Building and Deploying Intelligent AI Agents

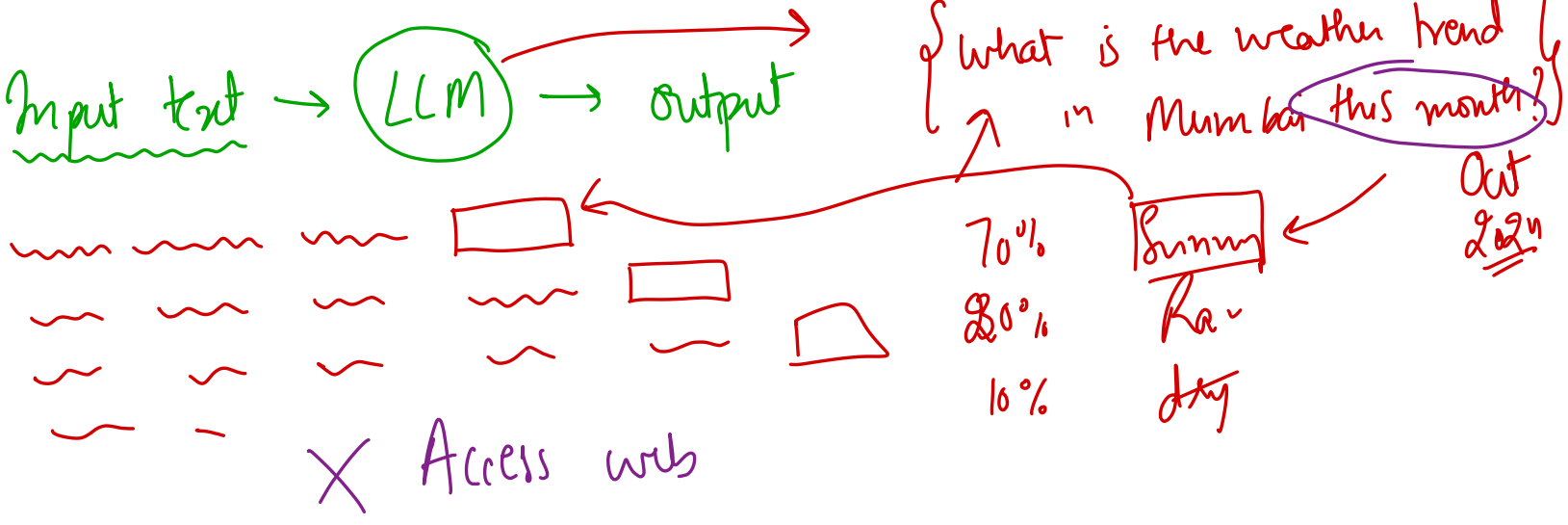
What if I told you that while we're sitting here, there's a company whose AI agents are handling 2.3 million customer conversations this month - that's the work of 700 full-time employees - and customers can't even tell the difference?

Thats Klarna’s Army of Agents.
Their AI agents are resolving customer issues in under 2 minutes, speaking 35 languages, and generating $40 million in annual profit improvement.

What is LLM (Large Language Model) ?

Brain Behind AI Agents

A Large Language Model (LLM) is a neural network with billions of parameters trained on massive text datasets to predict the probability distribution of the next token in a sequence, enabling it to generate coherent, contextually relevant text.



LLM → Talk Talk Talk.

What LLMs CAN do:

- Understand context:
- Generate creative content : Poem, stories, emails.
- Translate languages : Word by word, meaning
- Reason through problems : Step by Step Reasoning
- Adapt tone and style :

What LLMs CANNOT do (by themselves):

- Access real-time information: Weather,
- Perform actions: Accessing DB
- Remember conversations:
- Learn from you:
- Browse the internet:

Key Takeaway Message

LLM alone : Intelligent Advisor.

LLM + Capabilities : Intelligent executor.

Agent

From LLM to AI Agent: Adding Superpowers

Goal →

LLM vs AI Agents

LLM Alone
- Can Think

No awareness
No planning
No memory
No tools.

AI Agent
- Can DO

See Environment
Makes plans
Remember me
Uses tools.

