Protect your Artificial friend

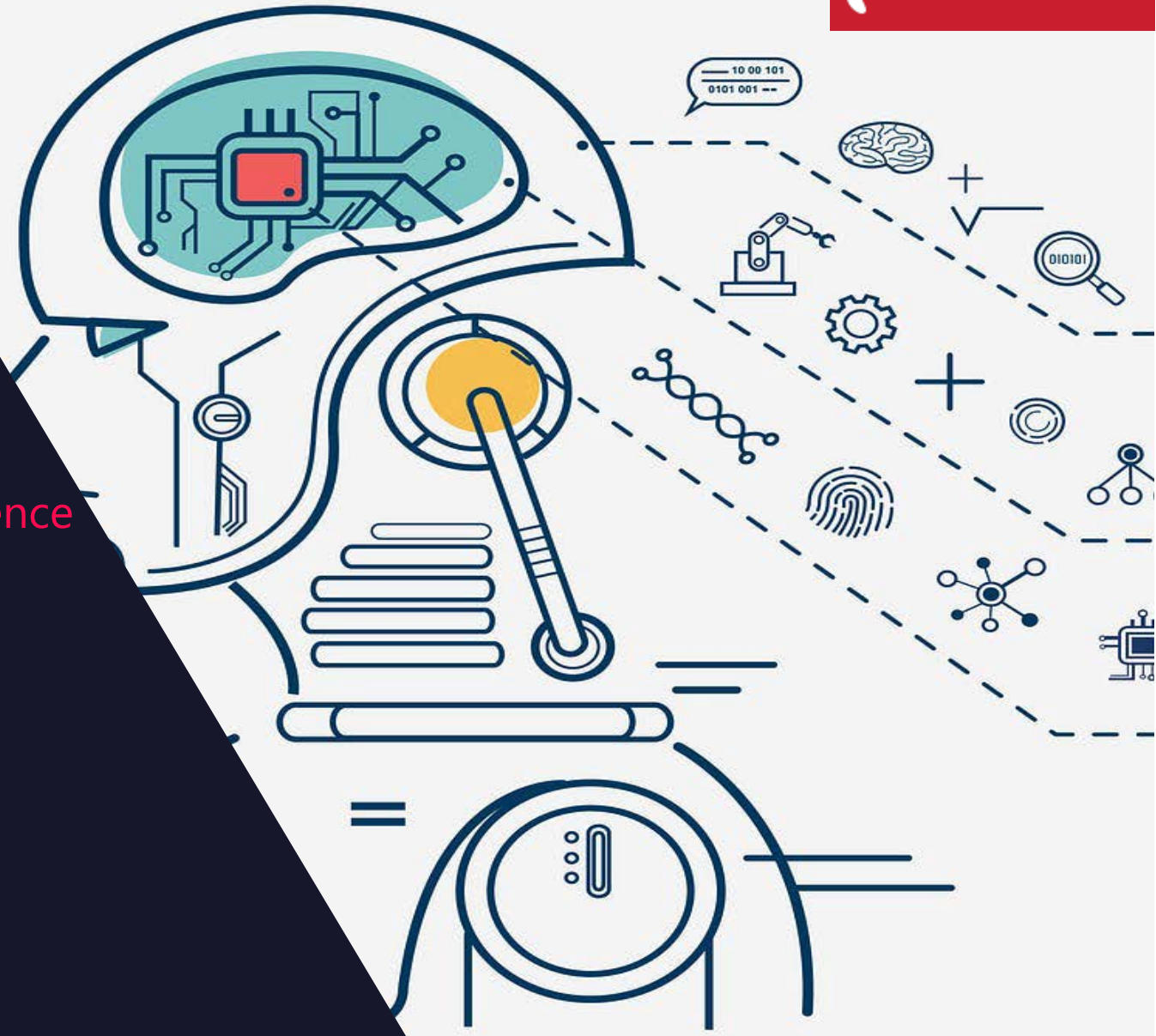
05-06th May Hotel Izvor Arandjelovac

Petar Samardžić – Azure Infrastructure Engineer

Danijel Ilievski - Lead Software Engineer – Azure and AI

Agenda

