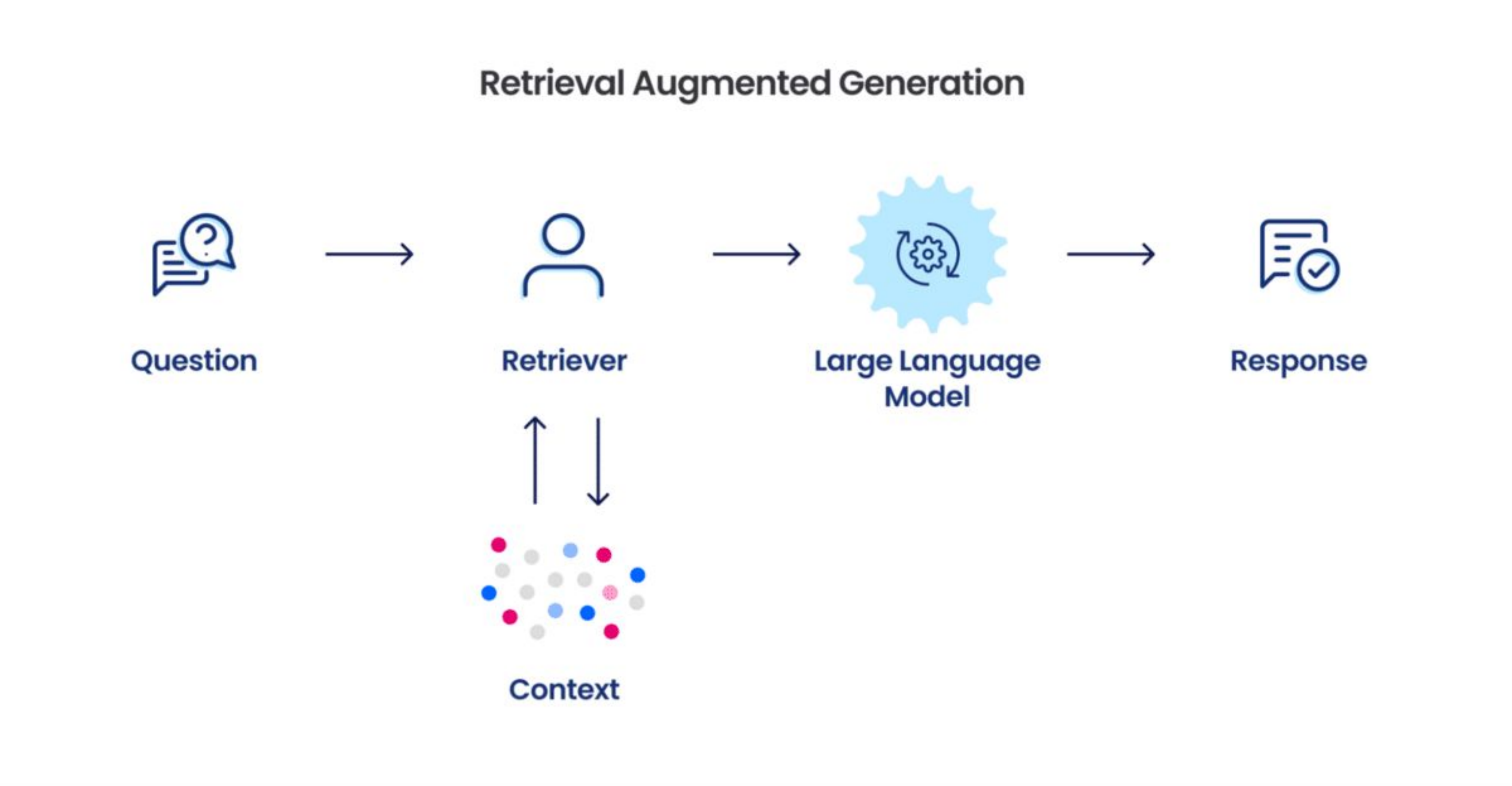
When can LLMs fail?