The 4 Core Components

1. PERCEPTION → See & hear
2. REASONING → Think & Plan
3. MEMORY → Remember
4. ACTION → Does thing

PERCEPTION :Perception is how agents understand what's happening around them - their senses.

Types of Perception:

Text Input: Email, Chat, documents

Data Access: DB, Spreadsheet

System Status: Inventory level, Server health.

External Signals: Weather data, Stocks

Example
Without Perception (LLM):
Customer: "Check my order"
LLM: "I'd love to help but I can't see your order information"

With Perception (Agent):
Customer: "Check my order"
Agent: Sees customer email → Identifies customer → Accesses order database
"I can see your order #12345 placed yesterday for the blue laptop"

REASONING : Reasoning is the agent's ability to think through problems step-by-step and make decisions.

Reasoning Patterns:

Chain of thoughts

Example
Customer: "My package hasn't arrived and I need it for tomorrow's presentation!"

Agent's Reasoning Process:
Thought 1: Customer is stressed, time-sensitive issue
Thought 2: Need to check current package location
Thought 3: If delayed, need alternative solution
Thought 4: Can offer store pickup or overnight replacement
Decision: Check package → Offer fastest solution

MEMORY : Memory allows agents to remember past interactions, learn patterns, and maintain context.

Three Types of Memory:

1. Short-term (Working Memory):

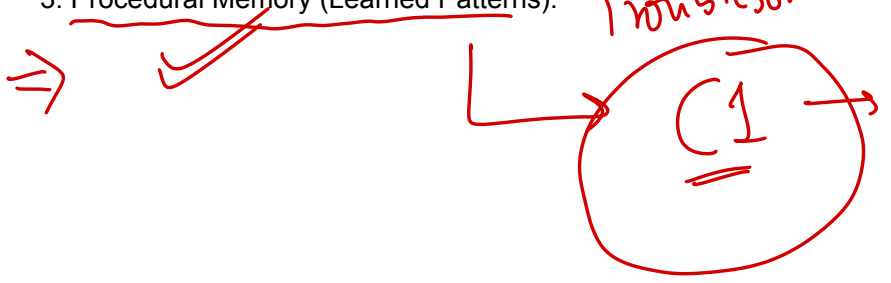
Agent Tell me about order 123

2. Long-term (Historical Memory):

Customer history → VIP Customer

3. Procedural Memory (Learned Patterns):

First Customer
Troublesome



Remote area → 2-3 days

5% - 20%

Example
Without Memory:
Monday: "I need help with order 123"
Agent: "Sure, what's your name?"
Tuesday: "Any update on my order?"
Agent: "What order? What's your name?"

With Memory:
Monday: "I need help with order 123"
Agent: "Hi John, I see order 123..."
Tuesday: "Any update?"
Agent: "Hi John! Following up on order 123 we discussed yesterday..."

ACTION : Actions are the tools and capabilities that let agents actually DO things, not just talk about them.

Integration Tools:

Call APIs
Trigger workflows
Update CRM
Sync systems.

Transaction Tools:

Processing Payment
Update Record
Modify order
Generate Report

Communication Tools:

Send Emails
Post Message
Create tickets
Schedule meeting

Information Tools:

Web search.
Analyse data
Read document
Check inventory.

Example

Customer: "Cancel my subscription and refund this month"

LLM Response: "To cancel, you need to go to settings, click subscriptions,
select cancel, then email billing for the refund..."

Agent Actions:
✓ Action 1: cancel_subscription(customer_id: 12345)
✓ Action 2: calculate_prorated_refund(days_used: 5)
✓ Action 3: process_refund(amount: $47.50)
✓ Action 4: send_confirmation_email()
✓ Action 5: update_crm_record()

"Done! Subscription cancelled, $47.50 refunded, confirmation sent."