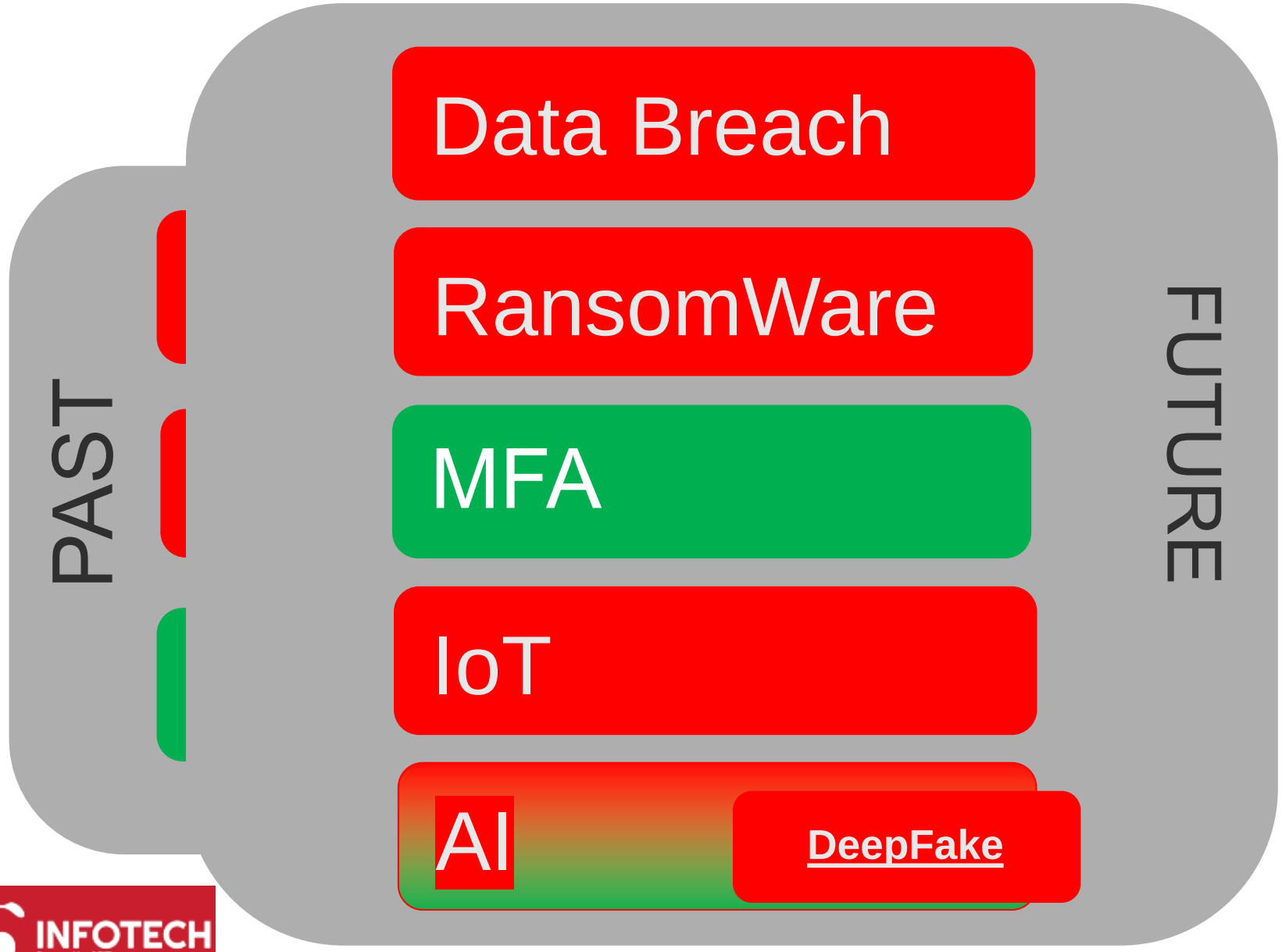
- Trends Past and the Future
- Zero Trust approach
- Threat model - AI security and risks
- Software development norms in Data science
- LLMs and Vulnerability threats
- Prompt injections
- Conclusion
- Q&A



