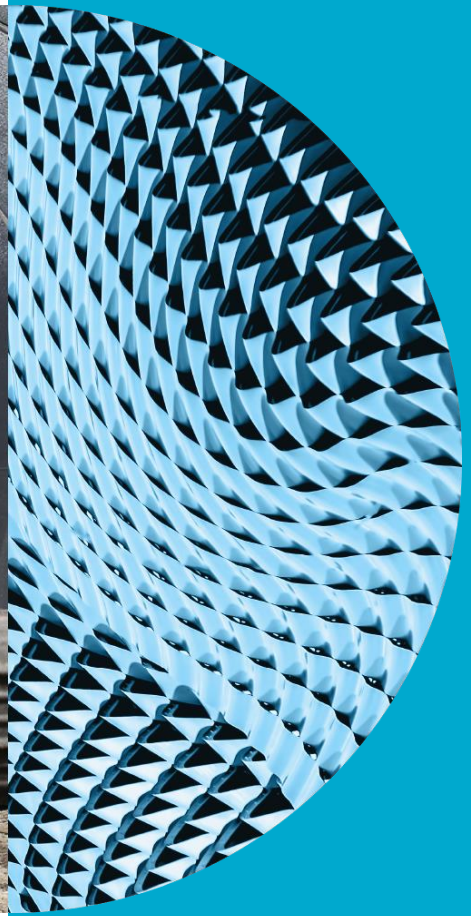
Safe and Responsible AIware Engineering



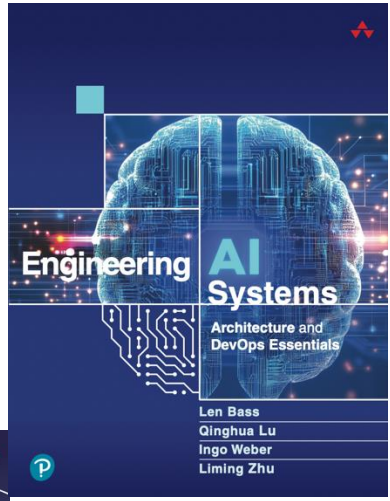
Qinghua Lu

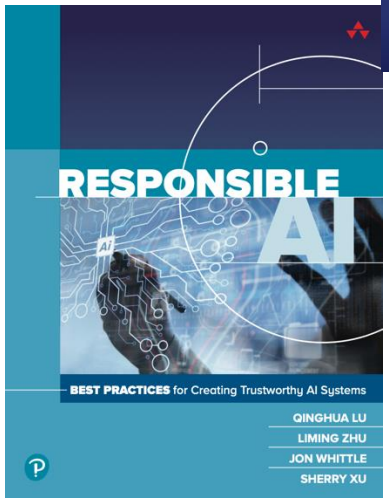
Responsible AI Science Team Leader
Data61, CSIRO

qinghua.lu@data61.csiro.au

CSIRO and Responsible AI Team

- CSIRO
  - Australia's national science agency
  - Formed in 1916
  - 5500 people
  - 50 sites (Australia, France, Chile, US)
  - Data61: Data and Digital RU
- Responsible AI team
  - Formed in 2022
  - ~30 full time research scientists/engineers
  - Diverse and multidisciplinary team
  - 6 scientists in the top 30 for Responsible AI





What is an AI system?

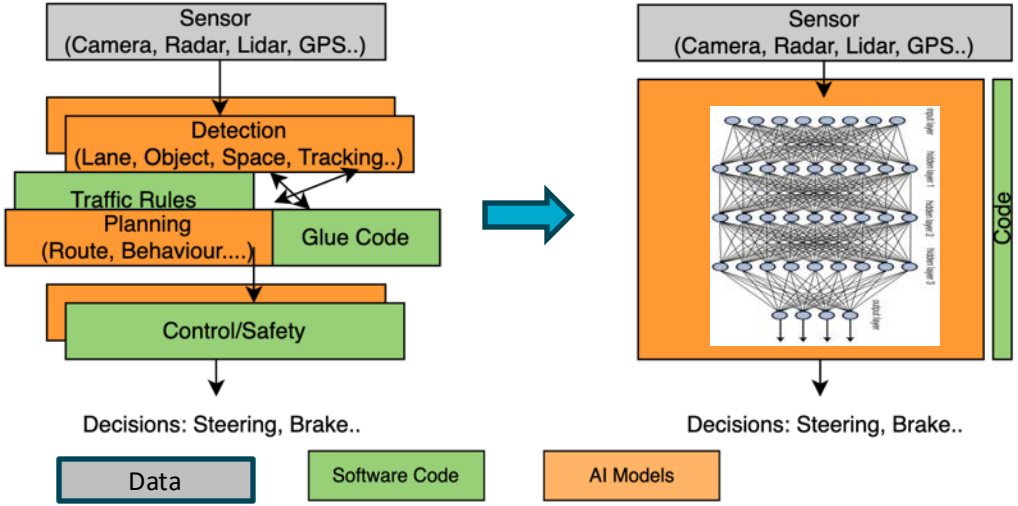
- *An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their level of autonomy and adaptiveness after deployment*. (OECD, 2023)

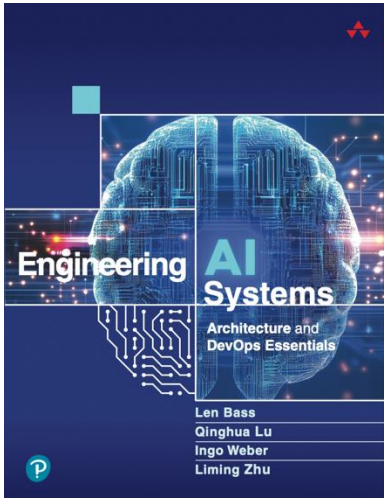
- *Artificial Intelligence (AI) is the research and development of mechanisms and applications of AI systems. (ISO 22989).*

What is an AIware?

- *AI-as-Software, also known as AIware, refers to AI systems where functions are primarily encapsulated within a single general AI model as parameters/weights, rather than distinct narrow AI models explicitly chained together by traditional business code logic. (Bass, 2025)*

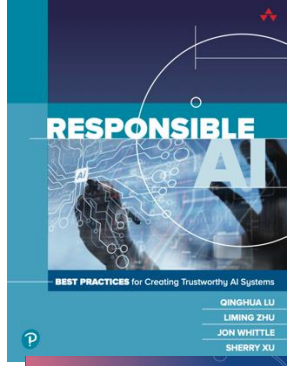
End-to-End AI: Data In, Decision out, No Code

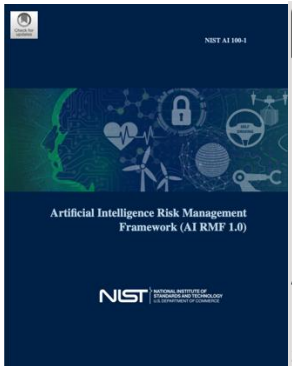


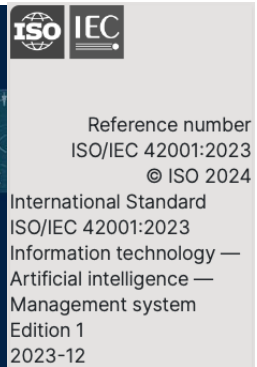


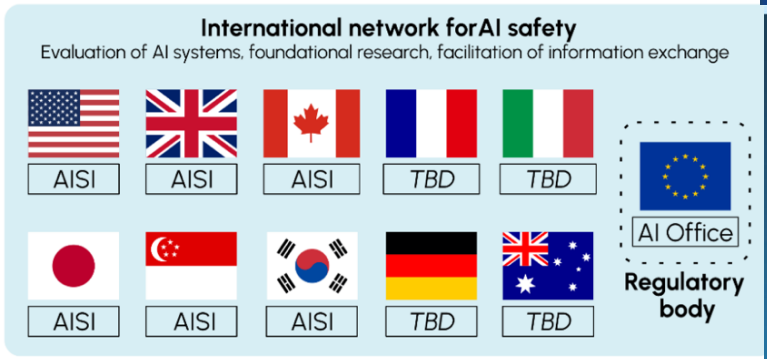
What is Responsible/Safe AI?

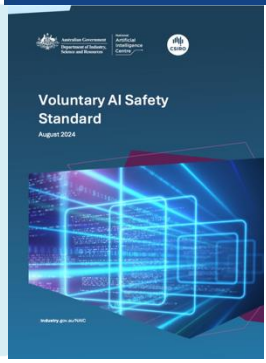
- ***Responsible AI*** *is the* ***practice of developing and using*** *AI systems in a way that provides benefits to individuals, groups, and wider society, while minimizing the risk of negative consequences.(Lu, 2023)*
- ***AI safety*** *is often used to describe* ***prevention of or protection against AI-related harms****.* (Bengio, 2024)