LLM
plan
validate

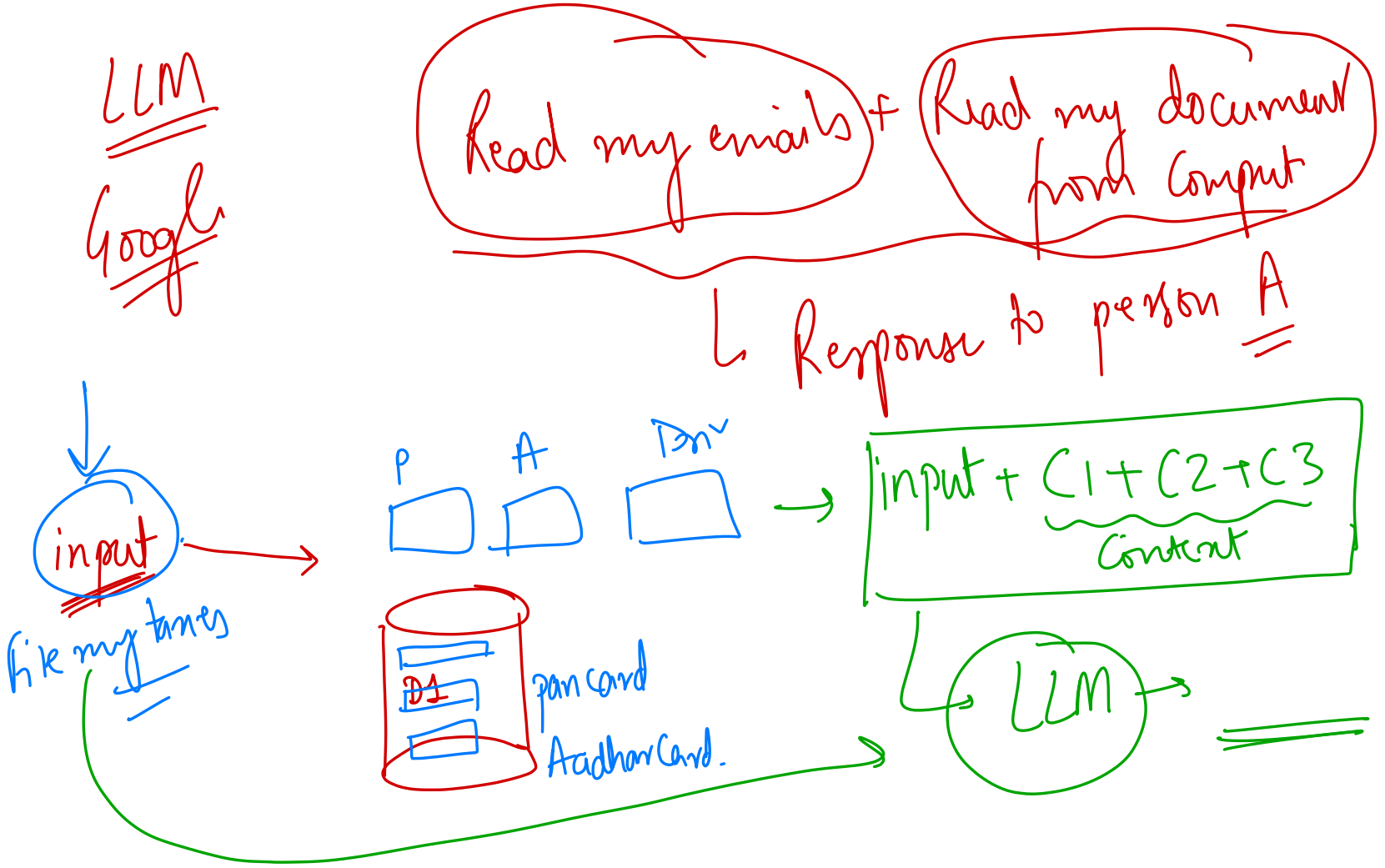
Tool Selection Logic

By Task Requirements:

Check order → database
Send email → Gmail tool
Calculate Refund - Calculator

By Confidence Level:

High → Direct execute
Med → Use with validation.
Low
Ultralow → Multiple


LLM
Google
Read my emails
+
Read my document from Comput
Response to person A
input
file my taxes
P
A
Drv
D1
pan card
Aadhar Card.
input + C1 + C2 + C3
Context
LLM

Real Life Scenario

SCENARIO: Customer emails: "I ordered 2 days ago, still nothing!"