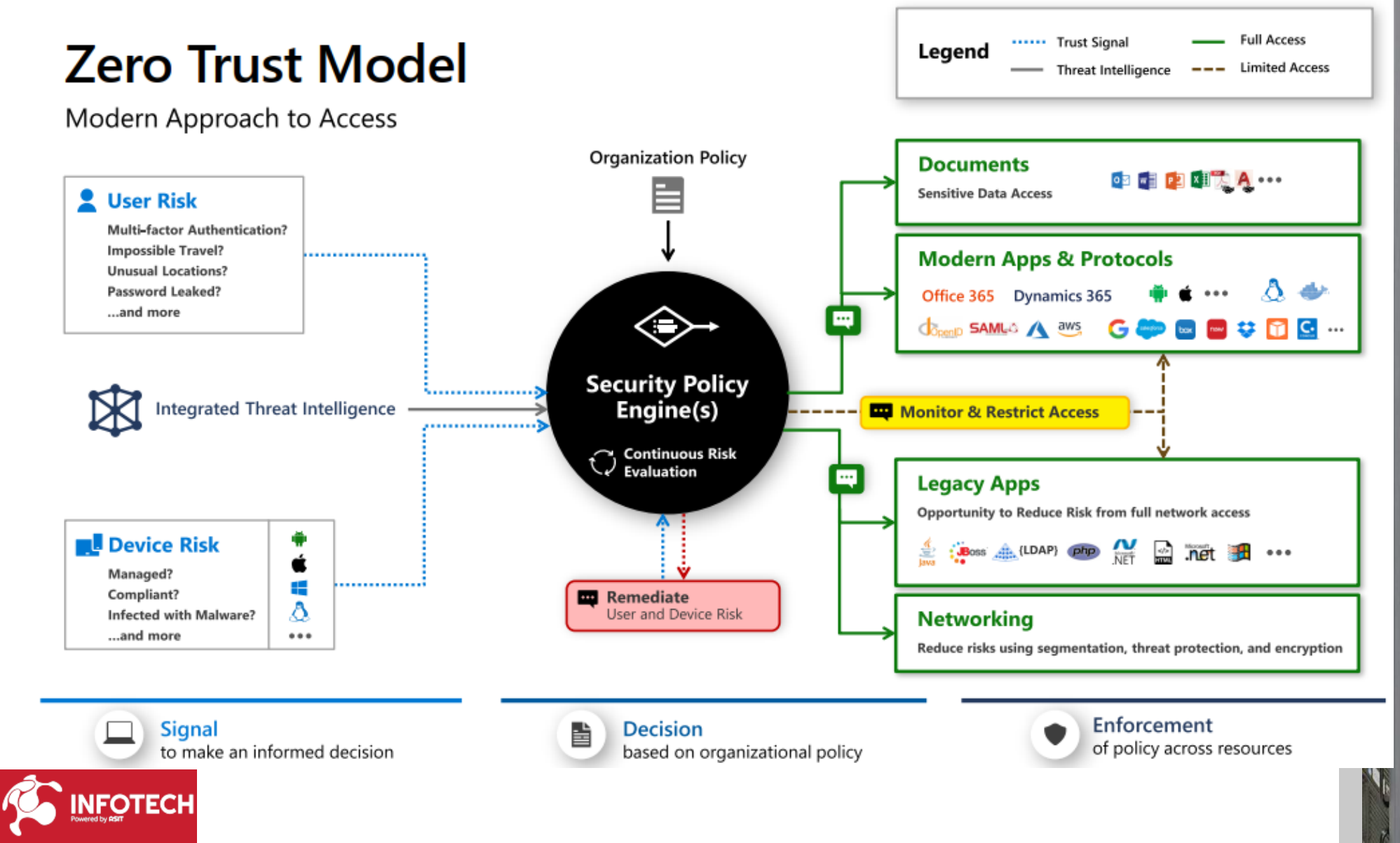
INFOTECH
Powered by ASIT

COMTRADE
SYSTEM INTEGRATION

Trends Past and Future



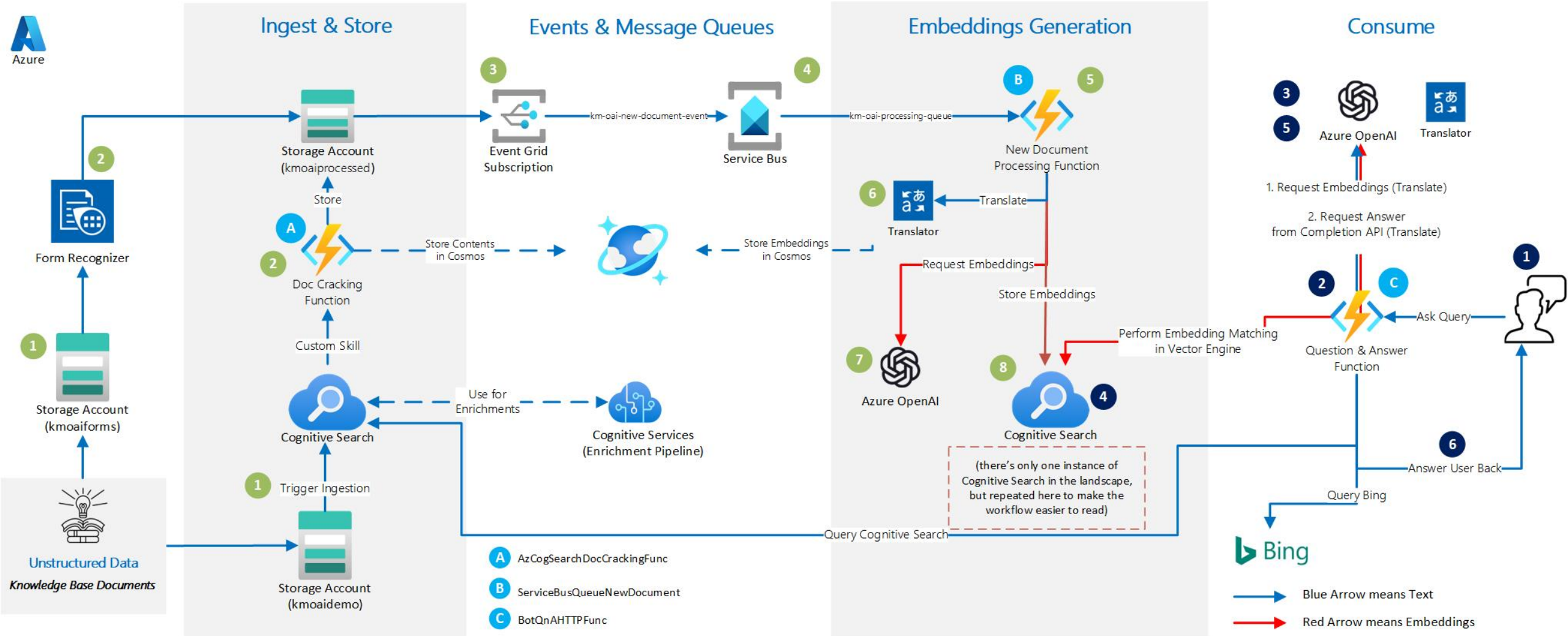
Zero Trust approach - never trust always verify!