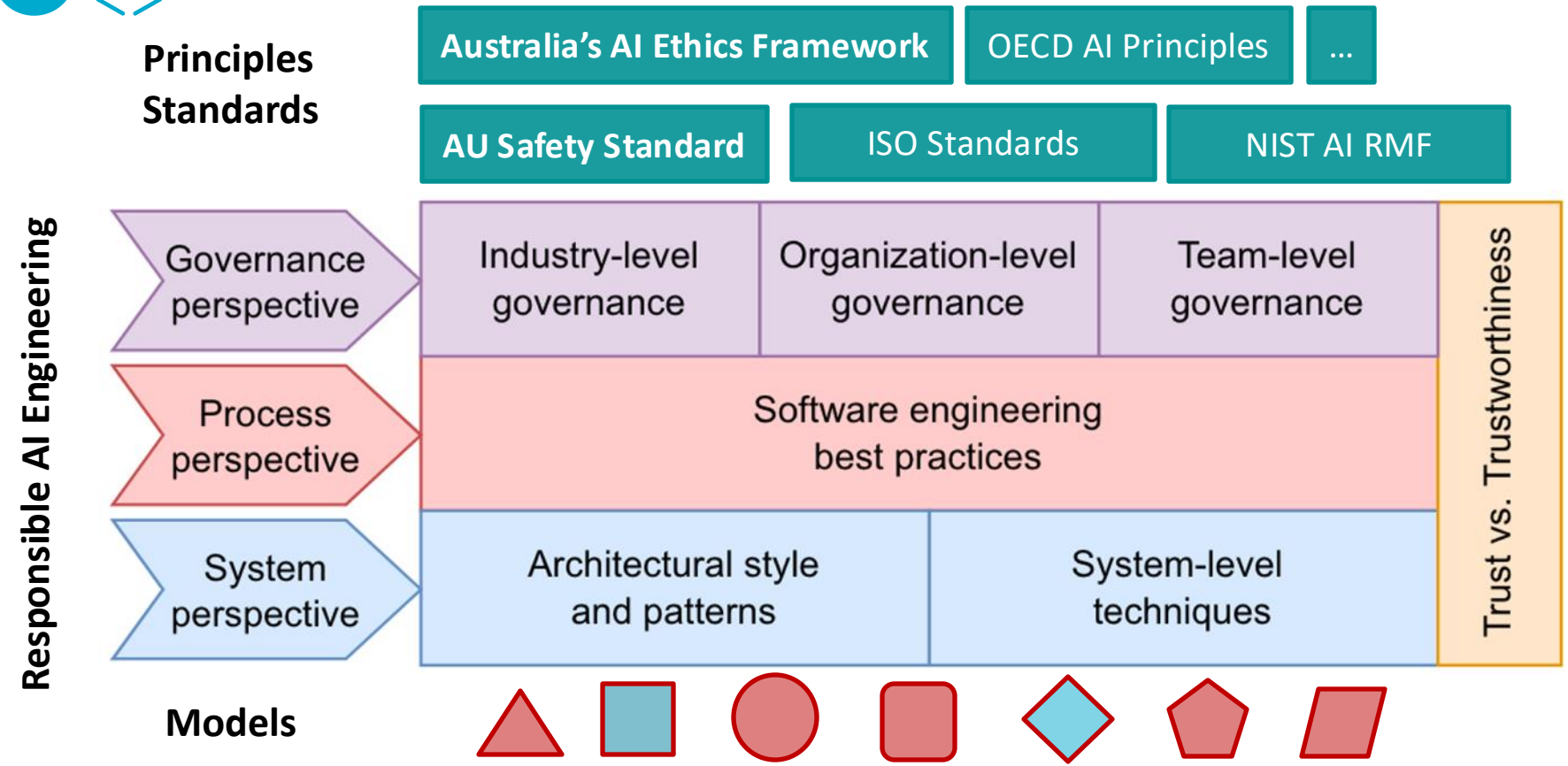






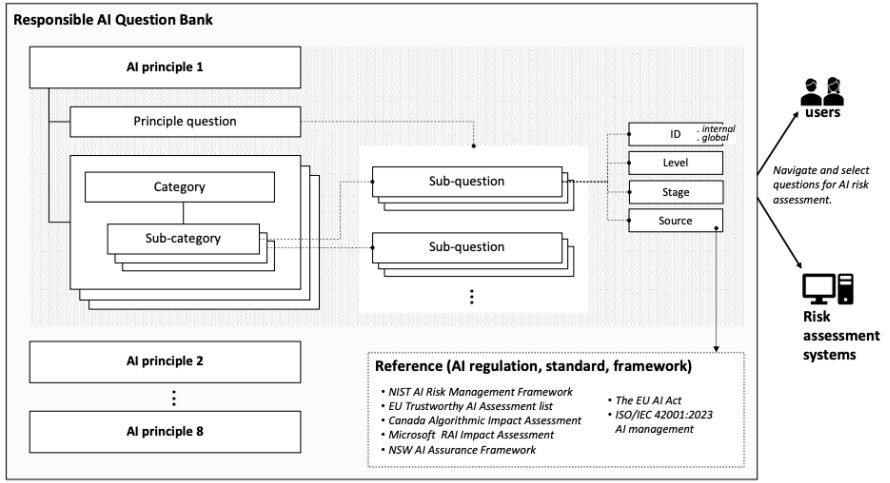






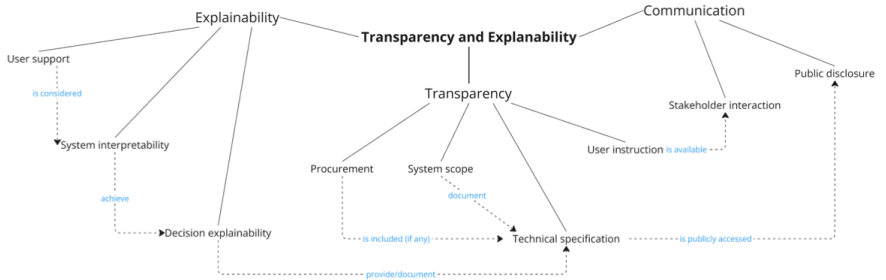
Lu, Q., Zhu, L., Xu, X., Whittle, J., Xing, Z., 2022. Towards a Roadmap on Software Engineering for Responsible AI, in: 1st International Conference on AI Engineering (CAIN)

Responsible AI Question Bank



Transparency and Explainability	Do you disclose information about AI systems, including related risks and opportunities, and inform stakeholders about the use of AI?

Concept map:



Sub-question (examples):

Category		Question	Level	Stage	Source
Explainability	System interpretability	Do you design the AI system with interpretability in mind from the start?	2	D	EU
Transparency	Technical specification	Do you comprehensively understand and document the AI system including intended purposes, potentially beneficial uses, etc.?	2	P	NIST
Communication	Stakeholder interaction	Do you establish processes that consider users' feedback and use this to adapt the system?	2	P	EU

P: planning, R: requirement, D: design, I: implementation, T: testing, DE: deployment, O: operation

Lee, S.U., Perera, H., Liu, Y., Xia, B., Lu, Q., Zhu, L., 2024. Responsible AI Question Bank A Comprehensive Tool for AI Risk Assessment. https://arxiv.org/pdf/2408.11820

Responsible AI Metrics Catalogue

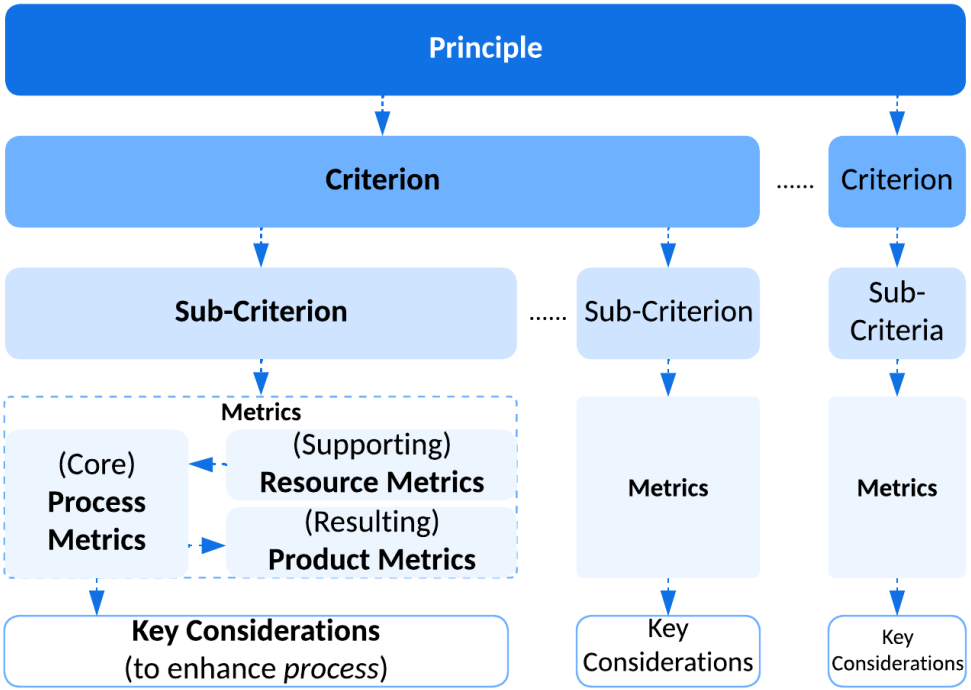
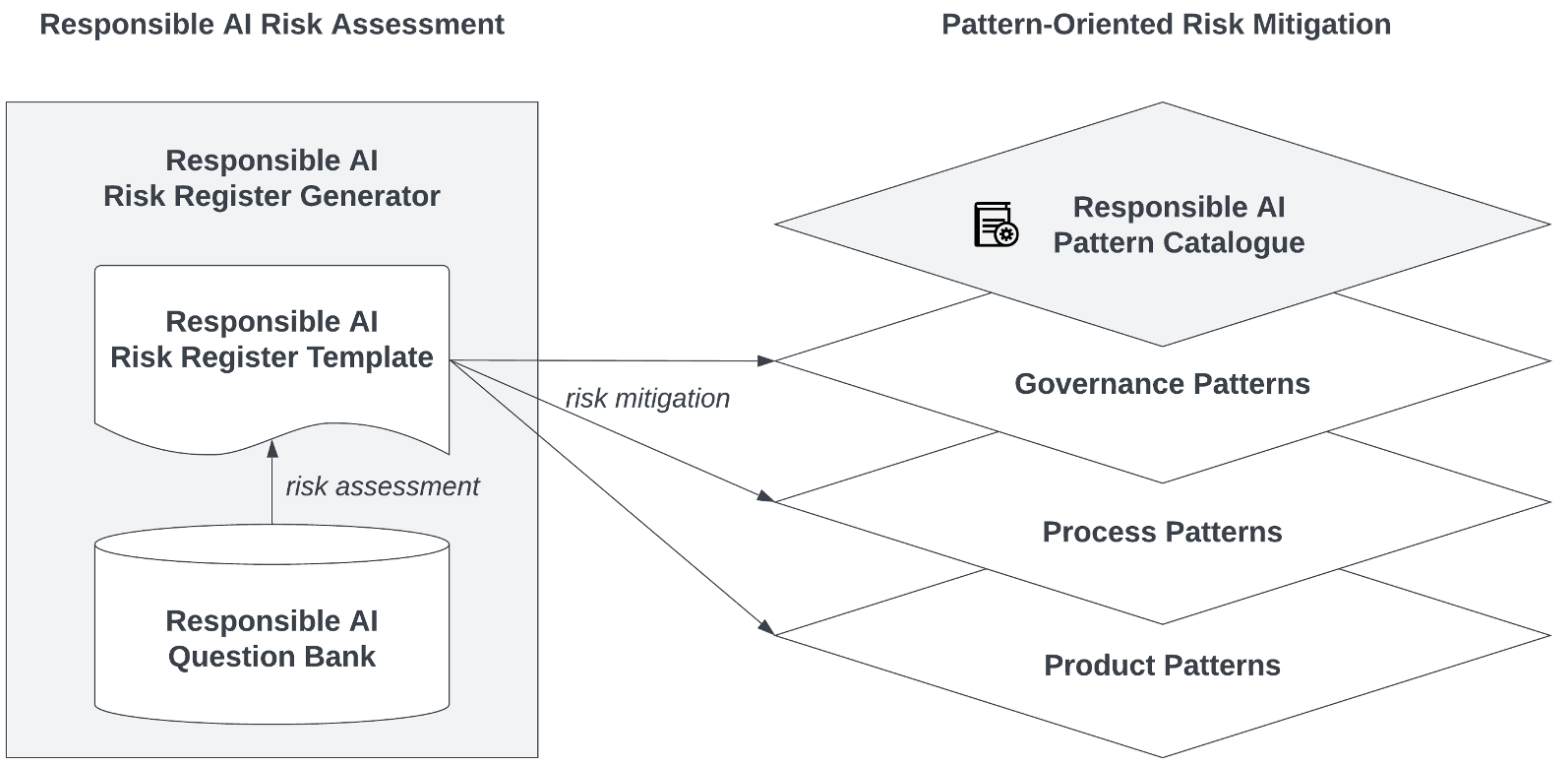


