ENGAGESMARTER
Making financial services more engaging

Assuring and evaluating AI Agents for financial services

Tom Dorrington Ward
CTO & Co-Founder, Engage Smarter AI

LSE Statistics Research Showcase
20 June 2024

INTRODUCTION

Generative AI is engaging, but can be unreliable

ChatGPT started a new wave of “generative AI” applications

- We see huge potential for generative AI to help people make smarter financial decisions – because it is accessible and engaging

Using conversational AI “Agents” for financial guidance requires assurance

- Risks with providing incorrect guidance
- Advice is a regulated activity

In this talk, we’ll survey some of the elements we think about when building AI Agents for financial services – particularly around assuring and evaluating them

Our Mission is to close the Financial Guidance gap

Around **26 Million UK consumers** underserved by FS firms

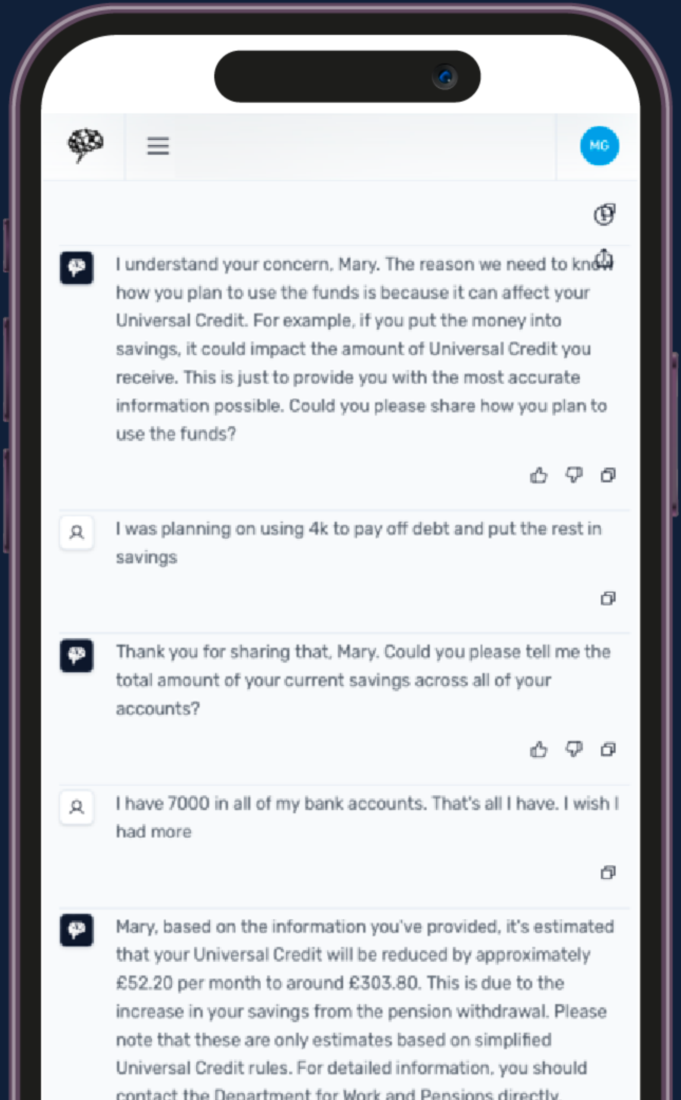
Financial Advice is too expensive for consumers

Human experts are too expensive for firms

Our Product

For customers

- ✓ Automated financial guidance
- ✓ Technical expert on financial products
- ✓ Less friction - Available 24/7 and expert help



I understand your concern, Mary. The reason we need to know how you plan to use the funds is because it can affect your Universal Credit. For example, if you put the money into savings, it could impact the amount of Universal Credit you receive. This is just to provide you with the most accurate information possible. Could you please share how you plan to use the funds?

I was planning on using 4k to pay off debt and put the rest in savings

Thank you for sharing that, Mary. Could you please tell me the total amount of your current savings across all of your accounts?