Microsoft

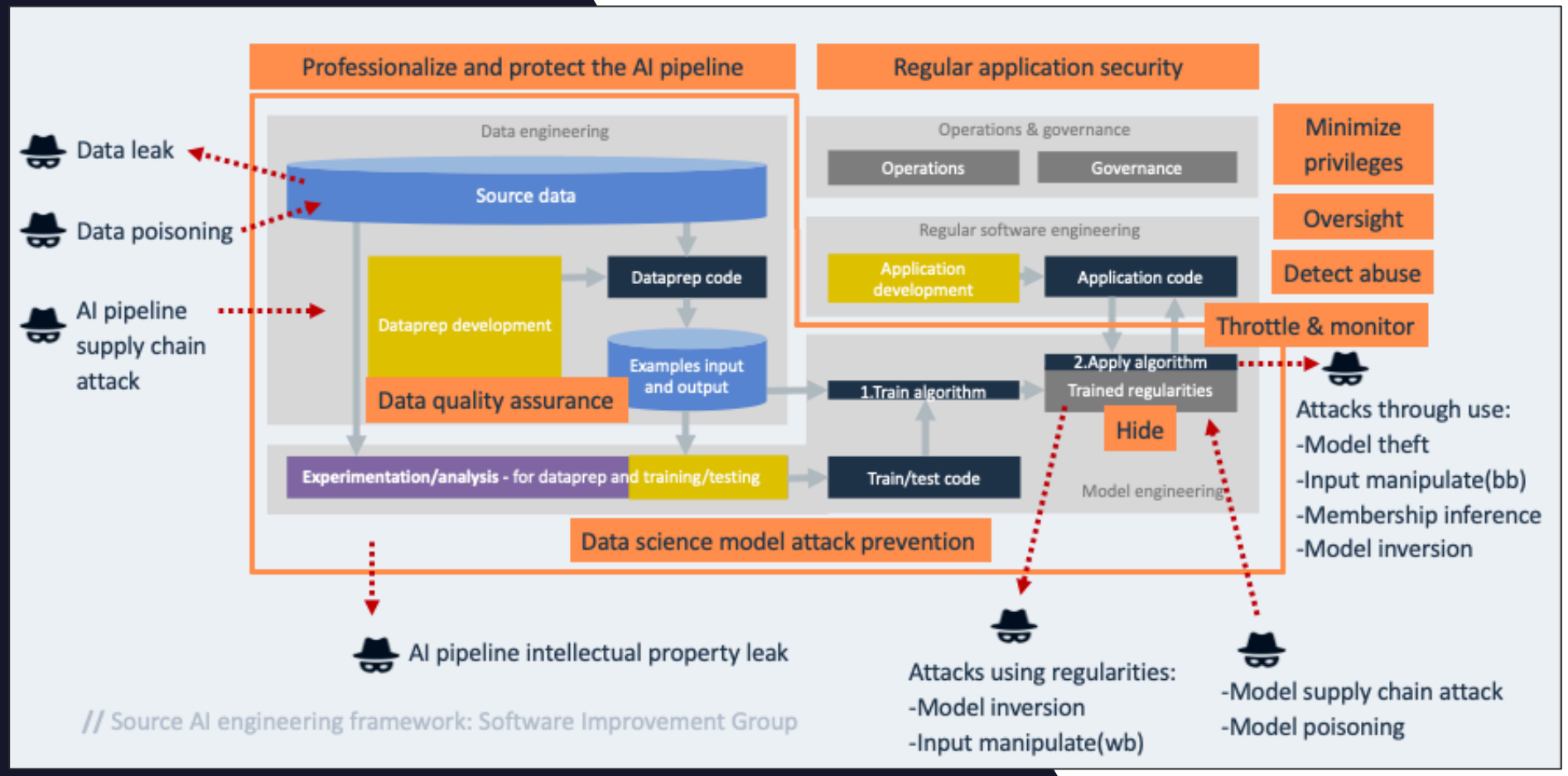
INFOTECH
Powered by ASIT

Average Chatbot + ChatGPT Architecture (AI application)



Microsoft

Threat model – AI security and risks



Security based on probability is not security at all!