Table 2: System-Level Metrics Catalogue for AI Accountability

Criteria	Sub-Criteria	Process Metrics	Key Considerations	Resource Metrics	Product Metrics
Responsibility	RAI Oversight	Roles and Responsibilities	• Comprehensive role clarity: - Design and development - Deployment and operations - Procurement and integration - Governance and compliance - AI as a service	• Soft laws (e.g., best practices, guidelines standards etc) • Hard laws (e.g., EU AI Act)	• Procedure Manuals • Contracts or agreements • Position descriptions • Recruitment practices • Workforce dev strategy
		RAI Governance Committee	• Multidisciplinary composition • Strategic leadership involvement		• Policy doc on Committee
		Organizational AI Risk Tolerance	• Tiered risk-based categorization • Balancing competing interests		• Policy doc on org's risk tolerance and mitigations
	RAI Competence	RAI Training	• Holistic training content • Targeted training for diverse roles • Adaptive and ongoing education		• Training certificates
		RAI Capability Assessment	• Multifaceted assessment • Standard alignment • Organizational RAI maturity • Continuous enhancement		• Assessment reports
Auditability	Systematic Oversight	Data Provenance	• Detailed data record-keeping • Data version control • Data integrity and risk mitigation • Legal and ethical compliance	• Soft laws (e.g., auditing guidelines and frameworks etc) • Hard laws (e.g., EU AI Act) • AI documentation tools (e.g., datasheets, model/system cards) • Technical tools (e.g., blockchain, knowledge graph)	• Provenance records • System features (e.g., auto-logging, version control
		Model Provenance	• Detailed model record-keeping • Model selection and validation • Model version control		
		System Provenance and Logging	• Detailed system record-keeping - System version control - Decision/Trade-off • Comprehensive operational logging - User interaction and system response - Incident and response - System configuration changes • Composition Management		• Provenance records (and logs) • System features (e.g., auto-logging, version control
	Compliance Checking	Auditing	• Diversified auditing strategy • Multi-dimensional audit techniques • Ethical and legal compliance • Regular audits • Verifiable audits • Audit-driven improvements		• Audit reports • Compliance certificates and licenses
Redressability	Redress-by-Design	Incident Reporting and Response	• Accessibility and Visibility • Structured Incident Management • Feedback Loop Integration	• Redundancy design case studies • Incident management tools	• Incident and response doc • System features (user feedback and report)
		Built-in Redundancy	• Multi-Modal Redundancy		• System features (redundant components/functionalities)

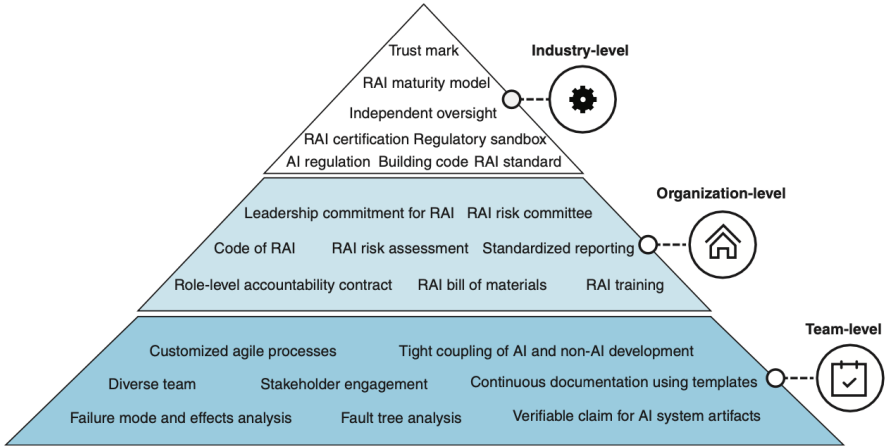
Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, Zhenchang Xing, Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability, submitted to CAIN'24.

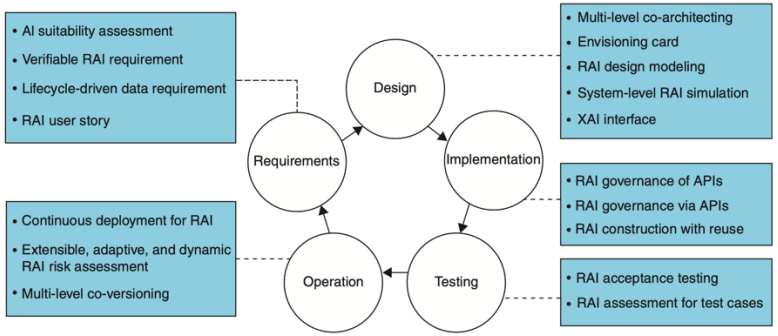
Responsible AI Pattern Catalogue

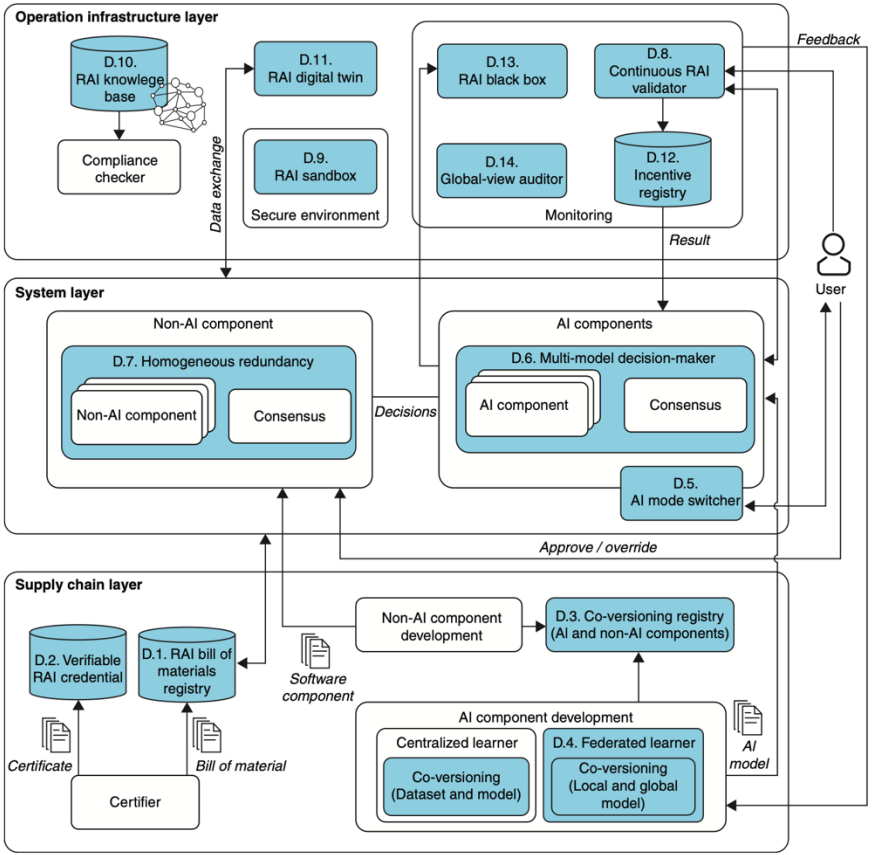


https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue /

Responsible AI Pattern Catalogue







https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue /

D.5. AI Mode Switcher

Adding an AI mode switcher to the AI system offers users efficient invocation and dismissal mechanisms for activating or deactivating the AI component when needed.