I have 7000 in all of my bank accounts. That's all I have. I wish I had more

Mary, based on the information you've provided, it's estimated that your Universal Credit will be reduced by approximately £52.20 per month to around £303.80. This is due to the increase in your savings from the pension withdrawal. Please note that these are only estimates based on simplified Universal Credit rules. For detailed information, you should contact the Department for Work and Pensions directly

For Financial Services firms

- ✓ 80-90% lower cost
- ✓ Pre-built agents with industry content
- ✓ Fulfill regulation Consumer Duty
- ✓ Significantly more accurate than human staff

AI agents used in Financial Services need to be both explainable and accurate

Accuracy	
98%	Our Pensions Expert AI agent
~80%	Human agents (actual answers)
~60%	GPT4 (OpenAI)

Expert moderated test result based on 400 real world pension questions

Our Roadmap

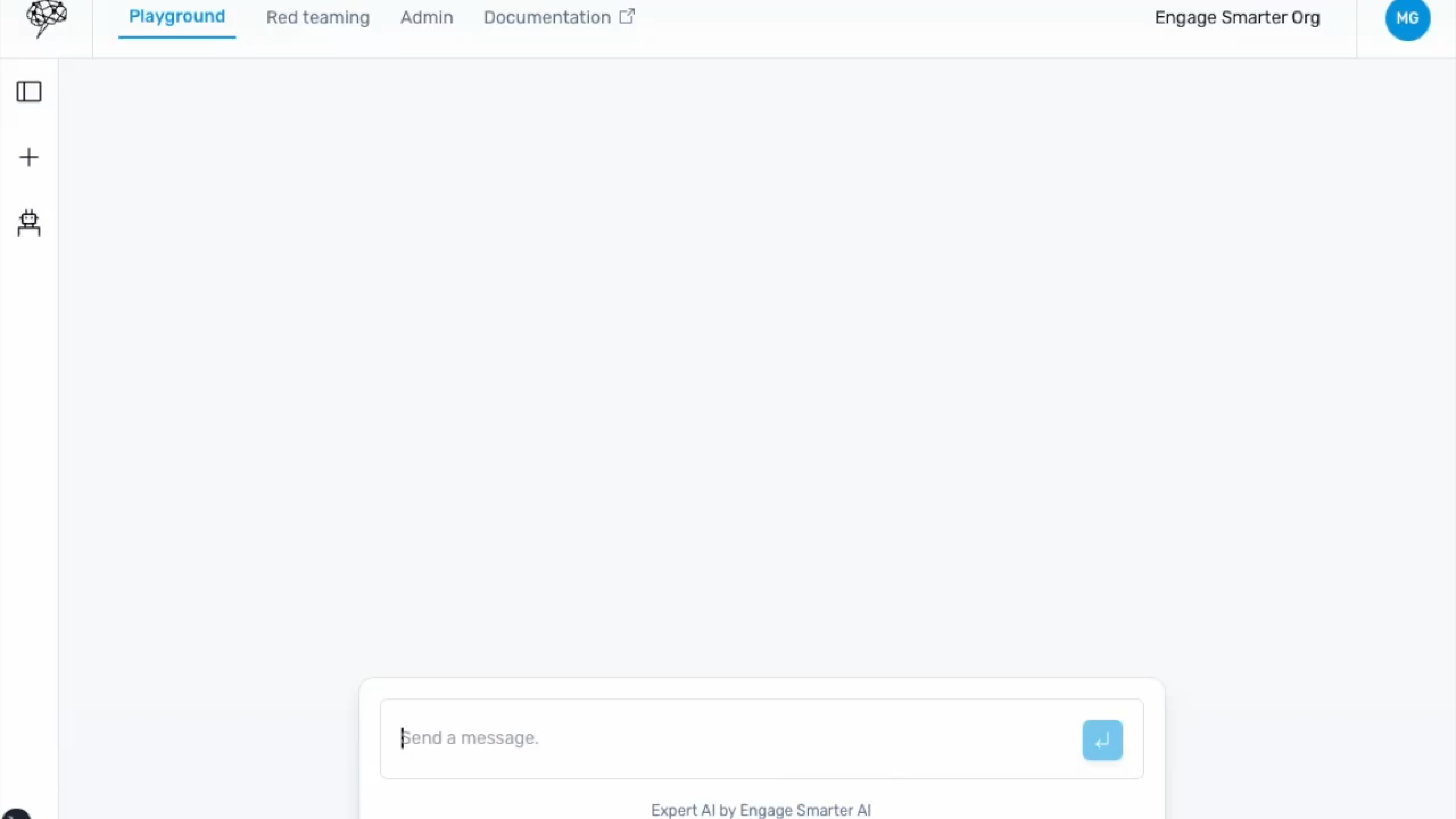
	Pensions	ISAs & GIAs	Investment	Debt	
Automated Expert Service Agents	✓	✓			
Technical Copilots (for staff)	✓				
Financial Guidance Agents	✓✓				
Commercially focused Guidance Agents	✓				

In beta and testing

Examples

“Can I pay some of my bonus into my pension? How much?”

