WSO2CON 2024

Architecting AI in the Enterprise: APIs and Applications



Malith Jayasinghe

VP of Research and AI, WSO2



Enterprise AI: Growth Forecast 2022–2030

- Global enterprise AI market was valued at $136.55 billion in 2022 (Forbes, 2022)
- Expected to grow at a CAGR of 36.6% from 2024 to 2030, reaching $1,811.8 billion by 2030 (Grand View Research, 2024)
- By 2026, over 80% of enterprises are anticipated to have used generative AI APIs/model and/or have deployed generative AI application in production (Gartner, 2023)
- More companies are experimenting with generative AI; however, the percentage of applications entering production is low (5% in 2023) , though it is increasing (Gartner, 2023)

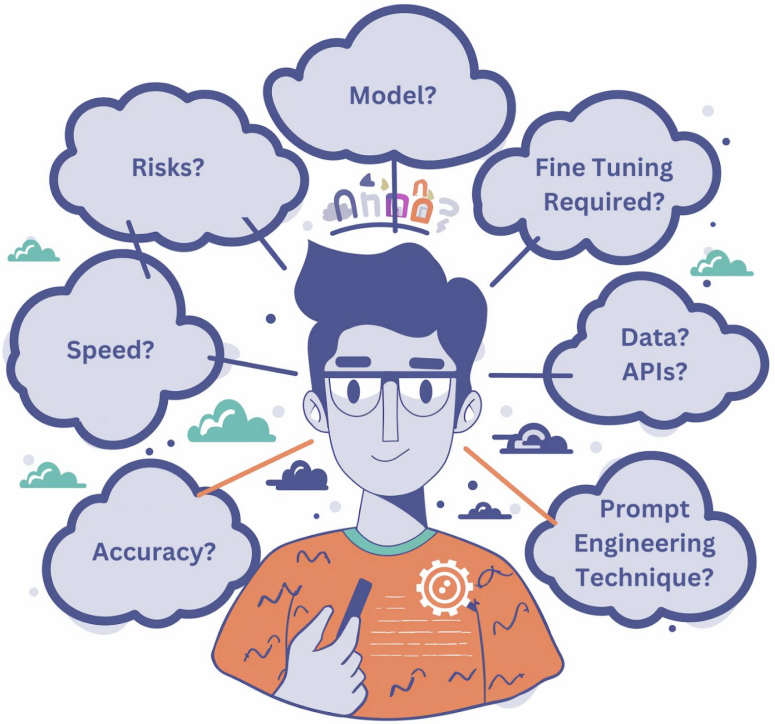
Value Proposition in AI Development

- Understanding AI's value is key to unlocking its potential. Strategic integration of AI can enhance efficiency, improve user experience, provide competitive advantages, and prepare organizations for the future

