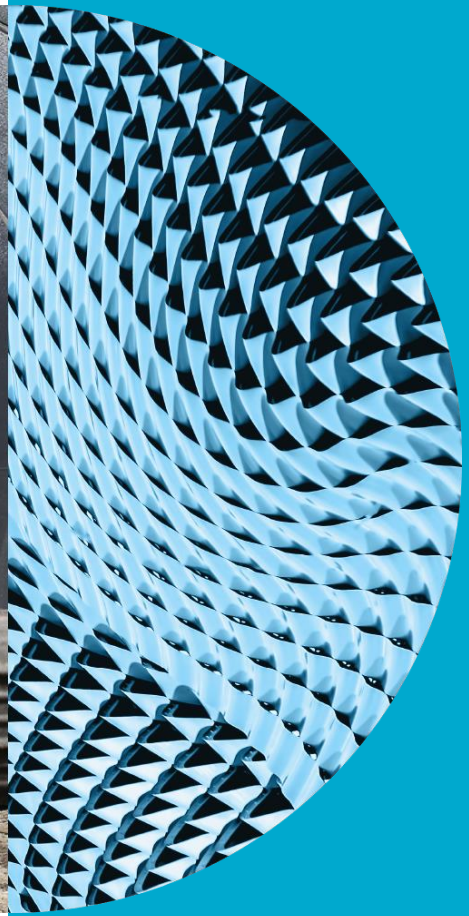


Software Systems Research Group

Qinghua Lu

Group Leader

Software Systems Research Group

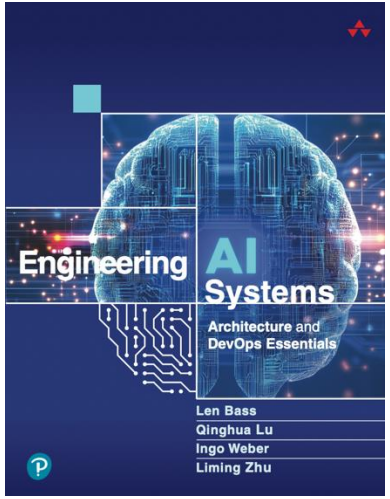
Data61, CSIRO

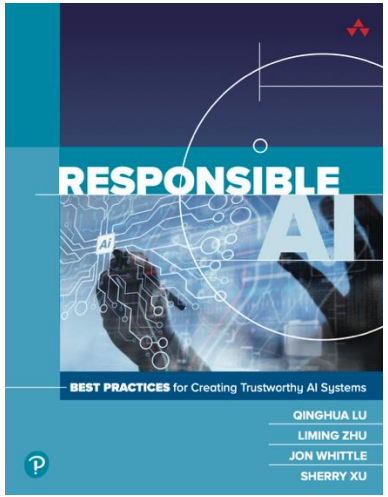
qinghua.lu@data61.csiro.au

https://research.csiro.au/ss/

Group Overview

- ~40 full time research scientists/engineers
- 5 Research Teams
  - Software Engineering for AI Research Team
  - Applied AI Systems Research Team
  - Trustworthy Processes Research Team
  - Architecture and Analytics Platform Research Team
  - AI Diversity and Inclusion Research Team
- 6 scientists in the global top 20 for Responsible AI





Global Leadership in Responsible AI & AI Engineering

- Contributing to International Working Groups
  - Frontier Model Forum
    - 20+ leading orgs including OpenAI, Google, Microsoft, Anthropic, Meta, AI Safety Institutes
  - International Network of AI Safety Institutes
    - Ongoing collaboration with US, UK, Canada, Japan, and Singapore AISI
  - AI Metrology
    - With Alan Turing Institute, MIT, Mila, etc
  - OECD.AI
  - EU General-Purpose AI Code of Practice
- Organising/Editorial Roles in Academic Conferences/Journals:
  - International Conferences on AI Engineering, AI Powered Software (AIware)
  - International Workshops on Responsible AI Engineering, Agentic Engineering
  - IEEE Transactions on AI, Engineering Applications of AI

Key Responsible AI and AI Engineering Projects

- Government Projects
  - DISR
    - Australia's AI Safety Standards (v1 and v2)
  - South Australian Gov and AIML
    - Responsible AI Research Centre
  - Responsible AI with Australian Sports Commission
- Industry Projects
  - ESG-AI project with Alphinity Investment
  - Tax copilot with Empathetic AI
  - Westpac/Cognitivo on RAG engineering