“I will help you understand the benefits and implications of transferring your pension pot to us.”

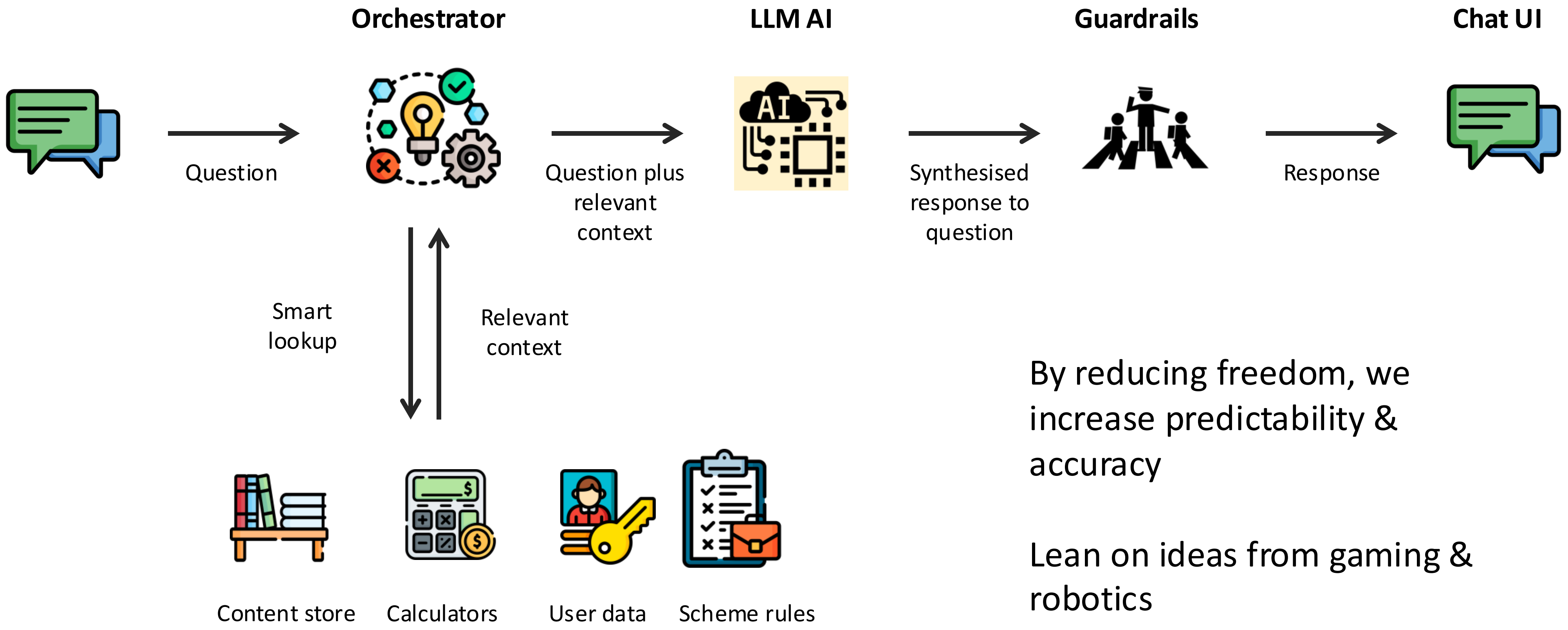

Playground
Red teaming
Admin
Documentation
Engage Smarter Org
MG
Send a message.
Expert AI by Engage Smarter AI

Let’s look at some dimensions for assuring these AI Agent systems

- ✓ Content, calculators and compound AI systems
- ✓ Guardrails
- ✓ Expert review and red teaming
- ✓ Iteration process
- ✓ Monitoring and metrics

DIMENSION 1

Content, calculators & compound AI systems



Orchestrator

LLM AI

Guardrails

Chat UI

Question

Question plus relevant context

Synthesised response to question

Response

Smart lookup

Relevant context

Content store

Calculators

User data

Scheme rules

By reducing freedom, we increase predictability & accuracy

Lean on ideas from gaming & robotics

DIMENSION 2

Guardrails are used for control flow

Guardrail examples

Scope

* Is conversation within our defined topics?