Context

Human autonomy is an individual's capacity for self-determination or self-governance, which should be supported in AI systems.

Problem

How can we enable human autonomy by allowing users to efficiently activate and deactivate the AI component when needed?

Solution

When to use AI at decision-making points can be a major architectural design decision when designing an AI system. In Figure 6.6, adding an AI mode switcher to the AI system offers users efficient invocation and dismissal mechanisms for activating and deactivating the AI component whenever needed, thus deferring the architectural design decision to the execution time that the end user or the operator of the AI system decides. The AI mode switcher is like a *kill switch* for the AI system that could immediately shut down the AI component and thus stop its negative effects (e.g., turning off the automated driving system and disconnecting it from the internet).

Benefits

Here are the benefits of the AI mode switcher pattern:

- **Increased trust:** An AI mode switcher gives users the choice to switch off the AI model when they do not trust the decision or recommendation provided by the AI component, thus increasing trust toward the AI system.
- **Contestability and autonomy:** The AI mode switcher enables human autonomy by allowing end users to switch off the AI component or override the decisions made by the AI component at runtime.

Drawbacks

Here are the drawbacks of the AI mode switcher pattern:

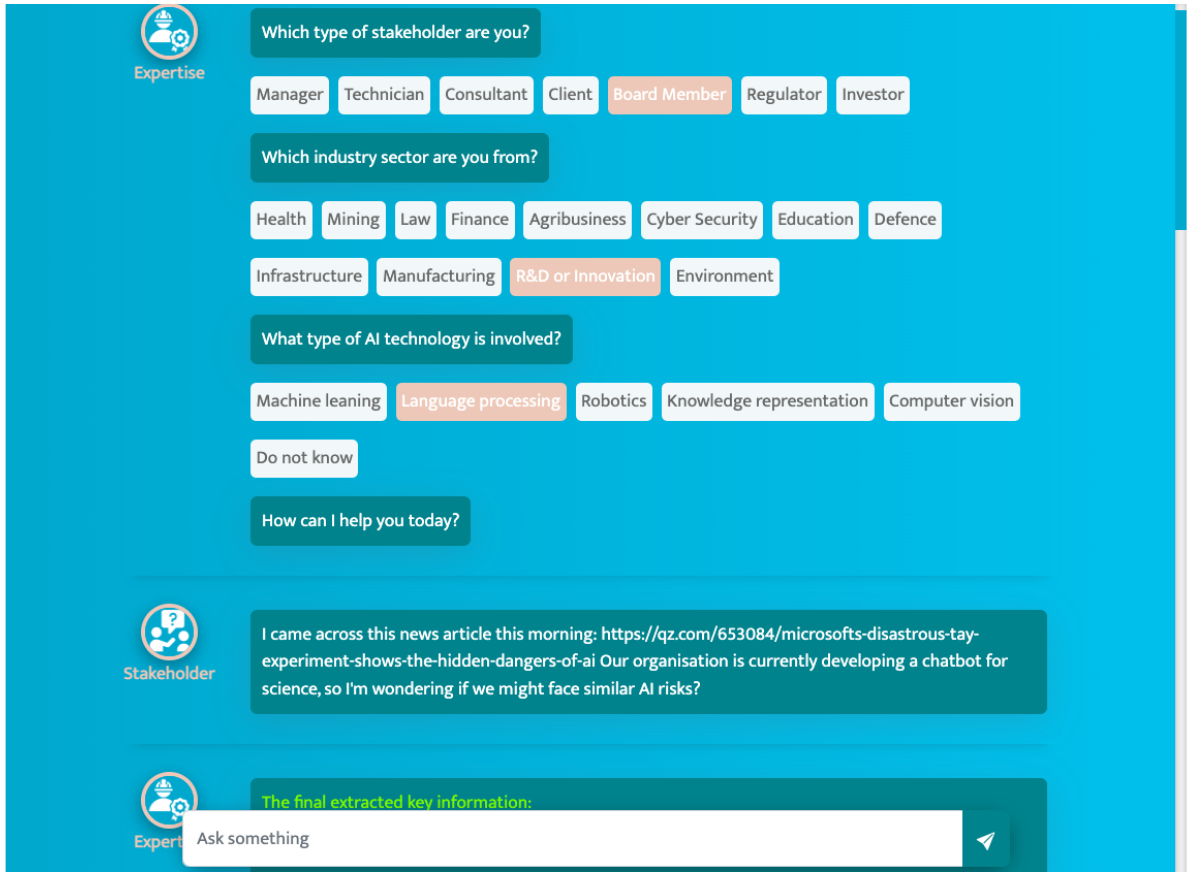
- **Efficiency:** Efficiency and performance of the decision-making points highly depend on the quality of other non-AI components involved.
- **Suitability to (near) real-time systems:** The use of an AI mode switcher in a (near) real-time system might be problematic. The performance of the system might be affected if the end user or the operator of the AI system keeps switching the AI component on and off.

Known Uses

Here are the known uses of the AI mode switcher pattern:

- Tesla Autopilot has multiple driver-assistance features that can be enabled or disabled during driving.[28] Users maintain control of the vehicles and can override the operation of these features at runtime.
- Waymo operates self-driving cars with an automated driving system that human safety drivers can override.[29]

Responsible AI Chatbot



You Are Using AS

Stakeholder

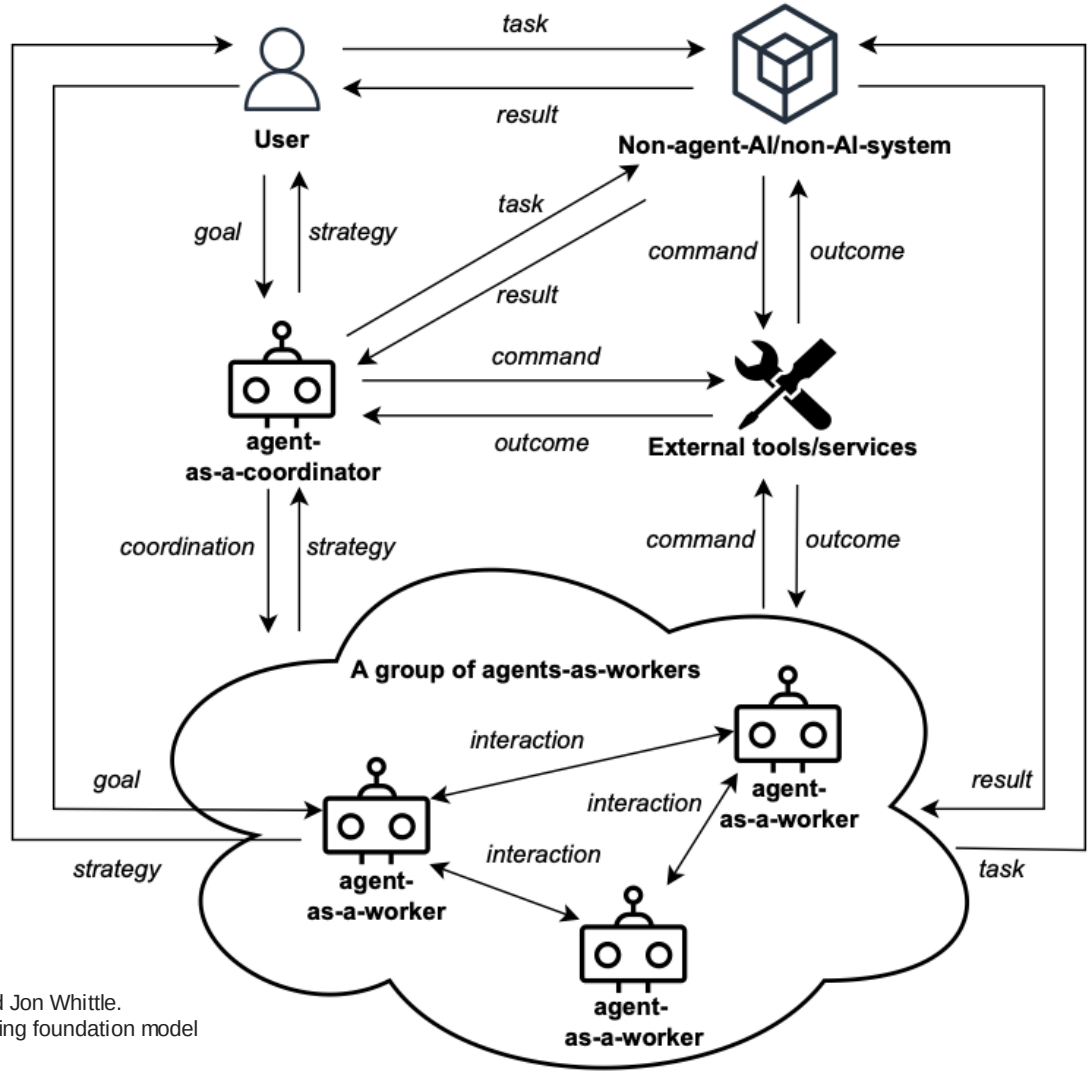
Console

```
AI: Now analyzing your input, please be patien t.

AI: Check whether there is hyperlink......

AI: Extracting hyperlin k......

AI: Accessing hyperlink ......

AI: Extracting Key Info rmation from the conten t......

AI: Summarizing your ce ntral ideas......

AI: Selecting incidents from database for you.. ....

AI: Generating explanat ions......
```