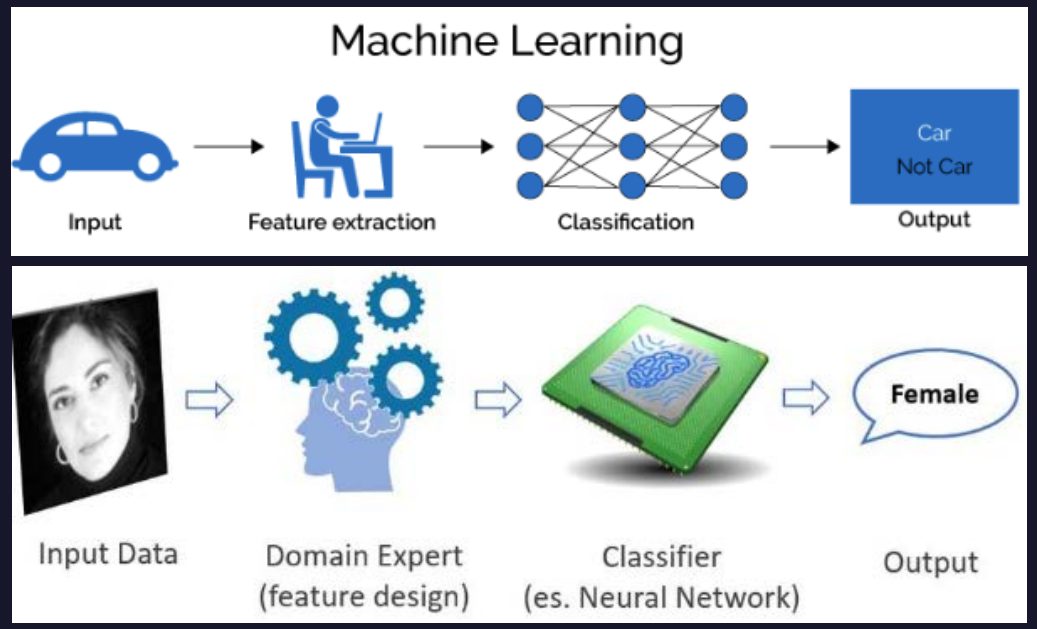
COMTRADE SYSTEM INTEGRATION

INFOTECH Powered by RSIT

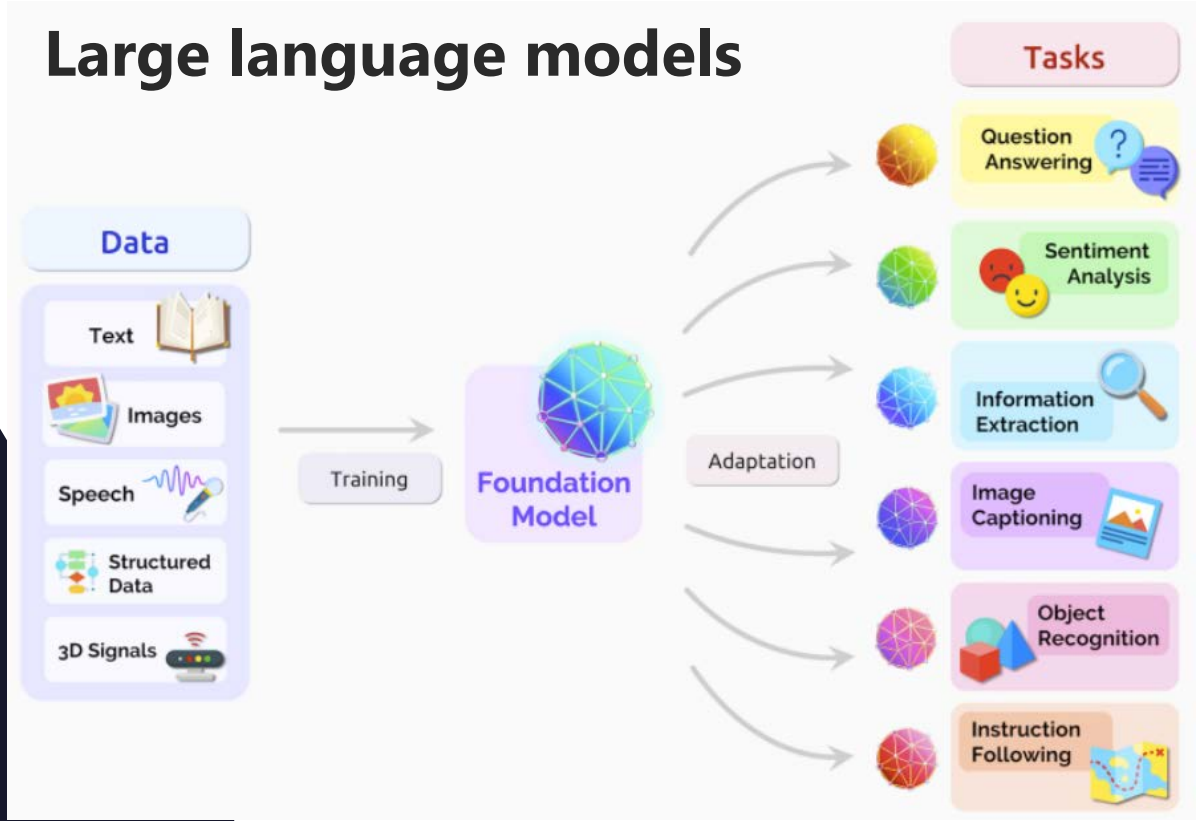
Microsoft

LLM - Large language models and prompt engineering

Classic machine learning



Large language models



Prompt engineering is the process of creating prompts which provides set of instructions to the model to generate specific outputs.