* Is the response inconsistent with the ground-truth knowledge base?

Content

* Is the content correct vs another source?

* Is the logic correct via a second check?

* Are any facts assumed in this response?

Context

* Is this response contradictory to the wider context of the conversation?

* Is there relevant background for this customer or conversation that would be expected to be known?

Tone

* Is the language in the right tone of voice for this service / client?

* legal / advice language check

* Any red flags on the customer questions? e.g. complaint or frustration?

Guardrails act upon different stages within the language engine pipeline to check and steer in real-time

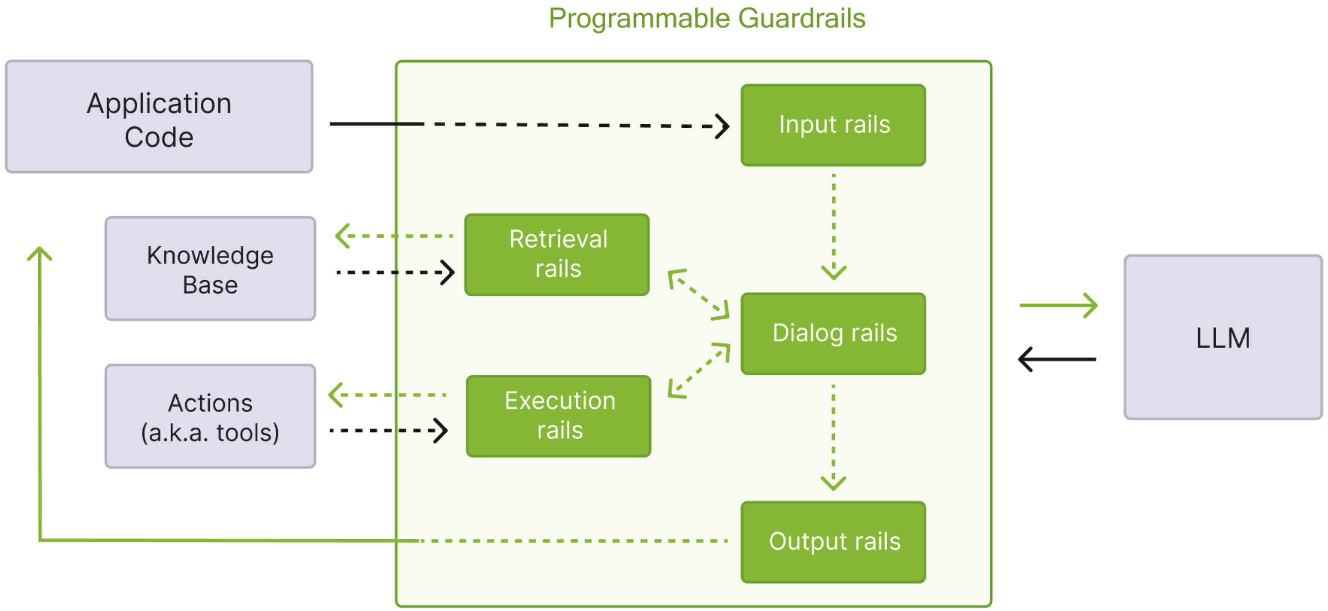
DIMENSION 2

Lots of possible guardrails

Lots of open-source packages to start from

Couple of examples

- NeMo by NVIDIA
- LLM-Guard - Library of different guardrails e.g. InvisibleText, BanSubstrings, PromptInjection



High-level flow through programmable guardrails.