- **Complex reasoning tasks**: LLMs have limited reasoning capacity, LLMs are good knowledge retrievers but not good reasoners

- **No dynamicity:** LLMs are static and unable to access real time information

- **Limited knowledge(hallucination)**: While trained on vast data, LLMs lack up to date knowledge

RAG - Retrieval Augmented generation



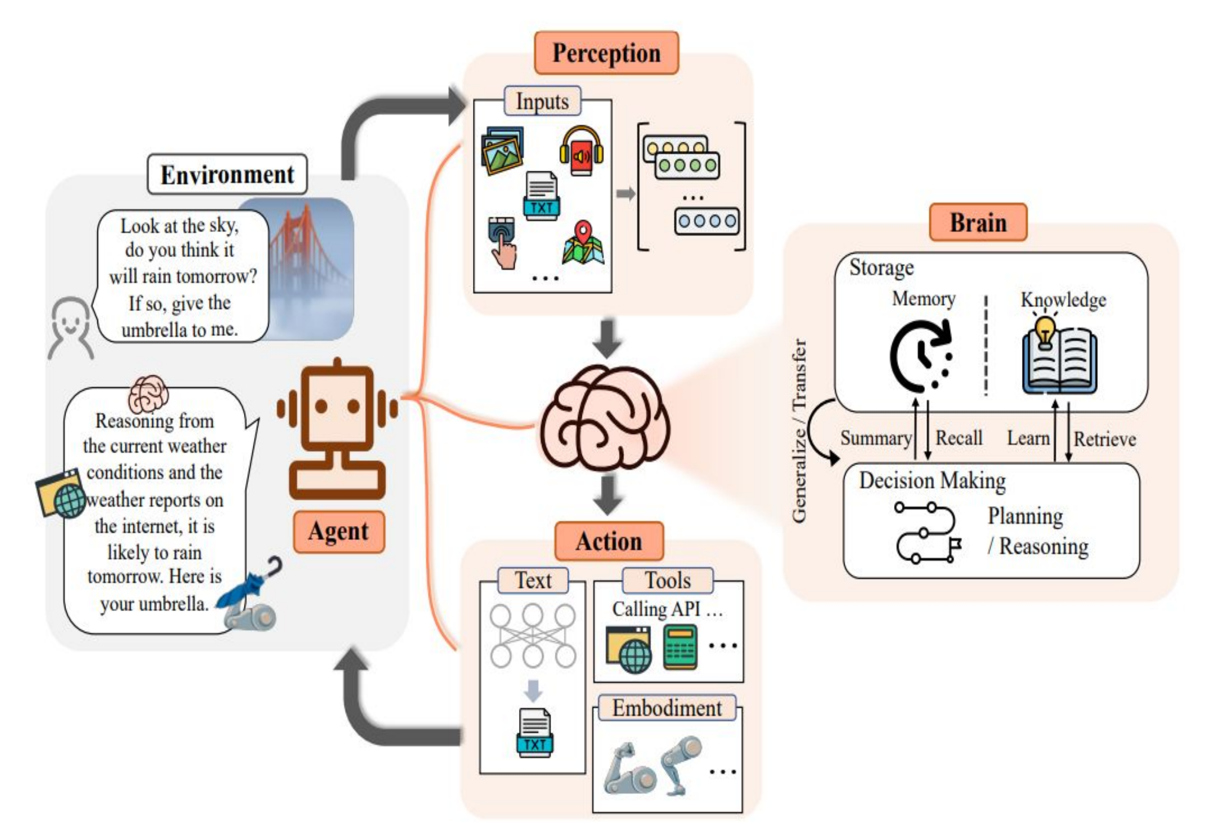
LLM Agents

LLM agents, also known as Large language model agents, **leverage LLMs to execute complex tasks by integrating them with essential components like planning and memory.**

LLM Agent = LLM + Tools + State

- LLM: Computational engine i.e “brain”
- Tools: agents ability to interact with the external world
- Memory/State: agents memory of previous message and results from used tools

Components of LLM Agents


Perception
Inputs
TXT
...
Environment
Look at the sky, do you think it will rain tomorrow? If so, give the umbrella to me.
Reasoning from the current weather conditions and the weather reports on the internet, it is likely to rain tomorrow. Here is your umbrella.
Agent
Brain
Storage
Memory
Knowledge
Summary
Recall
Learn
Retrieve
Generalize / Transfer
Decision Making
Planning / Reasoning
Action
Text
Tools
Calling API ...
Embodiment
...