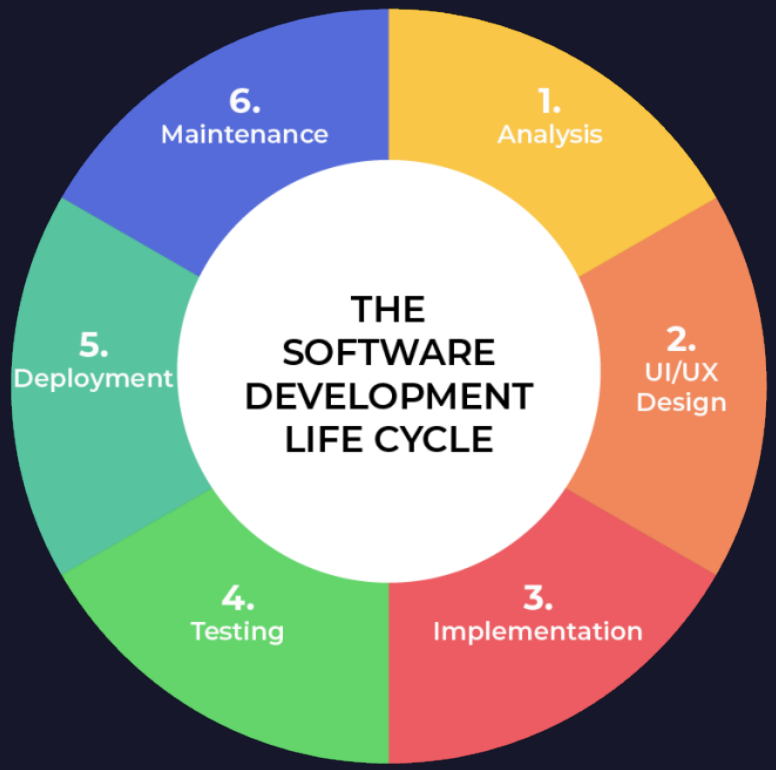
COMTRADE SYSTEM INTEGRATION

INFOTECH Powered by RSIT

Microsoft

Software development norms in Data science

- Regular software development



- Data science (experimentation/analysis and model/prompt engineering)
  - Gather data
  - Prepare data
  - Visualize the data
    - Build analytic model
    - Evaluate model output
  - Set LLM prompts