Unique Characteristics of Agents

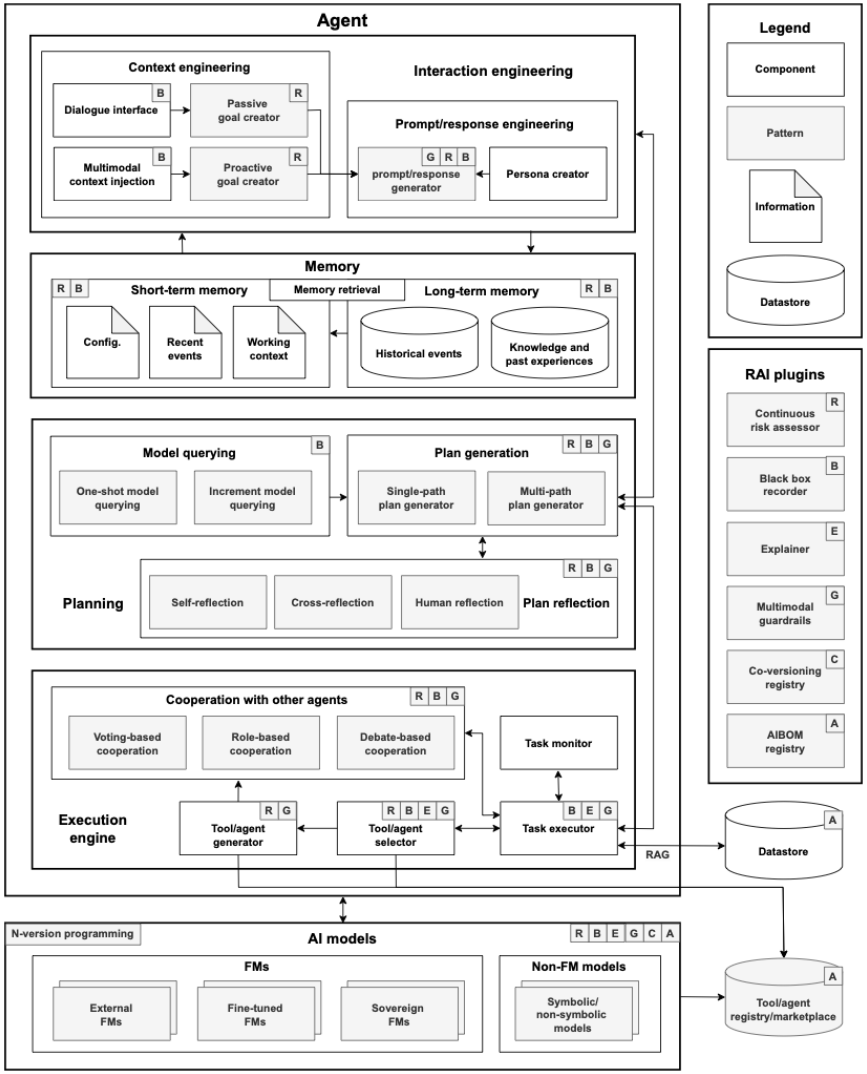
- Complex Architecture
- Autonomous Operation
- Non-Deterministic Behaviour
- Continuous Evolution

Architecture of an agent-based ecosystem



Qinghua Lu, Liming Zhu, Xiwei Xu, Zhenchang Xing, Stefan Harrer, and Jon Whittle. "Towards responsible generative ai: A reference architecture for designing foundation model based agents." ICSA'24. https://arxiv.org/pdf/2311.13148

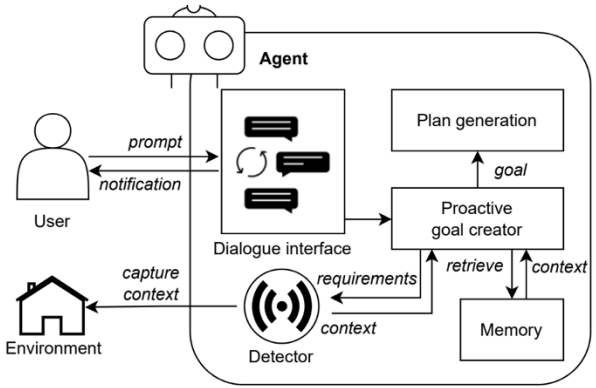
Reference architecture



Qinghua Lu, Liming Zhu, Xiwei Xu, Zhenchang Xing, Stefan Harrer, and Jon Whittle. "Towards responsible generative ai: A reference architecture for designing foundation model based agents." ICSA'24. https://arxiv.org/pdf/2311.13148

Agent Design Pattern Catalogue

Pattern: Proactive Goal Creator



Summary: Proactive goal creator anticipates users' goals by understanding human interactions and capturing the context via relevant tools.

Context: Users explain the goals that the agent is expected to achieve in the prompt.

Problem: The context information collected via solely a dialogue interface may be limited, and result in inaccurate responses to users' goals.

Forces:

- *Underspecification.* i) Users may not be able to provide thorough context information and specify precise goals to agents. ii) Agents can only retrieve limited information from the memory.
- *Accessibility.* Users with specified disabilities may not be able to directly interoperate with the agent via *passive goal creator*.

Solution: Fig. 4 illustrates a simple graphical representation of *proactive goal creator*. In addition to the prompts received from dialogue interface, and relevant context retrieved from memory, the *proactive goal creator* can anticipate users' goals by sending requirements to detectors, which will then capture and return the user's surroundings for further analysis and comprehension to generate the goals, for instance, identifying the user's gestures through cameras, recognising application UI layout via screenshots, etc. The *proactive goal creator* should notify users about context capturing and other relevant issues with a low false positive rate, to avoid unnecessary interruptions. In addition, the captured environment information can be stored in the agent's memory (or knowledge base) to establish "world models" [22, 23] to improve its ability to comprehend the real world.

Consequences:

Benefits:

- *Interactivity.* An agent can interact with users or other agents by anticipating their decisions proactively with captured multimodal context information.
- *Goal-seeking.* The multimodal input can provide more detailed information for the agent to understand users' goals, and increase the accuracy and completeness of goal achievement.
- *Accessibility.* Additional tools can help capture the sentiments and other context information from disabled users, ensuring accessibility and broadening the human values of foundation model-based agents.

Drawbacks:

- *Overhead.* i) *Proactive goal creator* is enabled by the multimodal context information captured by relevant tools, which may increase the cost of the agent. ii) Limited context information may increase the communication overhead between users and agents.

Known uses:

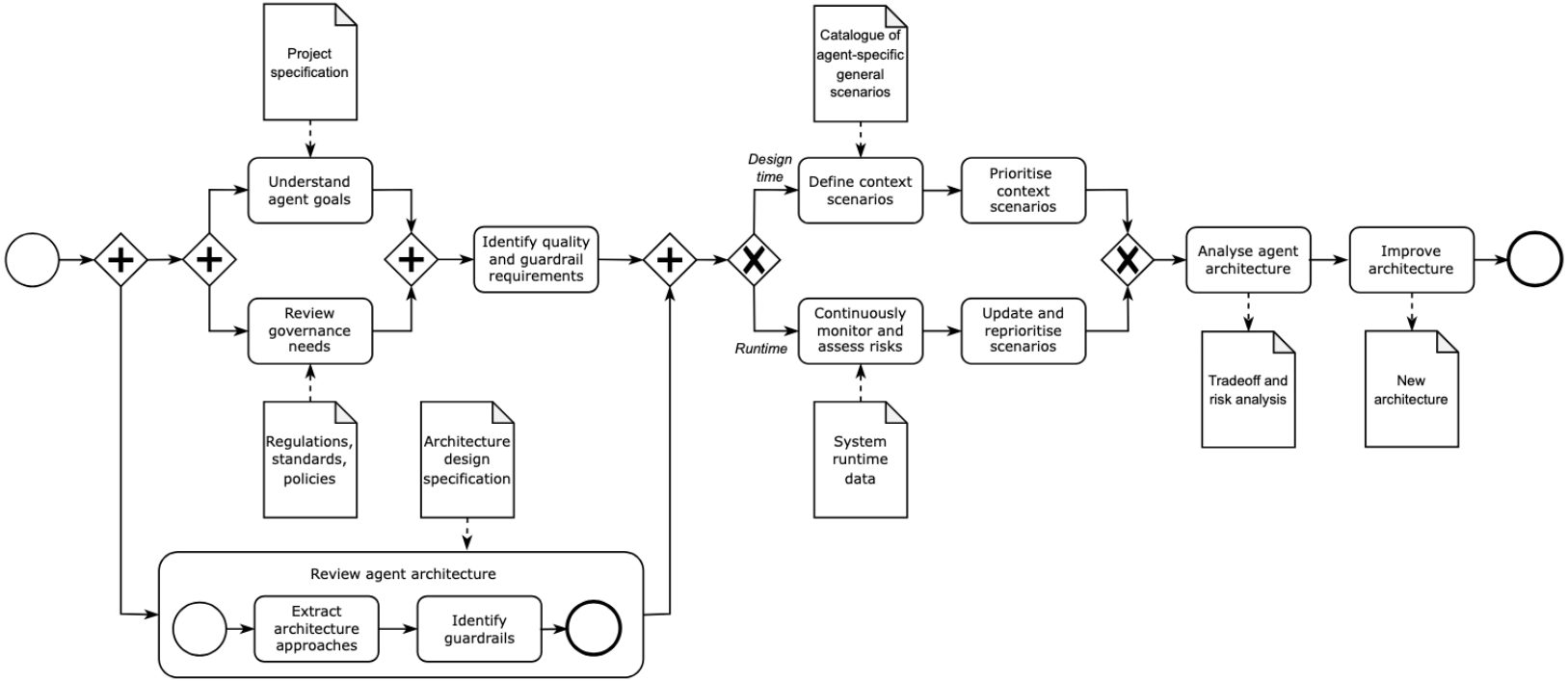
- *GestureGPT* [24]. GestureGPT can decipher users' hand gesture descriptions and hence comprehend users' intents.
- Zhao et al. [25] proposed a programming screencast analysis tool that can extract the coding steps and code snippets.
- *ProAgent* [26]. ProAgent can observe the behaviours of other teammate agents, deduce their intentions, and adjust the planning accordingly.

