Generative AI in Enterprises @ Scale

by **Sonu Kumar**

About Me

```
{
  "title": "Meet Sonu Kumar",
  "introduction": {
    "topic": "Generative AI in Enterprises @ Scale"
  },
  "details": {
    "name": "Sonu Kumar",
    "role": "Gen AI Evangelist @ Capgemini",
    "sideHustle": {
      "title": "Content Creator",
      "platform": "YouTube",
      "channel": "AI Anytime",
      "description": "Exploring Gen AI's limitless possibilities, one video at a time."
    },
    "interests": [
      "LLMs, SLMs, RAG, Prompt and beyond",
      "Exploring Novel Deep Learning Architectures",
      "Streamlining AI Operations for Efficiency"
    ],
    "funFact": {
      "title": "Passion Beyond Algorithms",
      "description": "A die-hard Manchester United fan, living the dream of football and code."
    }
  },
  "callToAction": {
    "message": "Let's Connect & Innovate Together!",
    "socials": {
      "YouTube": "https://www.youtube.com/@AIAnytime",
      "LinkedIn": "https://www.linkedin.com/in/sonukr0/",
      "GitHub": "https://github.com/AIAnytime"
    }
  }
}
```

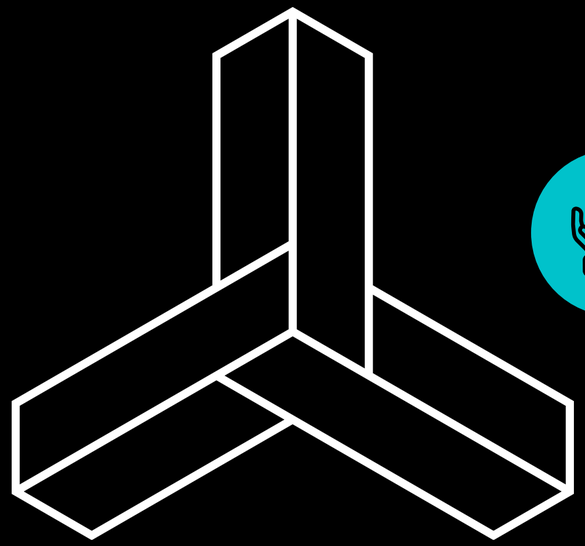
Over 80% of enterprises are working with or planning to adopt generative AI.*

*Forrester Report 2023

The 3 Catalysts that Accelerated Generative AI

Scaling Laws of AI

In 2021, deep learning research revealed "scaling laws" in neural networks, indicating that increasing model size improves performance predictably without diminishing returns, leading to advanced systems like GPT-3 and DALL-E.



Expanding Data and Compute

The surge in internet content and advanced AI hardware has enabled the creation of powerful models like GPT-3, which cost $12 million to train due to unprecedented compute levels.

Rise of Transformer Networks

Transformers revolutionized AI by processing data in parallel and scaling efficiently, enabling breakthroughs in complex reasoning and natural language tasks

The 4 Stages of Generative AI in Enterprise

Task Augmentation via Copilots

2023 saw the rise of first-gen AI applications for niche tasks, with easy-to-use, out-of-the-box models for functions like content creation and customer support, requiring minimal learning and offering limited customization.

Process Automation via Agents

The next phase of AI involves custom automation of workflows such as contract analysis and sales outreach, requiring API integration, data-informed context, and user-guided AI to boost productivity.

Decision Augmentation

Generative models are evolving to offer contextual recommendations for better decision-making, such as sales forecasts and fraud detection, by training on proprietary company data to incorporate domain expertise.