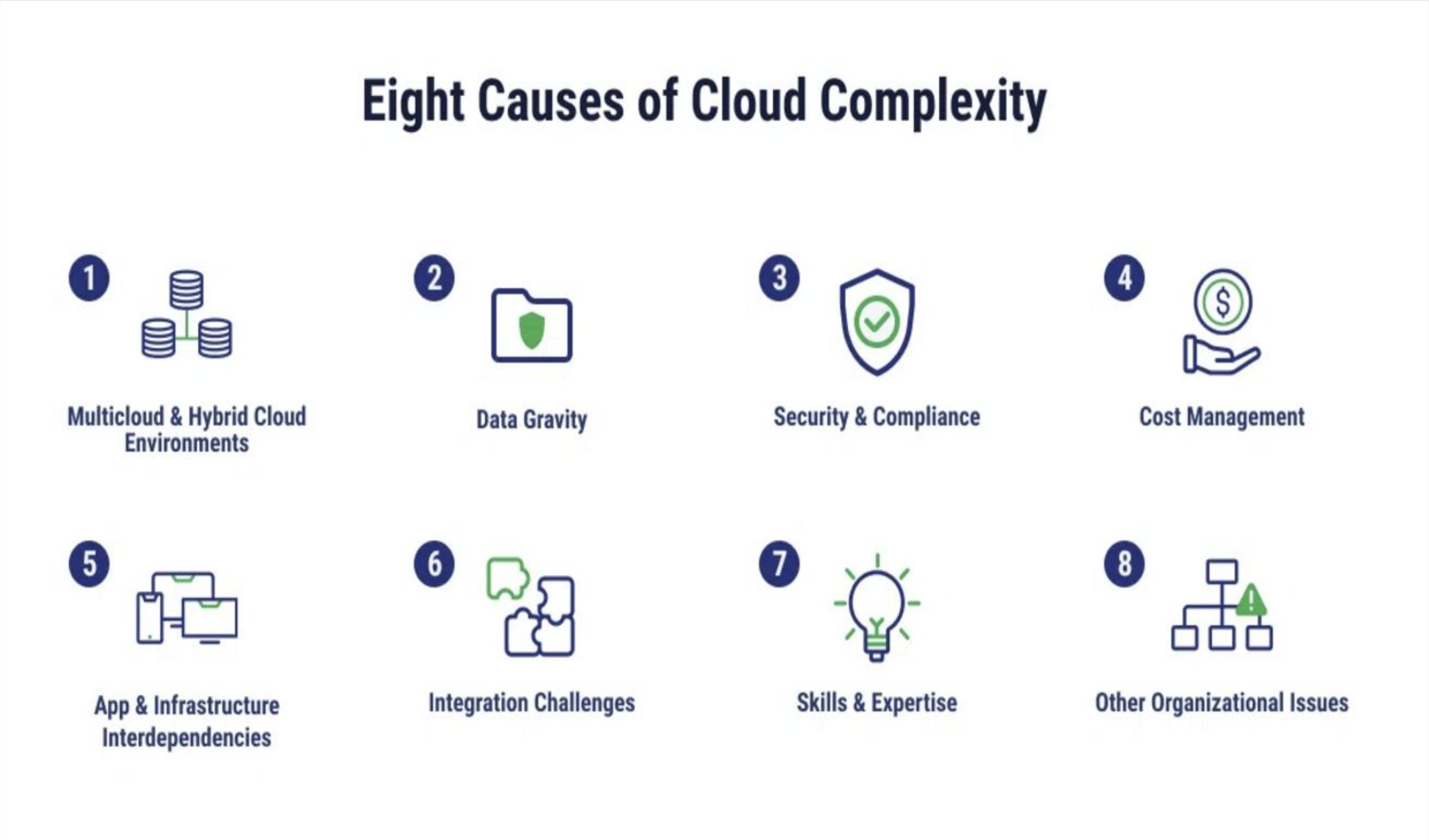
- **LLM** - Computational engine "brain"
- **Planning** - Chain of thought process to create a plan for executing tasks
- **Tools** - executable functions, APIs
- **Memory** - short term memory to retain agents thought, long term memory to retain context
- **Actions** - performs actions based on their environment and reasoning

“Imagine if your cloud infrastructure could predict and fix problems on its own using AI predictions—all while you sip your coffee. That’s not just the future; it’s happening now.”

A frustrated devops engineer like me(us)

Why GenAI in Cloud Native??

Complexity!!!