Related patterns:

- *Passive goal creator. Proactive goal creator* can be regarded an alternative of *passive goal creator* enabling multimodal context injection.
- *Prompt/response optimiser*.
- *Proactive goal creator* can first handle users' inputs and transfer the goals and relevant context information to *prompt/response optimiser* for prompt refinement.

Yue Liu, Sin Kit Lo, Qinghua Lu, Liming Zhu, Dehai Zhao, Xiwei Xu, Stefan Harrer, and Jon Whittle. "Agent Design Pattern Catalogue: A Collection of Architectural Patterns for Foundation Model based Agents." *Journal of Systems and Software* (2024). https://arxiv.org/abs/2405.10467

AgentArcEval: An Architecture Evaluation Method for Foundation Model based Agents



Qinghua Lu, Dehai Zhao, Yue Liu, Hao Zhang, Liming Zhu, Xiwei Xu, Angela Shi, and Tristan Tan. "AgentArcEval: An Architecture Evaluation Method for Foundation Model based Agents." (2024).
https://www.researchgate.net/publication/385660422_Evaluating_the_architecture_of_large_language_model-based_agents

Catalogue of Agent-Specific General Scenarios

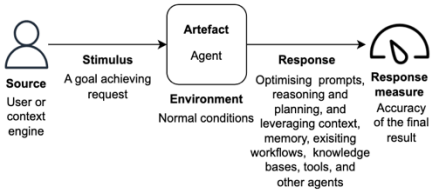


Fig. 2: Accuracy general scenario.

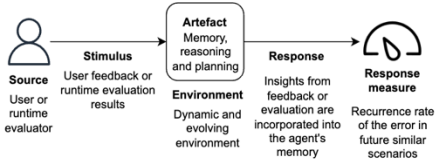


Fig. 3: Adaptability general scenario.

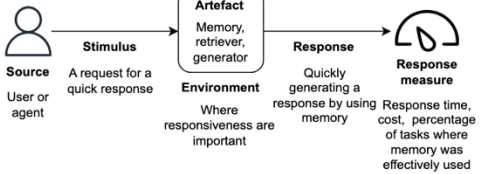


Fig. 4: Efficiency general scenario.

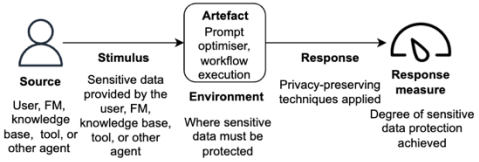


Fig. 5: Privacy general scenario.

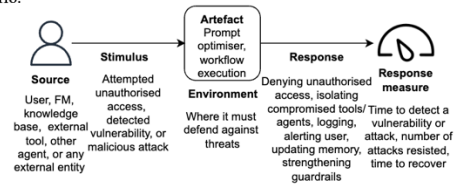


Fig. 6: Security general scenario.

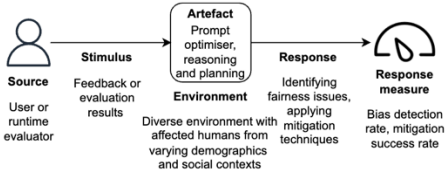


Fig. 7: Fairness general scenario.

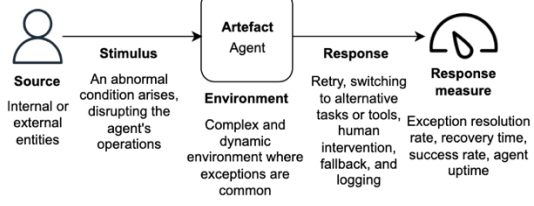


Fig. 8: Reliability general scenario.

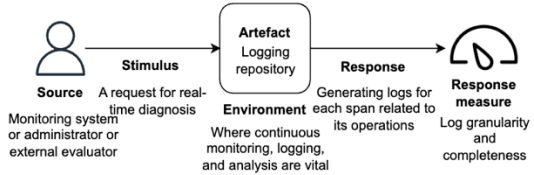


Fig. 9: Observability general scenario.

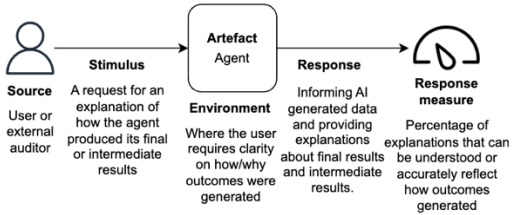
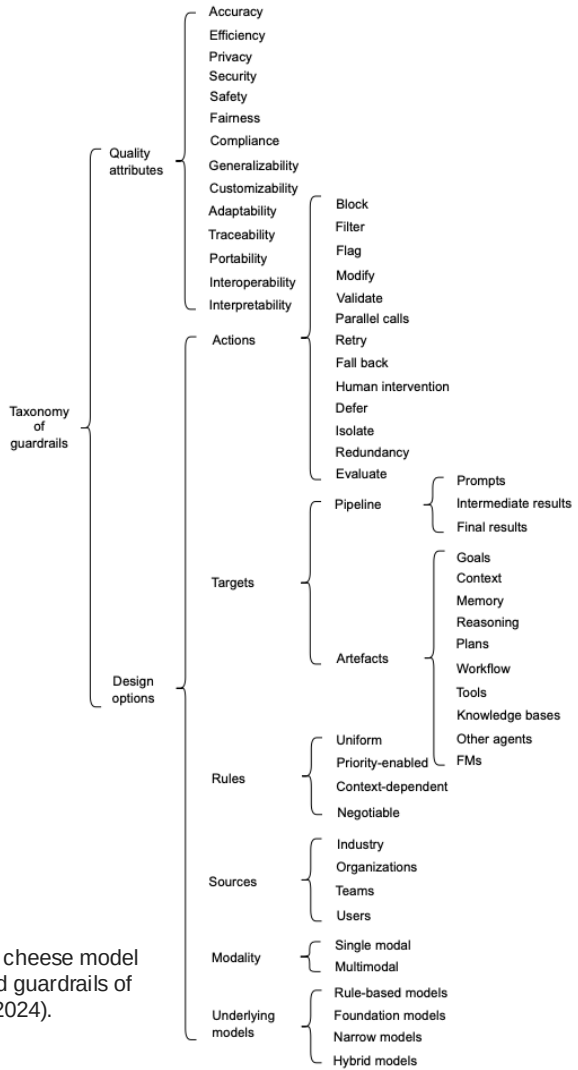


Fig. 10: Transparency general scenario.

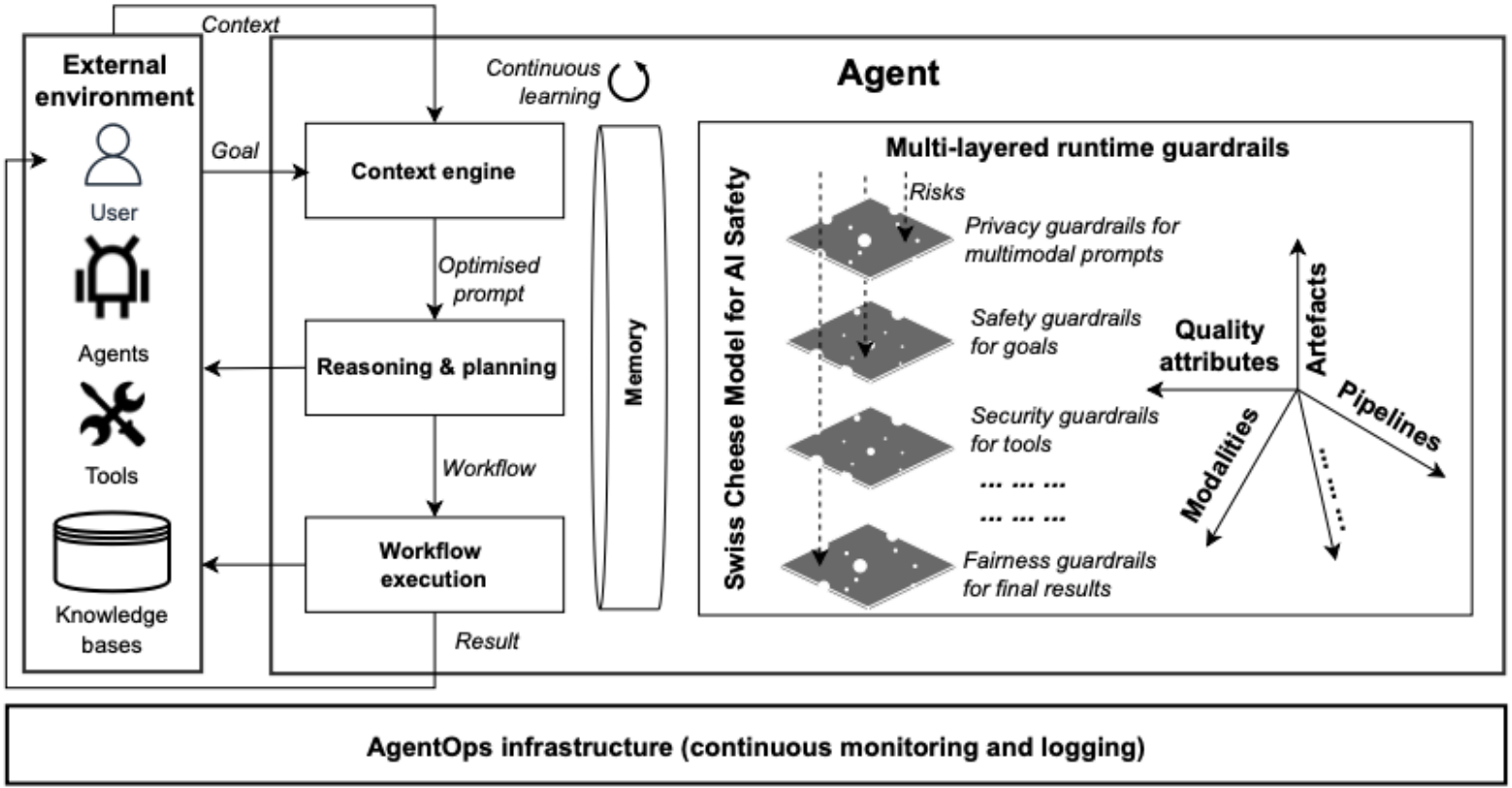
Qinghua Lu, Dehai Zhao, Yue Liu, Hao Zhang, Liming Zhu, Xiwei Xu, Angela Shi, and Tristan Tan. "AgentArcEval: An Architecture Evaluation Method for Foundation Model based Agents." (2024). https://www.researchgate.net/publication/385660422_Evaluating_the_architecture_of_large_language_model-based_agents