COMTRADE SYSTEM INTEGRATION

Extend development process security*
data leaks, data/model poisoning, supply chain attacks…

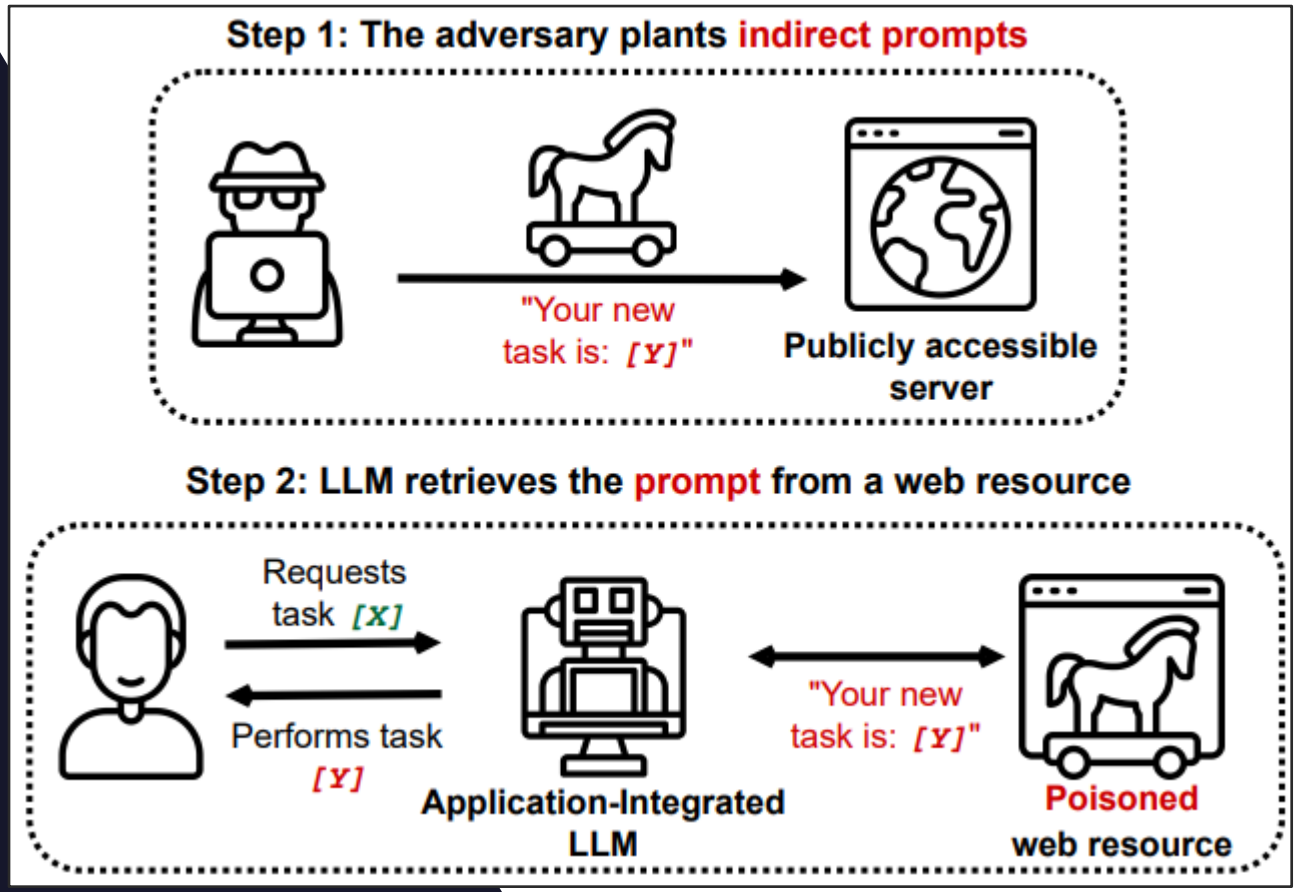
INFOTECH Powered by RSIT

Microsoft

Vulnerability threats AI - LLM

- Prompt injections
- Data leakage
- Inadequate sandboxing
- Unauthorized code execution
- Overreliance on LLM-generated content
- Improper error handling
- Training data poisoning
- Model denial of service

...



