Are we there yet?

Journey from automation to autonomous systems

Mangesh Pise | August 2024

Large Language Model (LLM) - Plethora of options

GPT
OpenAI

ChatGPT

- GPT 3
- GPT 3.5
- GPT 4
- GPT 4o
- GPT 4o-mini

Claude
Anthropic

- 3 / 3.5 Haiku
- 3 / 3.5 Sonnet
- 3 / 3.5 Opus

Llama
Meta/Facebook

- Llama 3
- Llama 3.1
- Llama Guard

Gemini
Google

- Gemini 1 / 1.5 Ultra
- Gemini 1 / 1.5 Pro
- Gemini 1 / 1.5 Flash
- Gemini 1 / 1.5 Nano

Command
Cohere

Bloom
BigScience

... 1000's more OpenSource

Key Factors driving new LLM Models

Token Window Size

GPT 3.5 (4,096 tokens) —> GPT 40 (128,000 tokens)

Why is this important?

- Better grounding
- Less hallucinations

Training Parameters

8 billion parameters to 405 billion (Llama)

Why is this important?

- Better instruction following
- Wider vocabulary, hence quality

Cost

50% to 78% reduction (consumption)

Why is this important?

- Adoption & Trust
- Competition

Model Size

GPT 4o vs 4o-mini | Gemini Ultra vs Nano

Why is this important?

- On-device embedding
- Better inference Speed

Multi-Modality

Chat completions to VQA

Why is this important?

- Towards Artificial General Intelligence
- Multimedia based interaction

GenAI Consumption
(general use)

→

Prompt Engineering
... still plays a BIG role

One-shot **Few-shot** **Chain-of-Thought**

evolving ...

ReAct
Reasoning & Acting

ART
Automatic Reasoning
And Tool Use

Reflexion
Self-Reflecting

BraggingRights!

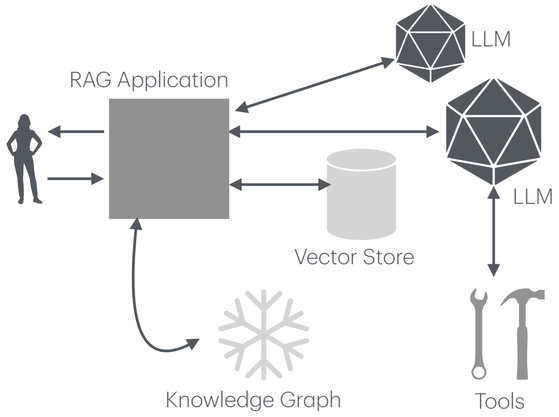
Deep-Dive
Simulate follow-ups

GenAI Consumption
(Application development use)

→

Dynamic Grounding Techniques
Trust Risk & Security (TRiSM)

Dynamic Grounding Techniques



RAG
With Vector Store

RAG
With Knowledge Graph

RAG
Hybrid

BraggingRights!

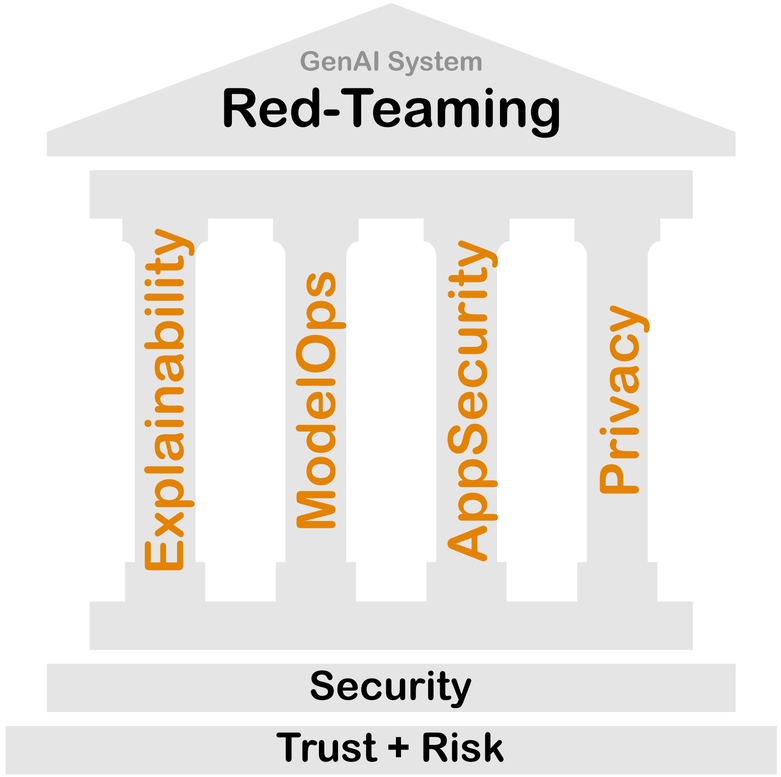
BroadAI
Multi-Agent System

Prompt-Chains
LLM orchestration

Agentic
With Tool Calling

Fine-Tuning
Improve tone/format/quality

Trust Risk & Security (TRiSM)



Current affairs & Path forward

2023 — 2024 — 2025+

Analyst's Perspective

- Organizations are not ready for risk management & governance with GenAI
- GenAI has entered trough of disillusionment, based on Gartner's hype-cycle for AI (2024)
- Excitement is being tempered by reality – only value-led use cases will lead to successful scaling
- "Gen AI: too much spend, too little benefit?" Goldman Sachs Global Macro Research

Business Executive's Perspective

- Top 3 goals to invest in GenAI for:
  - operational efficiency,
  - employee productivity,
  - Revenue growth
- 83% executives plan to continue increasing investment in AI over next 3 years.
- Concerned about data, ethics, privacy & regulatory uncertainty.
- Only 16% organizations have workforce equipped to use GenAI.

Developer's Perspective

- Developers are excited about using GenAI for:
  - Coding (64%)
  - Documentation (29%)
  - Writing test scripts (23%)
- GenAI Software Development Lifecycle missing guidance on:
  - Model selection
  - Model deployment / use
  - Managing LLM lifecycle
  - Testing
  - Security
  - Etc.

Take aways!

Models evolution is constant

AI strategy must include aspects to manage this constant change

Executives believe in outcome, but realize there is long runway

Risk Management & Governance around data practices, regulatory changes, and organizational up-skilling

Scope for driving architectural maturity

Standard reference architecture is still a moving target due to evolution in GenAI & related technology components

Cost-Value is at front and center

Ungoverned costs* on GenAI without guaranteed value-add is a big hurdle for continued investment

* **Cost components:** LLM (Tokens), specialized infrastructure (GPU), supporting technologies based on use-cases (embedding models, vector & graph stores, caches, guardrails, etc.), training, personnel, etc.

Thank You!

Mangesh Pise
https://www.linkedin.com/in/mangeshpise

References

- https://kpmg.com/kpmg-us/content/dam/kpmg/corporate-communications/pdf/2024/kpmg-genai-survey-august-2024.pdf
- https://www.pwc.com/us/en/tech-effect/ai-analytics/responsible-ai-survey.html
- https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-state-of-gen-ai-q3.pdf
- https://www.goldmansachs.com/images/migrated/insights/pages/gs-research/gen-ai--too-much-spend,-too-little-benefit-/TOM_AI%202.0_ForRedaction.pdf
- https://devops.com/survey-surfaces-widespread-reliance-on-generative-ai-among-developers/