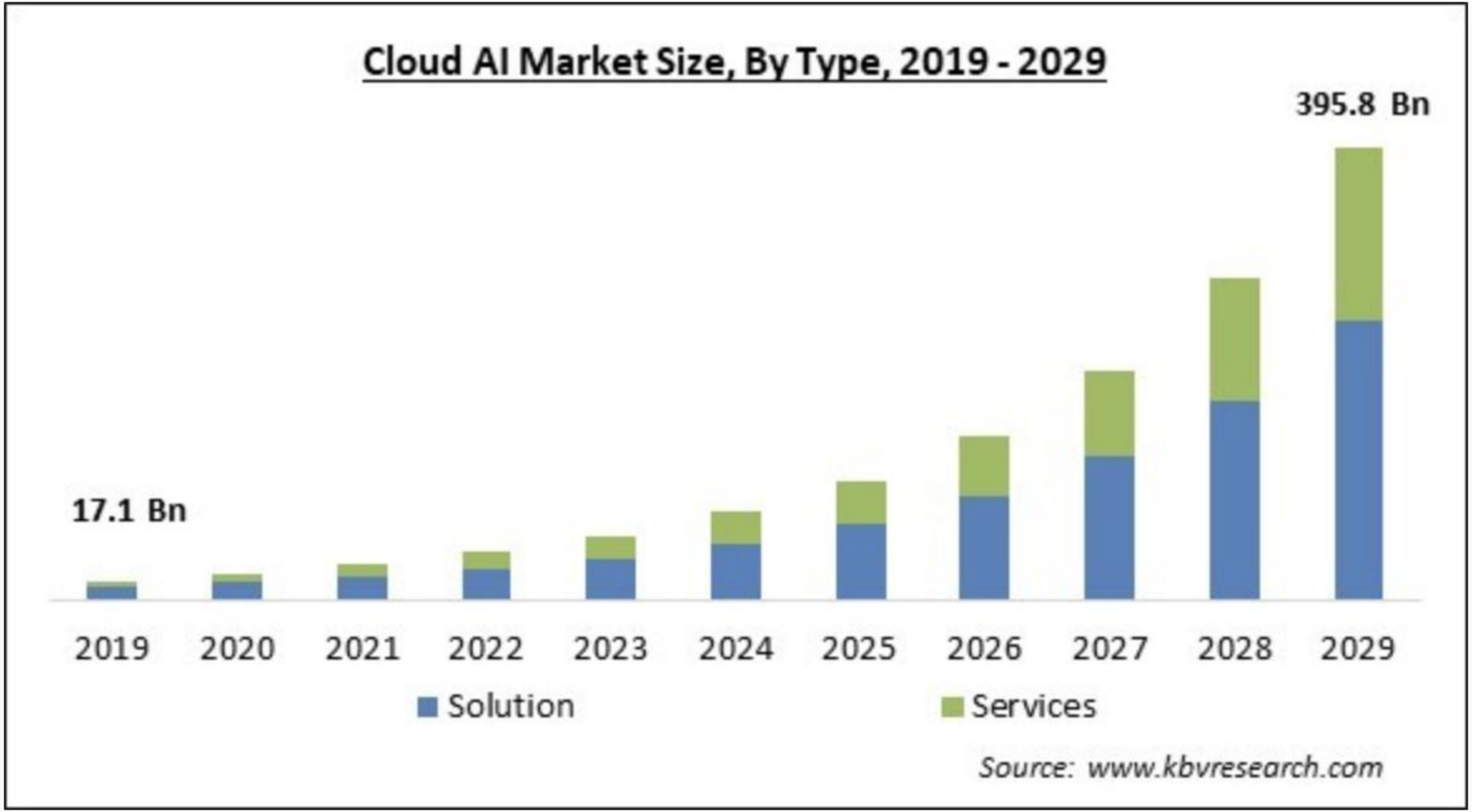


Eight Causes of Cloud Complexity

1. Multicloud & Hybrid Cloud Environments
2. Data Gravity
3. Security & Compliance
4. Cost Management
5. App & Infrastructure Interdependencies
6. Integration Challenges
7. Skills & Expertise
8. Other Organizational Issues

What does future hold??? (A prediction)

- AI-native Kubernetes clusters that "just work" without manual intervention.
- Integration of GPT-5-like models into every aspect of cloud management.
- The rise of fully autonomous DevOps workflows driven by Generative AI.
- The ability to automatically optimize cloud resources, reduce downtime, and cut costs will transform how companies manage their cloud infrastructure, making them more competitive in the market.



Cloud AI Market Size, By Type, 2019 - 2029

Source: www.kbvresearch.com

Thank You!

Say hi to me!!!

X (Twitter): @mhshx_

Linkedin: Mahesh Kasbe