PERCEPTION:
- Reads email
- Identifies customer from email address
- Sees frustration in tone

REASONING:
- This is a shipping delay issue
- Customer is frustrated
- Need to check order and provide solution
- Priority: High (time-sensitive)

MEMORY:
- Recalls: This customer had delays before
- Recalls: They preferred overnight shipping last time
- Notes: VIP customer, lifetime value $10,000

ACTION:
- Queries order database
- Checks tracking system
- Calculates delivery options
- Processes overnight shipping upgrade
- Sends apology email with tracking

RESULT: "I sincerely apologize for the delay. I've upgraded you to overnight shipping at no charge. Your order will arrive tomorrow by 10 AM.
Tracking: XYZ123.
As a VIP customer, I've also added a 20% discount to your next order."

Time taken: 4 minute.
Human time: 4 hour

How Agents Plan?

How AI Agents Plans?

→ Breakdown problem into smaller task
→ Determine sequence + dependencies
→ Allocate Resources + tools.
→ Checkpoint + validation.

Example

Task: Process refund request

1. Validate customer identity ← Tool: DB
2. Check return eligibility ← Tool: Policies
3. Calculate refund amount ← Tool: Calculator
4. Process payment ← Tool: Payment API
5. Send confirmation ← Tool: Gmail

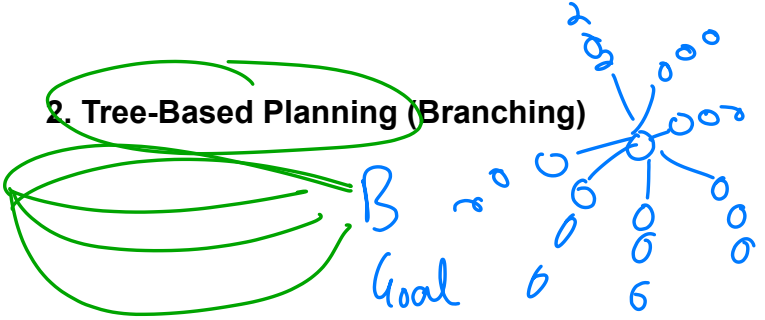
Fallback → Escalate to Human

Planning Approaches

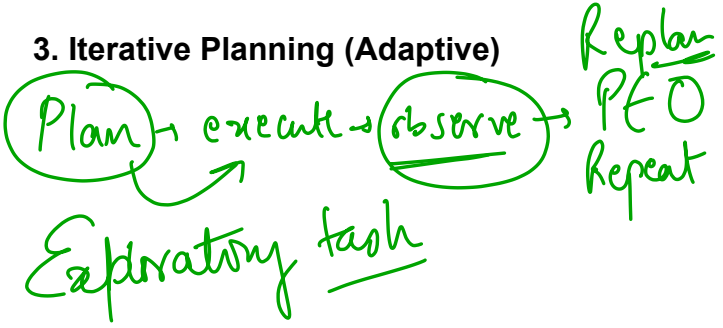
1. Linear Planning (Simple)

S1 - S2 → S3

2. Tree-Based Planning (Branching)



3. Iterative Planning (Adaptive)



The Evolution Journey (2022-2025)

2022: The ChatGPT Era

- Simple Q&A interactions
- No memory between conversations
- No ability to take actions

2023: The Plugin Phase

- Basic tool usage introduced
- Limited chaining of actions
- Memory still session-based

2024: The Agent Revolution

- Autonomous decision-making
- Complex multi-step workflows
- Persistent memory systems
- Production deployments begin

2025: The Production Era

- 51% of enterprises have agents in production
- Multi-agent systems are becoming standard
- Industry-specific agents dominate

Why NOW is the Moment

Four forces

1. LLM Capabilities : GPT-5 and Claude can now reliably understand complex instructions

2. Tool Integration: APIs everywhere - your CRM, payment systems, databases all have APIs We've built the 'hands' for AI to use

3. Framework Maturity: LangChain: 112,000 GitHub stars Production-ready, not experimental. Microsoft Autogen and others.

4. Economic Pressure : Labor costs rising, customer expectations increasing Companies NEED 10x productivity gains