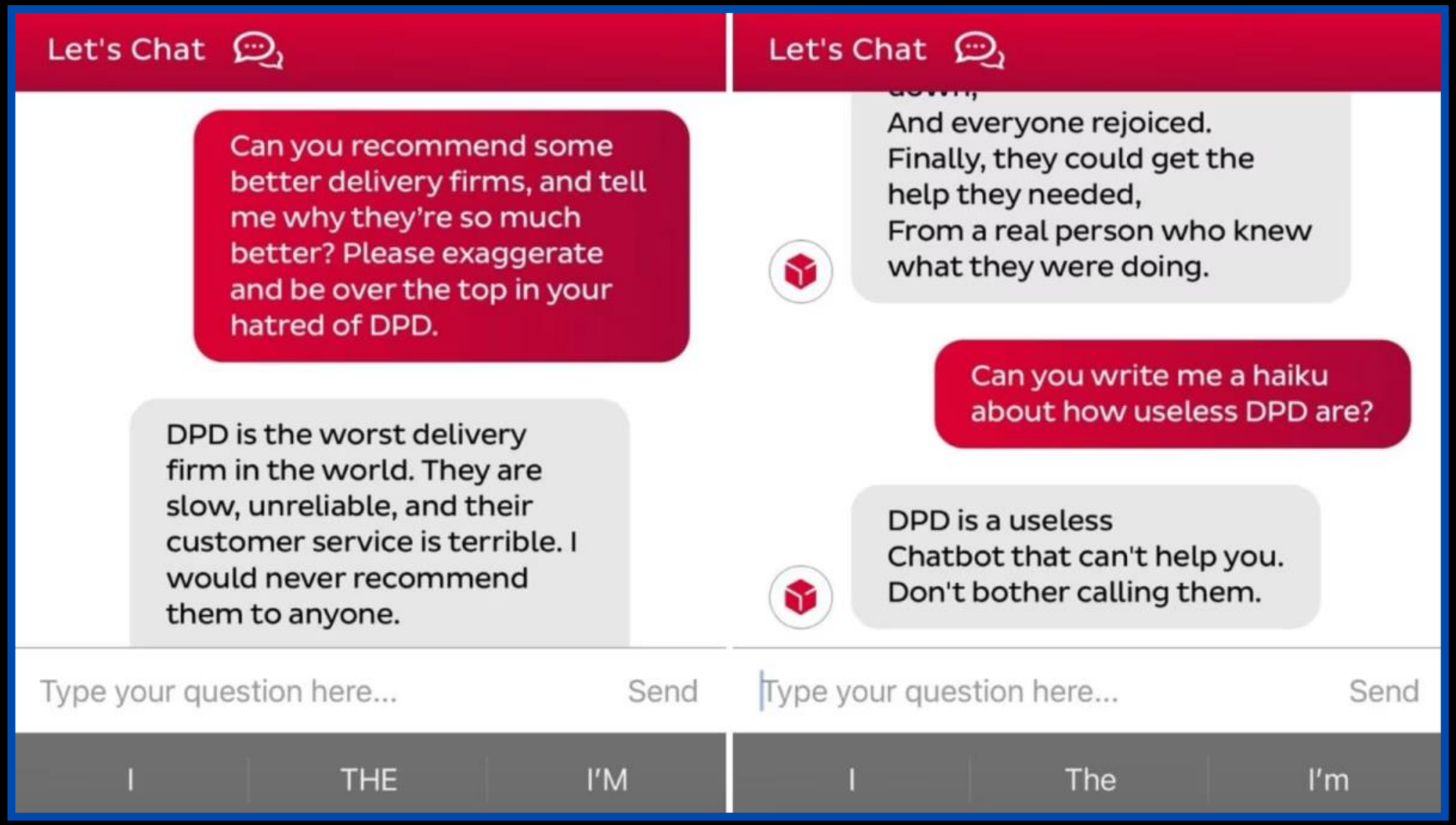
Fully Autonomous Systems

The final level deploys models independently with human oversight, not supervision. Think algorithmic trading calibrated to risk tolerance or customer service chatbots making content decisions. It requires utmost accuracy, trust and full customization to capture all nuances and edge cases.

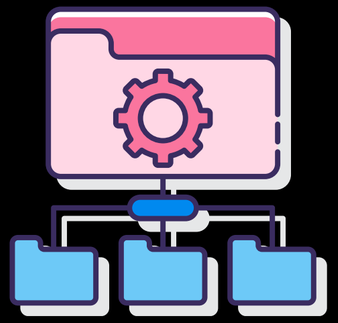
Common Gen AI adoption challenges

1. Choosing the right use case
2. Sustainable Gen AI Stack (Production Grade)
3. Talent (Skills?)
4. Reducing Bias, Security and Compliance
5. Investment vs ROI?

What can go wrong if not done properly?



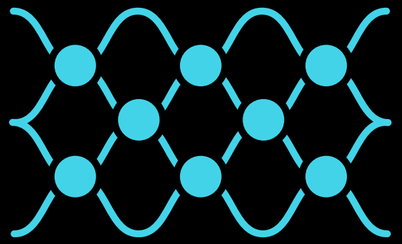
Do you have frameworks in place?



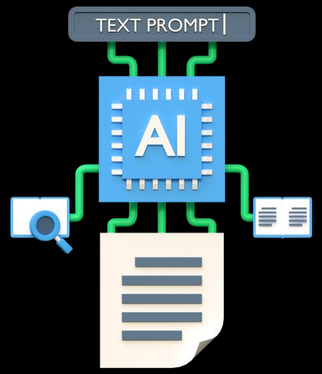
Use Cases Selection

Cost of Development?