Design Taxonomy of Runtime Guardrails for LLM Agents

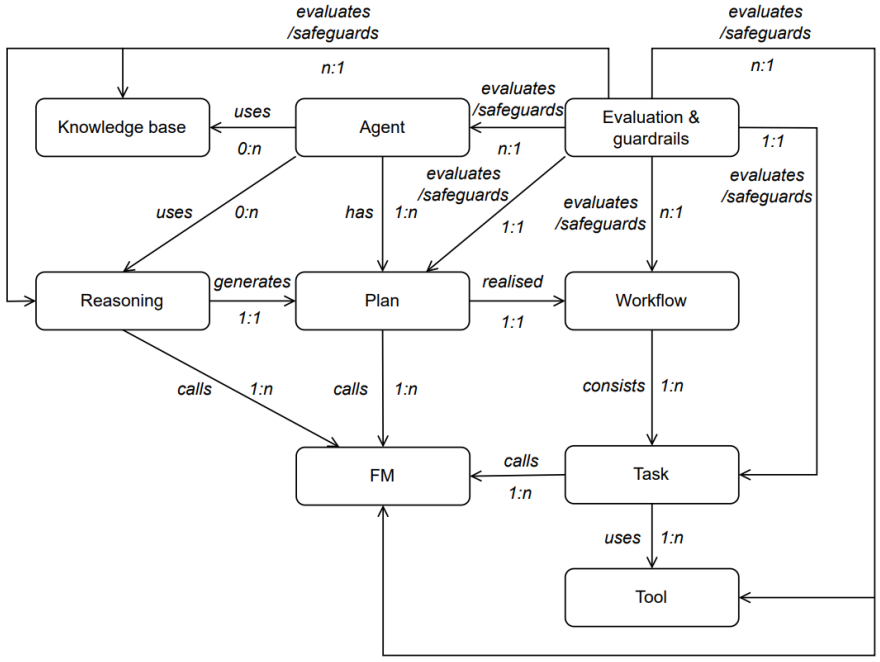


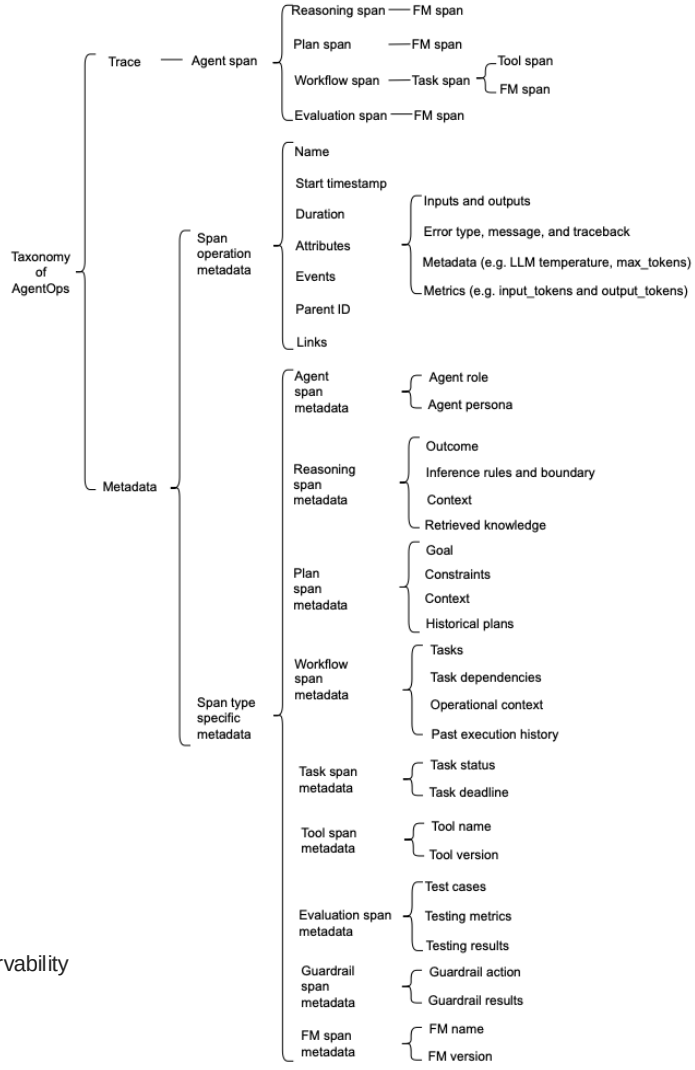
Md Shamsujjoha, Qinghua Lu, Dehai Zhao, and Liming Zhu. "Swiss cheese model for ai safety: A taxonomy and reference architecture for multi-layered guardrails of foundation model-based agents." *arXiv preprint arXiv:2408.02205* (2024). https://arxiv.org/abs/2408.02205

Swiss Cheese Model for AI Safety – Multi-Layered Guardrails for LLM Agents



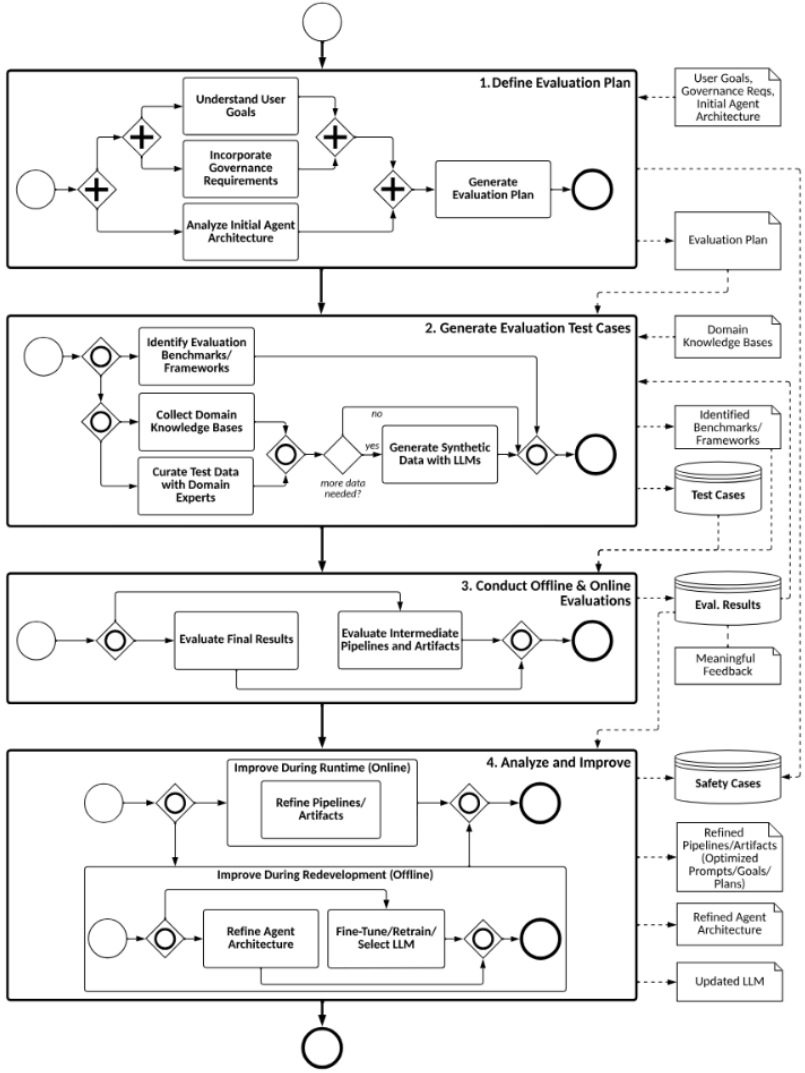
Md Shamsujjoha, Qinghua Lu, Dehai Zhao, and Liming Zhu. "Swiss cheese model for ai safety: A taxonomy and reference architecture for multi-layered guardrails of foundation model-based agents." *arXiv preprint arXiv:2408.02205* (2024). https://arxiv.org/abs/2408.02205