Responsible AI Methods/Frameworks

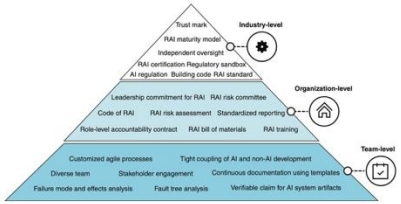
Responsible AI Question Bank

- 300+ questions
- 100+ metrics for concrete measurement of AI risks

https://arxiv.org/pdf/2305.09300.pdf
https://arxiv.org/abs/2311.13158

Responsible AI Pattern Catalogue

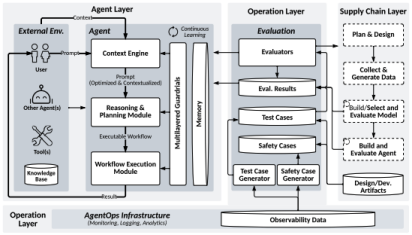
- 60+ reusable best practices for different stakeholders
- Governance, process, product



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

AI System Safety Evaluation

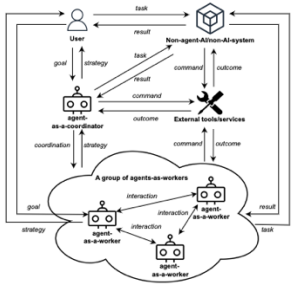
- Evaluation-driven learning
- Agent architecture evaluation
- AISI joint testing