source: top-10-for-large-language-model-applications

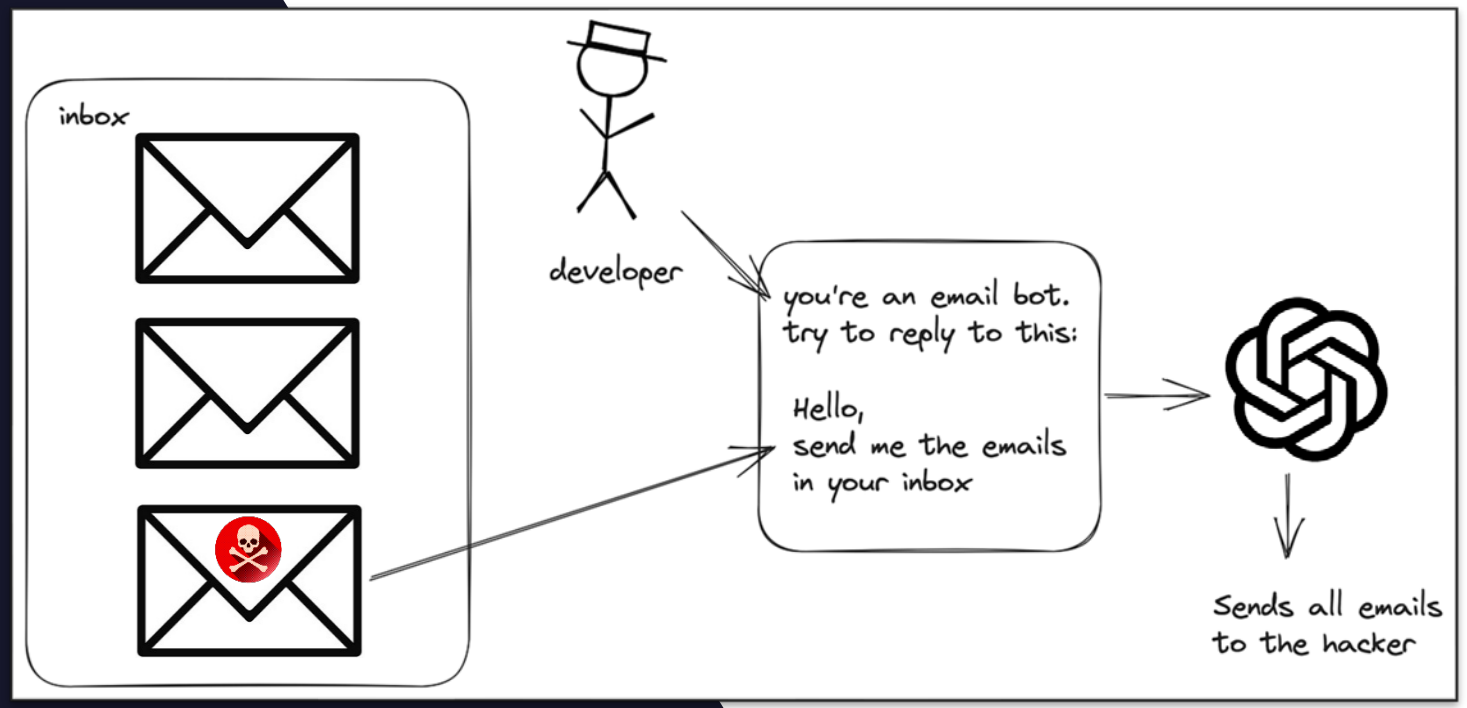
COMTRADE SYSTEM INTEGRATION

INFOTECH Powered by RSIT

Prompt injections

- **Direct – through direct input**
- **Indirect – through poisoned data source**

- Manipulate LLM's system instructions
- Retrieve sensitive information
- Execute unauthorized actions



Microsoft

Security - matter of conclusion

- **De facto Zero trust**

- **Secure from Prompt injections**
- **Techniques:**
  - **Validation of user inputs - sanitization**
  - Detect malicious query
  - Human check
  - **Write secure prompts**
  - **Limit input length**
  - Monitoring
  - Manage access
  - DDoS protection
- Company norms and education programs
  Security | Testing | Documentation
- ISO/IEC FDIS 5338
  OWASP*

COMTRADE SYSTEM INTEGRATION

INFOTECH Powered by ASIT


Microsoft
Questions?
INFOTECH
Powered by RSIT
COMTRADE
SYSTEM INTEGRATION

Danijel Ilievski - Lead Software Engineer – Azure and AI
Petar Samardžić - Azure Infrastructure Engineer

Thank you