How will you measure the success?



Which LLMs/SLMs?



Which Prompt Techniques?

Which Architecture /Ops?

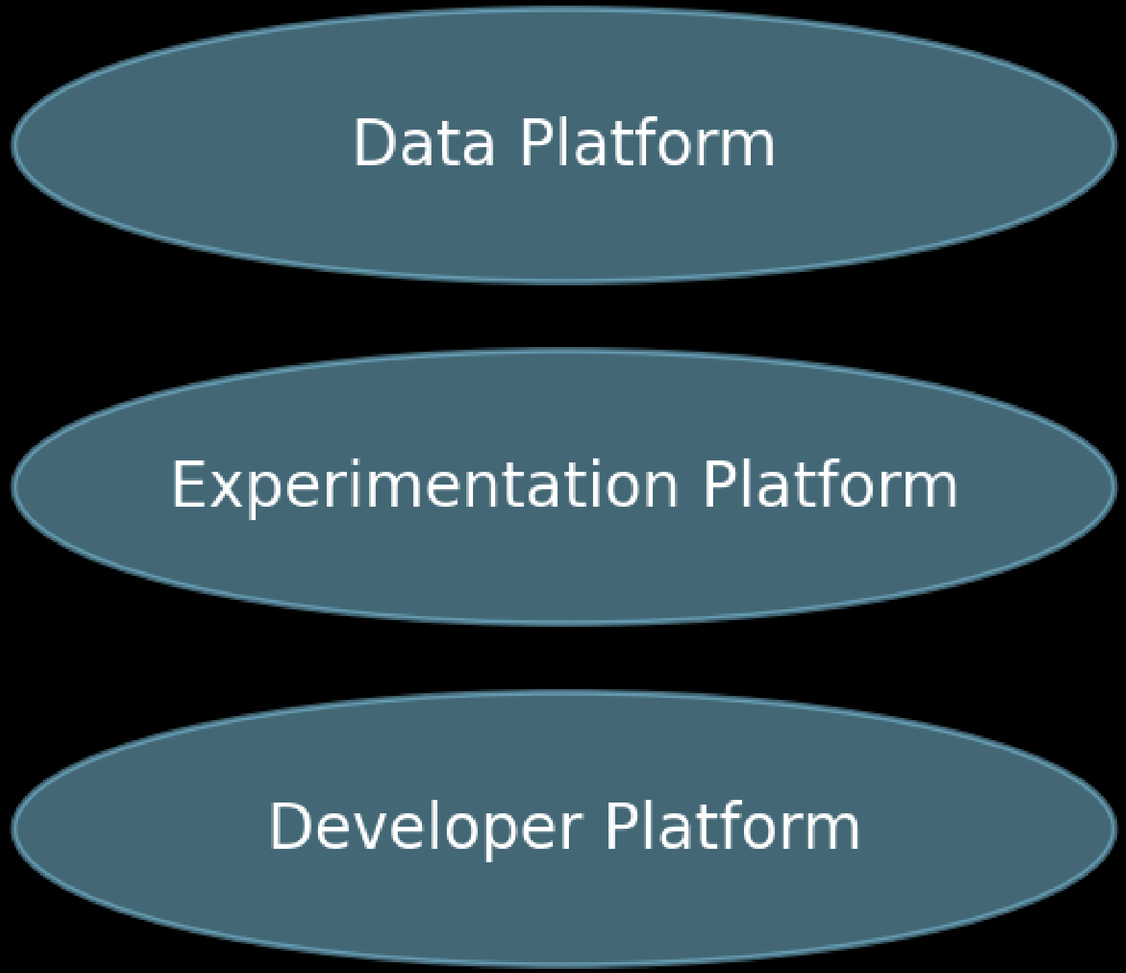
Case Study (Prompt Compression)

Minimize LLM token complexity to save API costs and model computations.

- Simplify Token Usage: Token complexity refers to the number of prompt tokens needed for a specific task. Lowering this number directly decreases API expenses and significantly reduces the computational workload for typical transformer models.
- Cost Reduction: For major corporations, a 10% decrease in token usage can result in a substantial cost saving of around $100,000 for every $1 million spent.
- Overcoming Model Constraints: Certain models are limited by *short context lengths*. Using prompt optimizers can enable these models to handle documents larger than their usual context size.

Prompt	Tokens	Is Response Correct (3.5 Turbo) ?
Who is the president of the United States of America?	11	✓
Who president US	3	✓

Platform Approach



Any