https://arxiv.org/abs/2411.13768
https://arxiv.org/abs/2404.05388

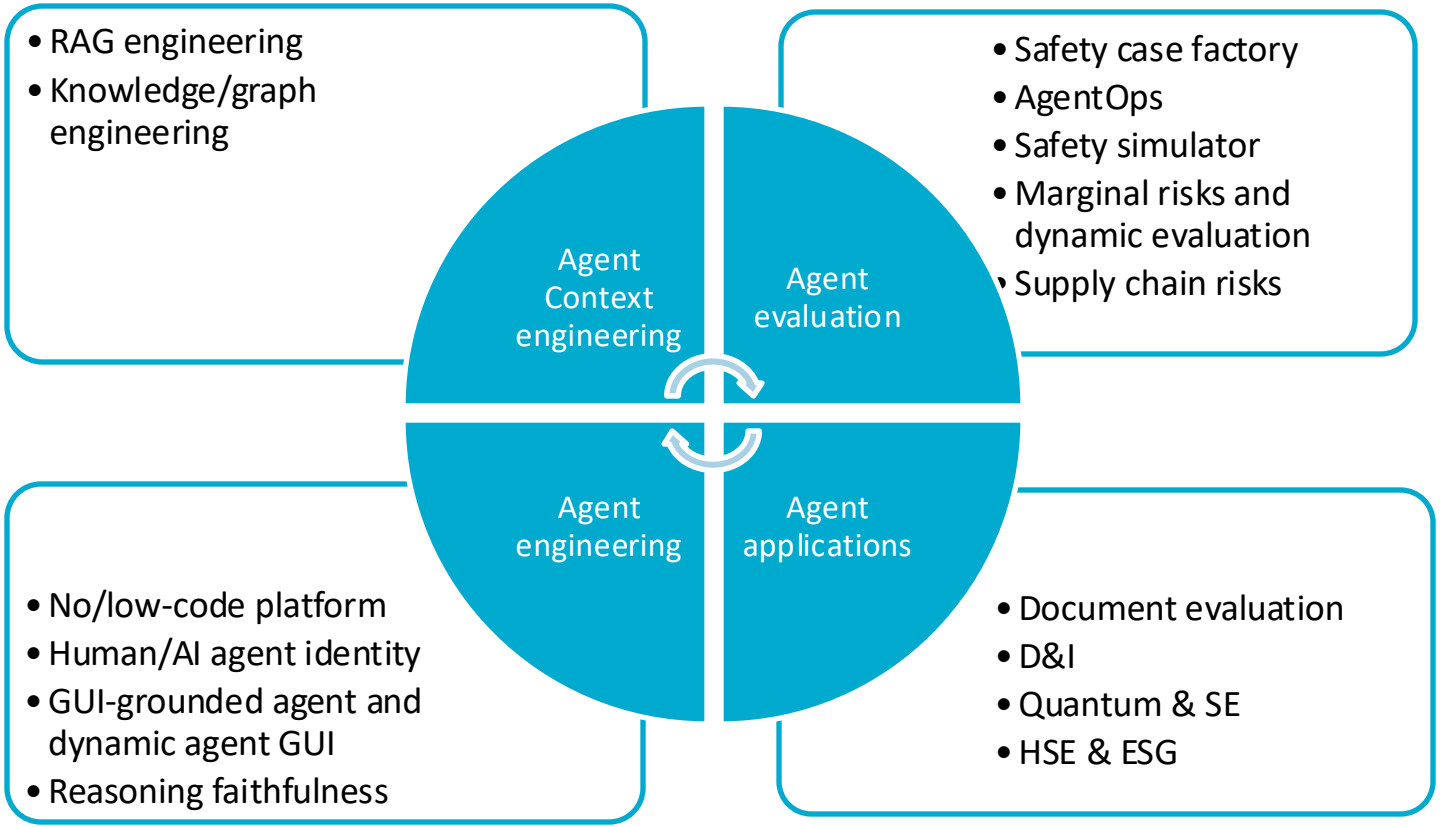
AI-Safety-by-Design Architecture

- Agent reference architecture
- Agent design pattern catalogue
- Swiss Cheese Model for multi-layered guardrails



https://arxiv.org/abs/2311.13148
https://arxiv.org/abs/2405.10467
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5266496

Agent Engineering



Thank you.