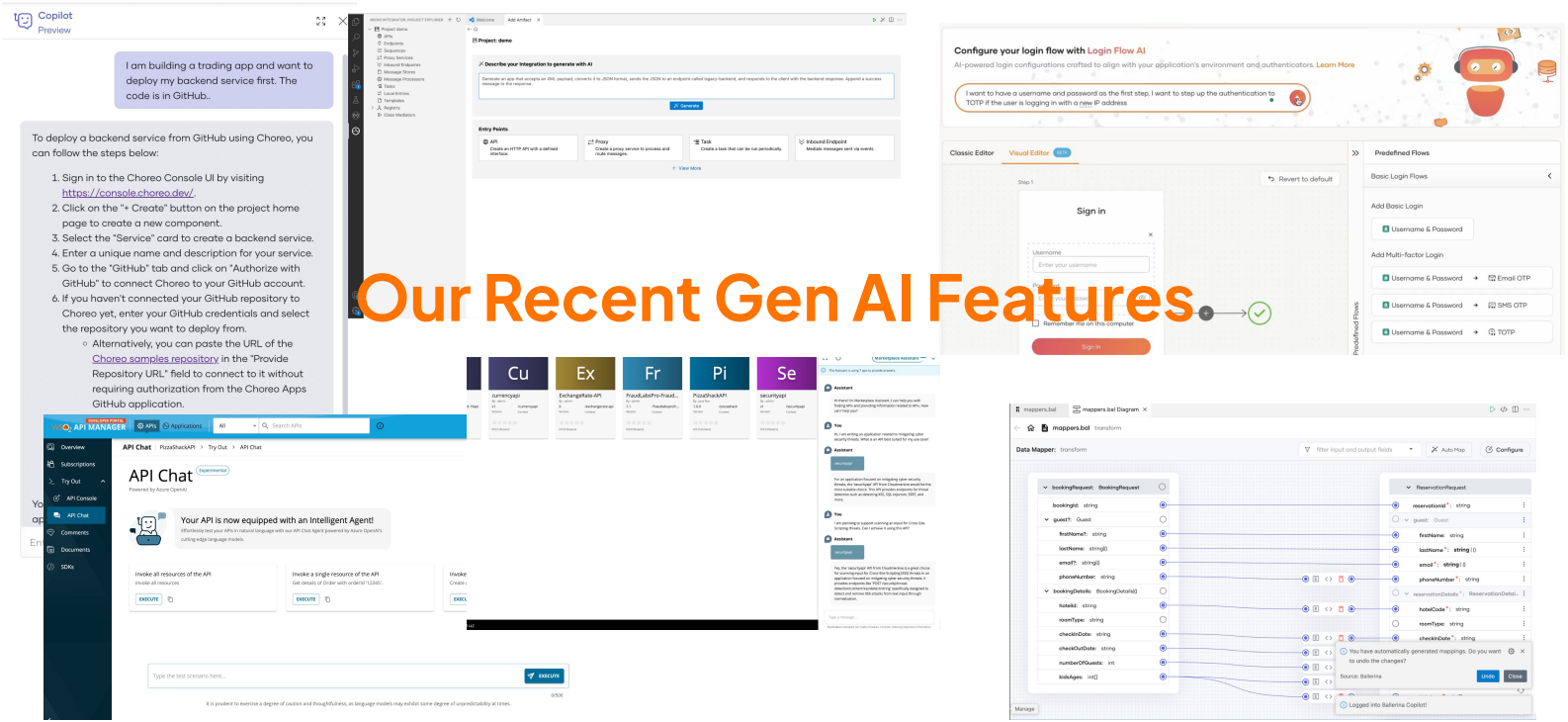


AI Development Challenges

- Objective: Provide insights into potential challenges and strategies for addressing them (Specific Focus: generative AI applications)
- Challenges: As you develop your AI application and attempt to move it into production, you will need to address many challenges
- UI/UX Considerations
  - Initiate design early
  - Consider design aspects deeply



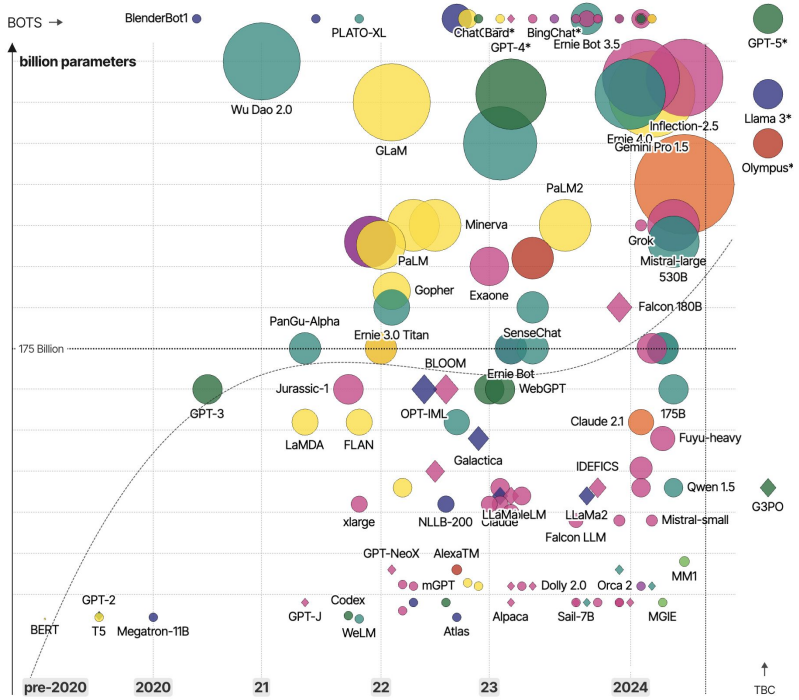
AI Development Challenges



Our Recent Gen AI Features

Selecting the Right Model

- Model selection criteria
  - Accuracy (Accuracy vs. speed?)
  - Cost (Accuracy vs. cost?)
  - Speed (Accuracy vs Speed?)
  - Risks
  - Other Considerations: Self-hosted model or Public cloud hosted, Model availability in specific regions
- Dependency on use case
- Trade-offs in model selection
- Given AI feature can use multiple models