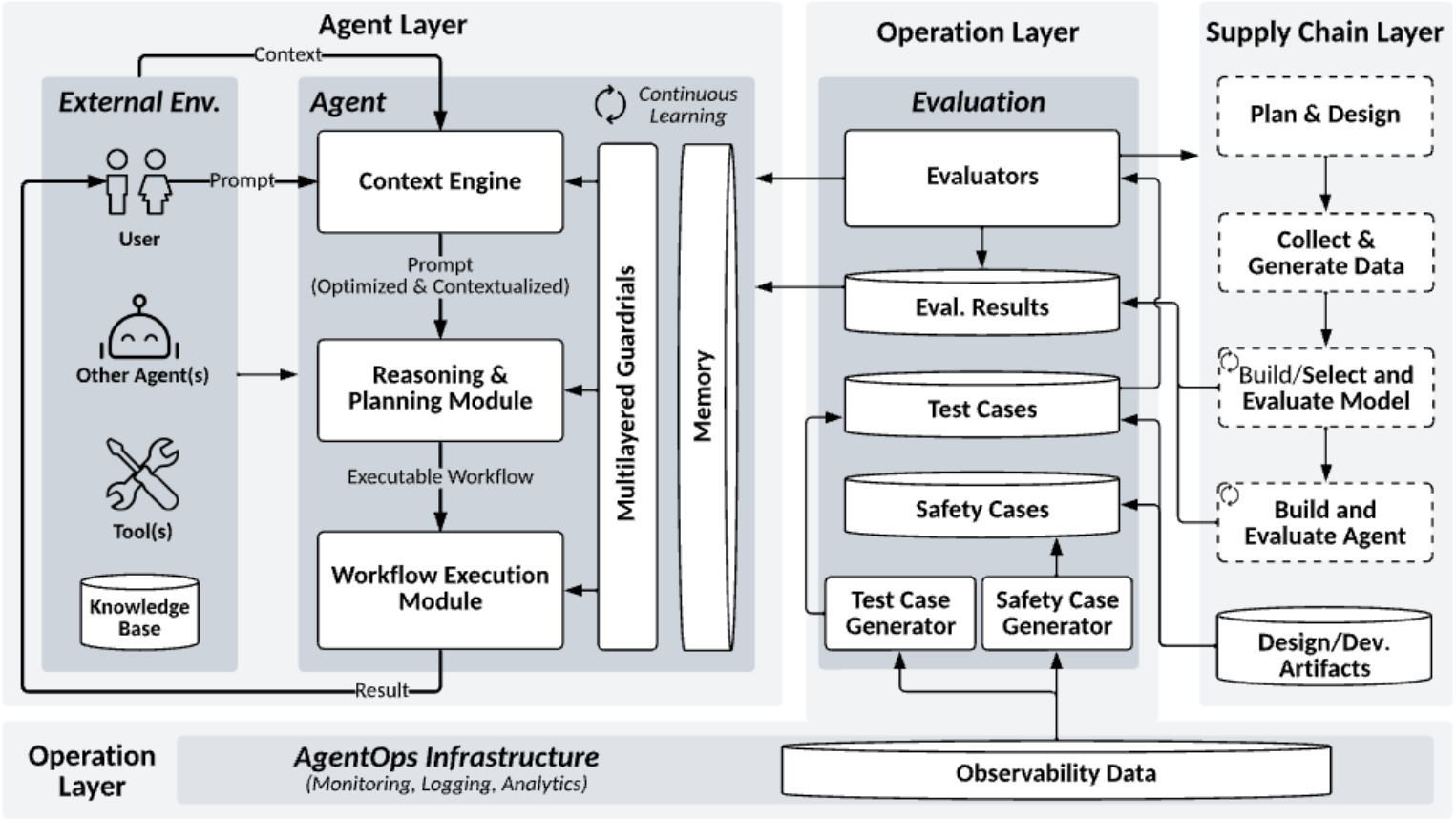
Liming Dong, Qinghua Lu, and Liming Zhu. "A Taxonomy of AgentOps for Enabling Observability of Foundation Model based Agents." *arXiv preprint arXiv:2411.05285* (2024).
https://arxiv.org/abs/2411.05285

Process Model for LLM Agent Evaluation



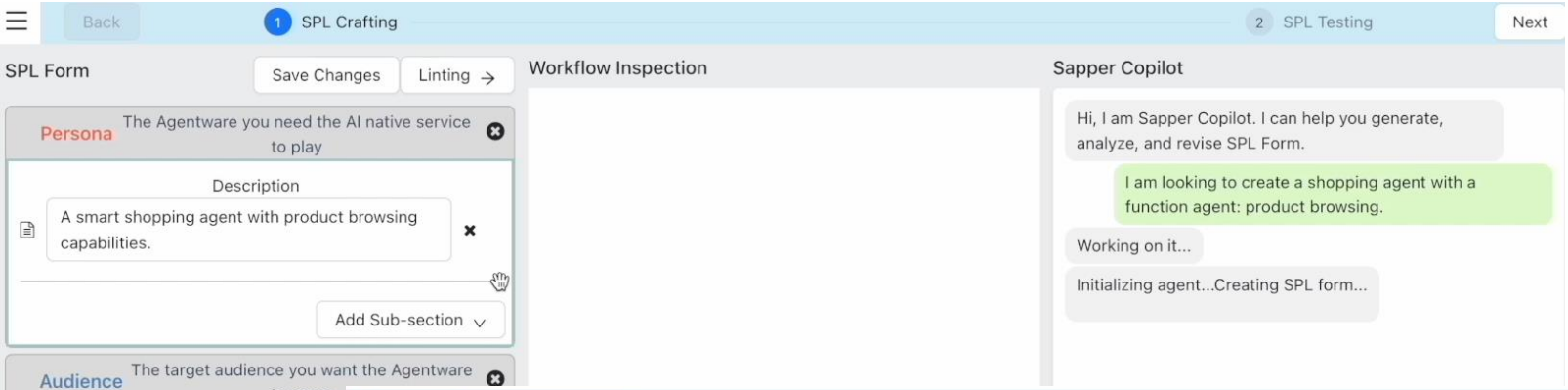
Xia, Boming, Qinghua Lu, Liming Zhu, Zhenchang Xing, Dehai Zhao, and Hao Zhang. "Evaluation-Driven Design of LLM Agents: A Process Model and Reference Architecture." *arXiv preprint arXiv:2411.13768* (2024). https://arxiv.org/abs/2411.13768

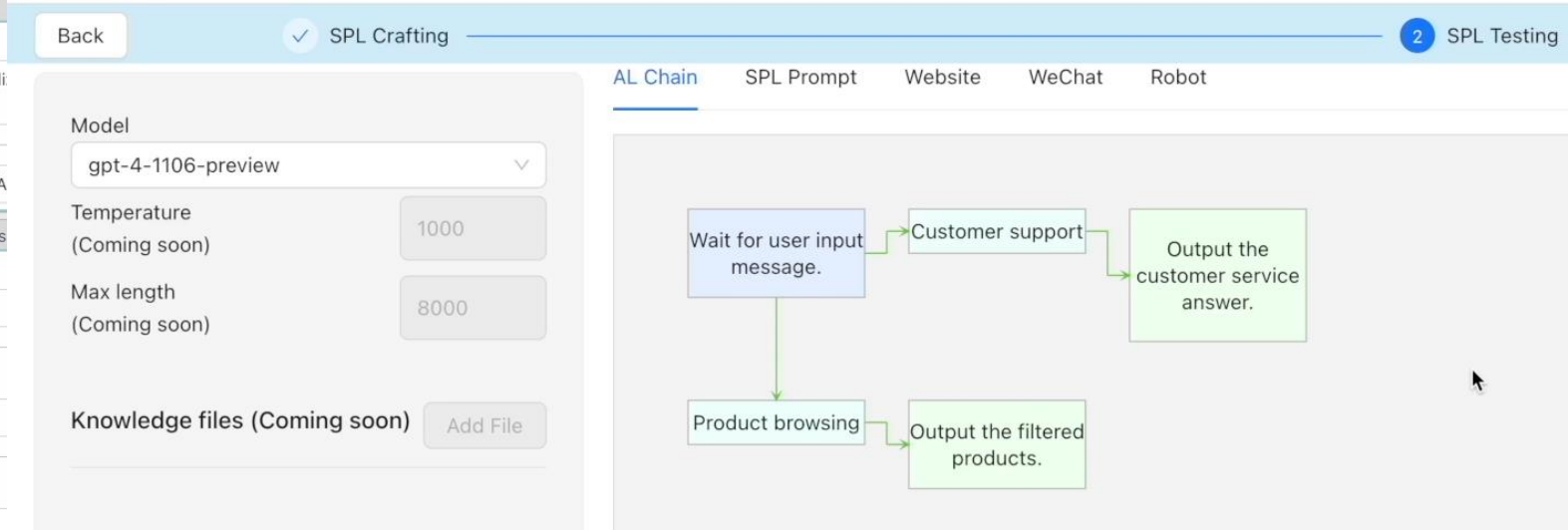
Evaluation-Driven Design for LLM Agents



Xia, Boming, Qinghua Lu, Liming Zhu, Zhenchang Xing, Dehai Zhao, and Hao Zhang. "Evaluation-Driven Design of LLM Agents: A Process Model and Reference Architecture." *arXiv preprint arXiv:2411.13768* (2024). https://arxiv.org/abs/2411.13768

Science Digital: Agent Platform





Responsible/Safe AIware Engineering at system-level and across supply chain

Thank you.

Qinghua Lu

Responsible AI Science Team Leader

qinghua.lu@data61.csiro.au

https://research.csiro.au/ss/team/se4ai/responsible-ai-engineering/