Qinghua Lu

Group Leader

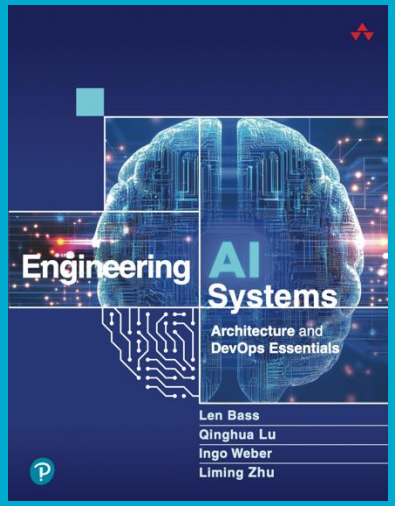
Software Systems Research Group

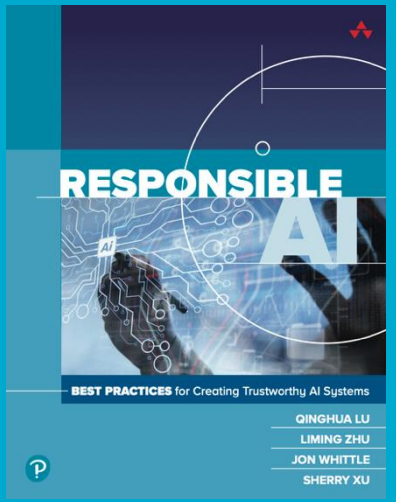
Data61, CSIRO

qinghua.lu@data61.csiro.au

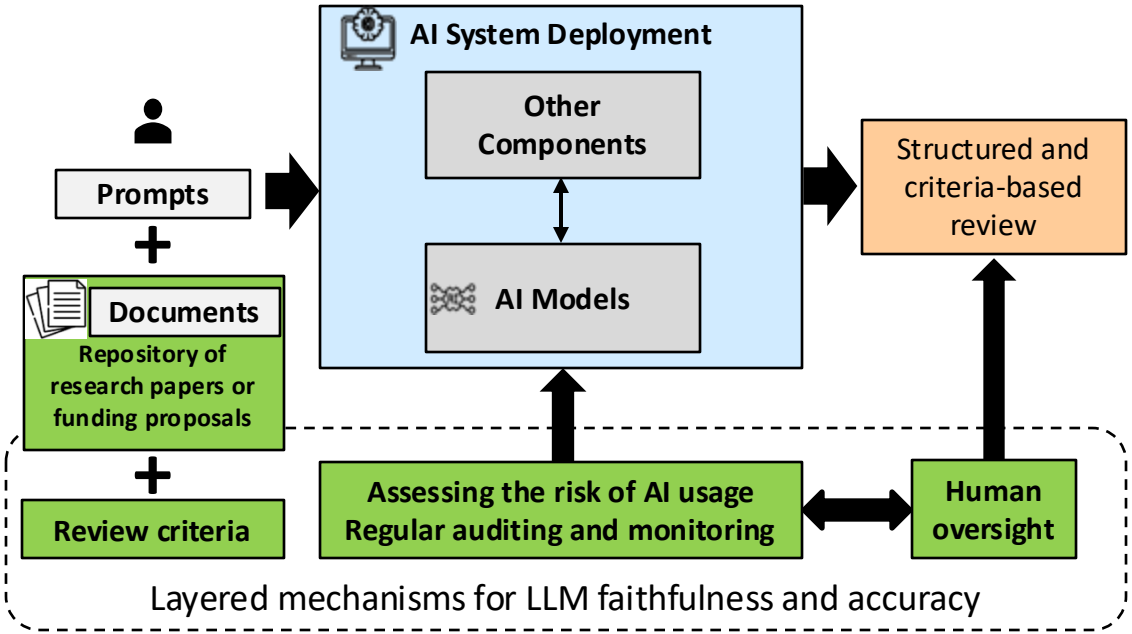
https://research.csiro.au/ss/

Australia's National Science Agency





AI for Document Evaluation



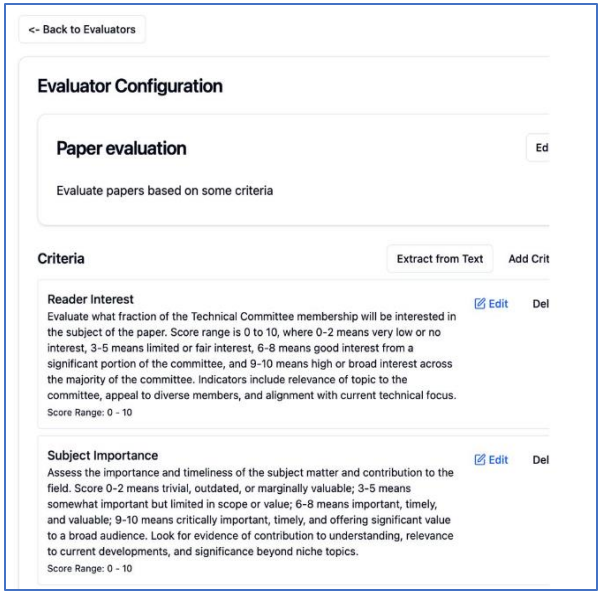
Value Proposition

1. An AI tool to **review large volumes of documents**—such as research papers or funding proposals—against specific rubrics or policy criteria.
2. **Human oversight** to ensure relevance, quality, and trust in generated reviews

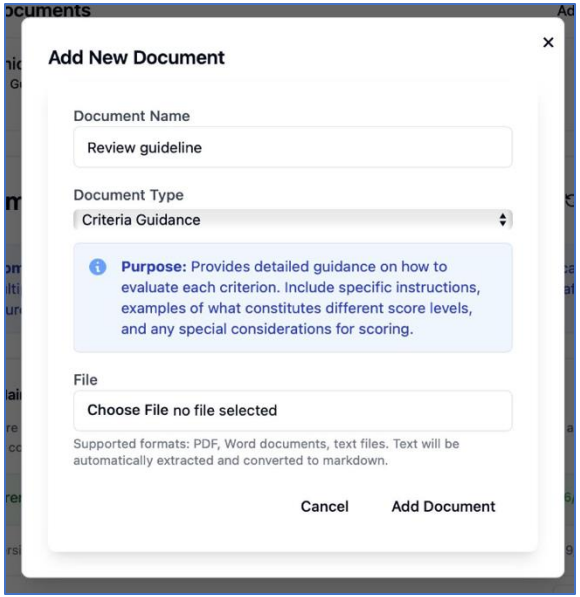
Technical Elements

1. Use large language models like GPT and Claude
2. Implement different levels of mechanisms to ensure the LLM faithfulness and accuracy

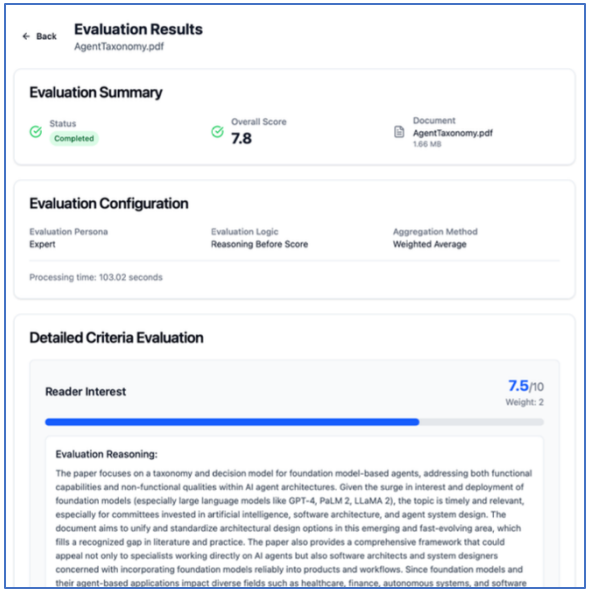
Demo



Customizable Criteria Based Evaluation



Multi-Source Document Support



Transparent Evaluation with Justification