Source: https://informationisbeautiful.net/

Identifying Data Sources

- Model improvements (build a RAG, Fine tune, pre-train)
- Identify the available data sources (e.g., internal R&D, product documentation)
- Data cleaning and preprocessing
  - Conduct data cleaning
  - Data leakage: LLM accidentally reveals sensitive information (PIIs, proprietary algorithms)
- Do I need to collect new data?



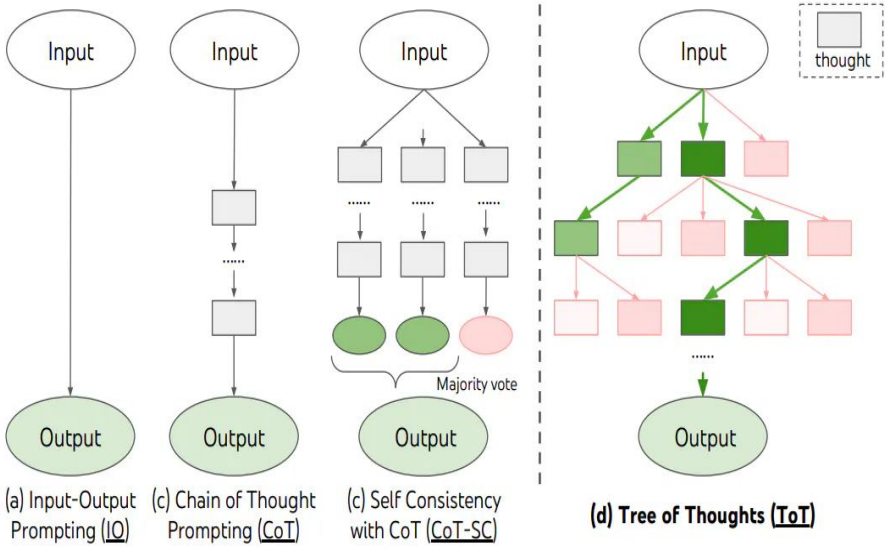
AI Interaction with APIs

- What are the available APIs?
- APIs can differ in type and have different representations
  - RESTful APIs
    - OpenAPI specification
  - GraphQL APIs
    - GraphQL Schema
- Well-documented APIs perform better when integrated with AI
- AI can bridge the gaps in missing documentation

```yaml
/pet:
  post:
    tags:
      - pet
    summary: Add a new pet to the store
    description: ''
    operationId: addPet
    consumes:
      - application/json
      - application/xml
    produces:
      - application/json
      - application/xml
    parameters:
      - in: body
        name: body
        description: Pet object that needs to be added to the store
        required: true
        schema:
          $ref: '#/definitions/Pet'
    responses:
      '405':
        description: Invalid input
    security:
      - petstore_auth:
          - write:pets
          - read:pets
```

Prompt Engineering

- Prompting engineering
  - Prompt engineering is guiding generative AI to produce desired outputs
  - A prompt is a natural language instruction describing how generative AI should perform (acts as the interface between human intent and machine output)
- Prompting
  - Zero-shot prompting, Few-shot prompting
  - Chain-of-Thought (COT), Tree of Thoughts (TOT)
  - ReAct (Reason and Act)



https://www.promptingguide.ai/

Prompt Engineering: ReAct (Reason and Act)

- LLM generation provides information, guidance, or suggestions. Their output doesn't (inherently) drive actions
- ReAct Agents can perform actions (e.g., calling APIs)
- ReAct Agents with the assistance of the model, can execute natural language commands using a collection of tools registered with them
- Tools can retrieve data or perform tasks, and they come in different forms, such as functions, API, etc.

Use Case: Train Booking System



Users

Chat UI
(Web site)

Train Booking Service
(Agent Service)

invokes

Existing Business APIs

Train API

Email API

Payment API

Initialization

Swagger Specs
(OpenAPI)

Natural Language Query:

Book me the next train to Galle, and send an email confirmation

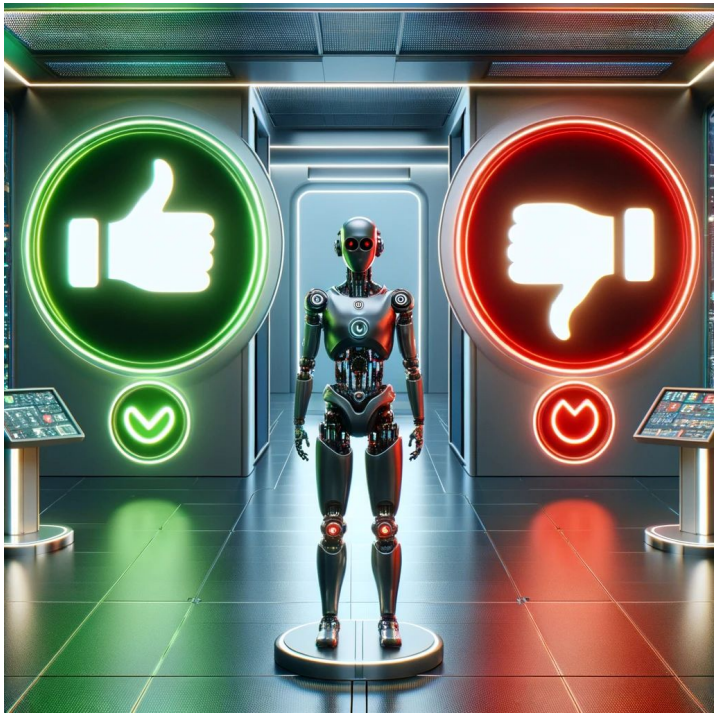
Evaluating Accuracy

- Create a dataset for evaluation purposes
- Select a metric (use existing one or develop new metrics if necessary)
- Automate the evaluation of accuracy
- Re-evaluate after any changes to confirm that accuracy has not diminished
- Even minor changes to prompts can significantly affect accuracy

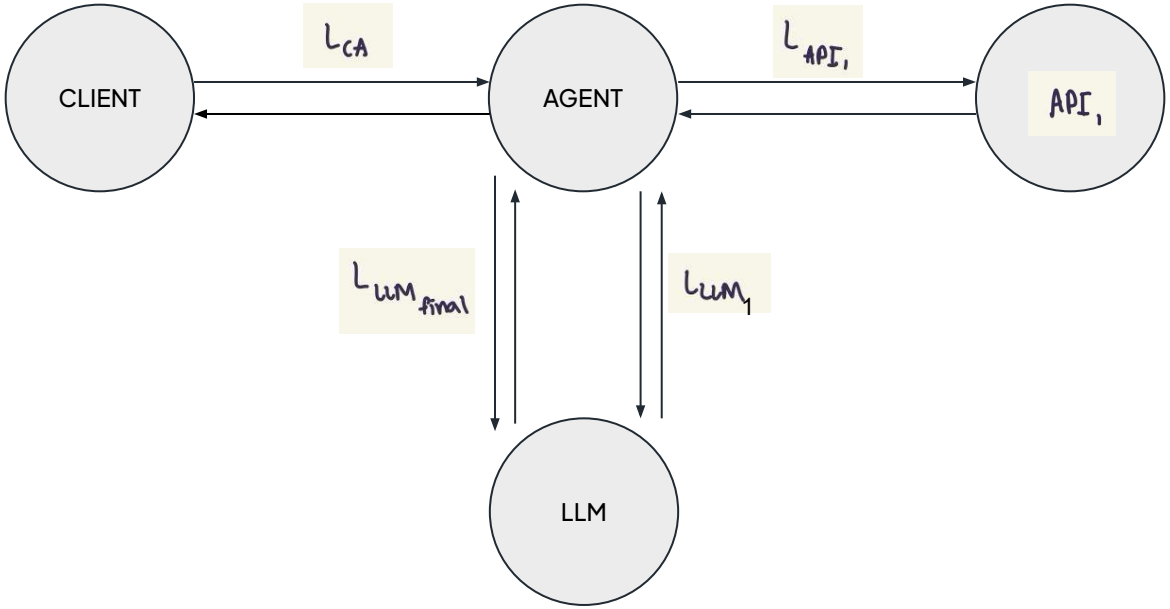


Collecting User Feedback + Improving the Accuracy

- Collect data: Gather feedback on AI outputs, user interactions, and engagement metrics
- Compliance: Ensure user consent and follow privacy laws. Anonymize sensitive data
- Improve models: Update vector DBs, fine-tune models, improve prompts



Performance (Latency, Throughput)

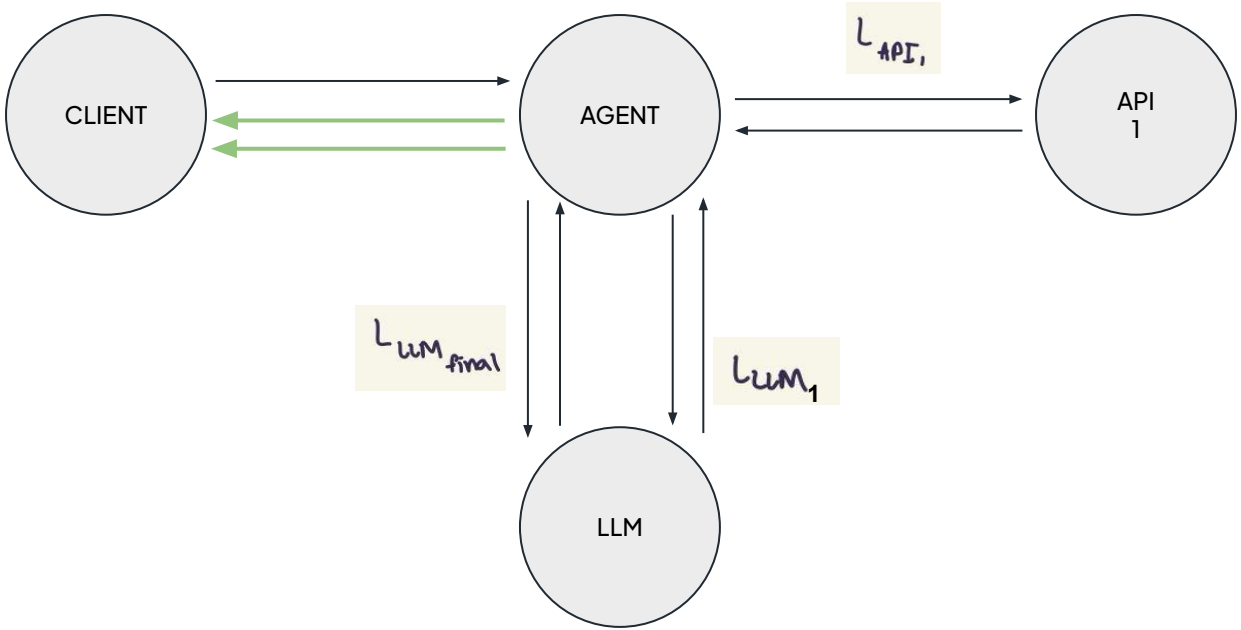


$$L_{Total} = L_{LLM} + L_{API_1} + L_{LLM_{final}} + L_{CA}$$

Enhancing User Experience for Tasks with Long Processing Times

- Complex Tasks: AI systems execute complex tasks that require multiple reasoning iterations and involve invoking multiple APIs
- Optimizations: Techniques such as caching are used, but latency can still be high
- User Satisfaction: Prolonged round-trip times can result in user dissatisfaction
- Strategies to Improve User Experience:
  - Incremental Display: Progressively display partial results as AI systems process different APIs, thereby reducing wait times for final outcomes

Enhancing User Experience for Tasks with Long Processing Times



Closing Remarks

- Purposeful Beginnings: We start with a clear definition of the problem and a keen understanding of the application's value, guiding our development journey from the outset
- User-Centric Design: Prioritizing UI/UX considerations ensures that our applications are accessible, intuitive, and meet user expectations
- Technical Execution
  - Model Selection/Model Improvements
  - Dataset Utilization/API Integration
  - Prompt Engineering
  - Accuracy evaluation/Continuous feedback
  - Scalability and Performance
- Forward Outlook: The development of generative AI applications is a dynamic, ongoing process requiring constant learning and adjustment to meet evolving technological trends and user demands

Question Time!



Thank You!