DIMENSION 3

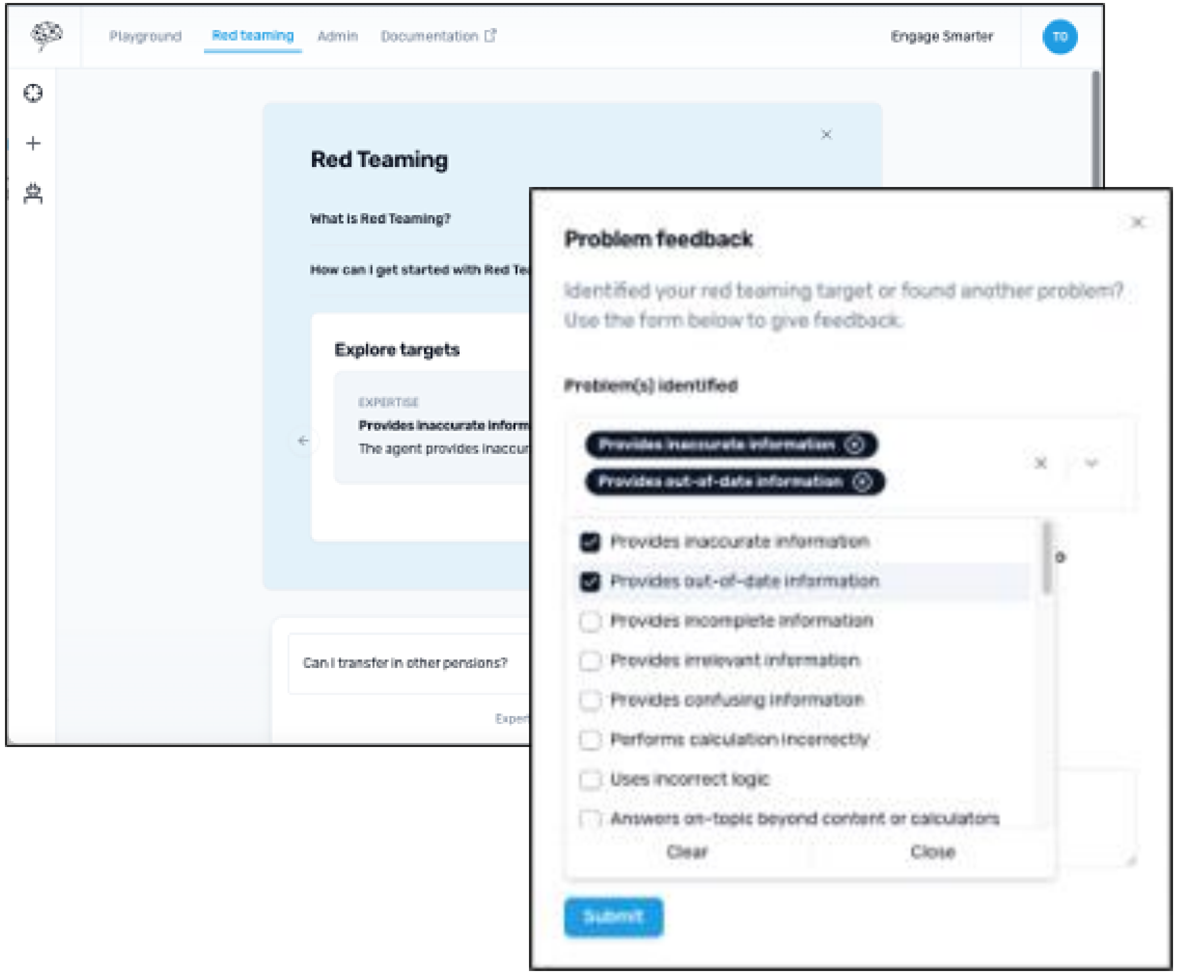
Expert review and red teaming

Expert review – Domain expert provides feedback on outputs

- End-to-end and module level (e.g. just a guardrail)

Red teaming – Trying to break the agent, so that its flaws can be found before an adversary finds them

- Useful for building internal support within client orgs



Red Teaming

What is Red Teaming?

How can I get started with Red Te

Explore targets

EXPERTISE

Provides inaccurate inform

The agent provides inaccur

Can I transfer in other pensions?

Exper

Problem feedback

Identified your red teaming target or found another problem? Use the form below to give feedback.

Problem(s) identified

Provides inaccurate information

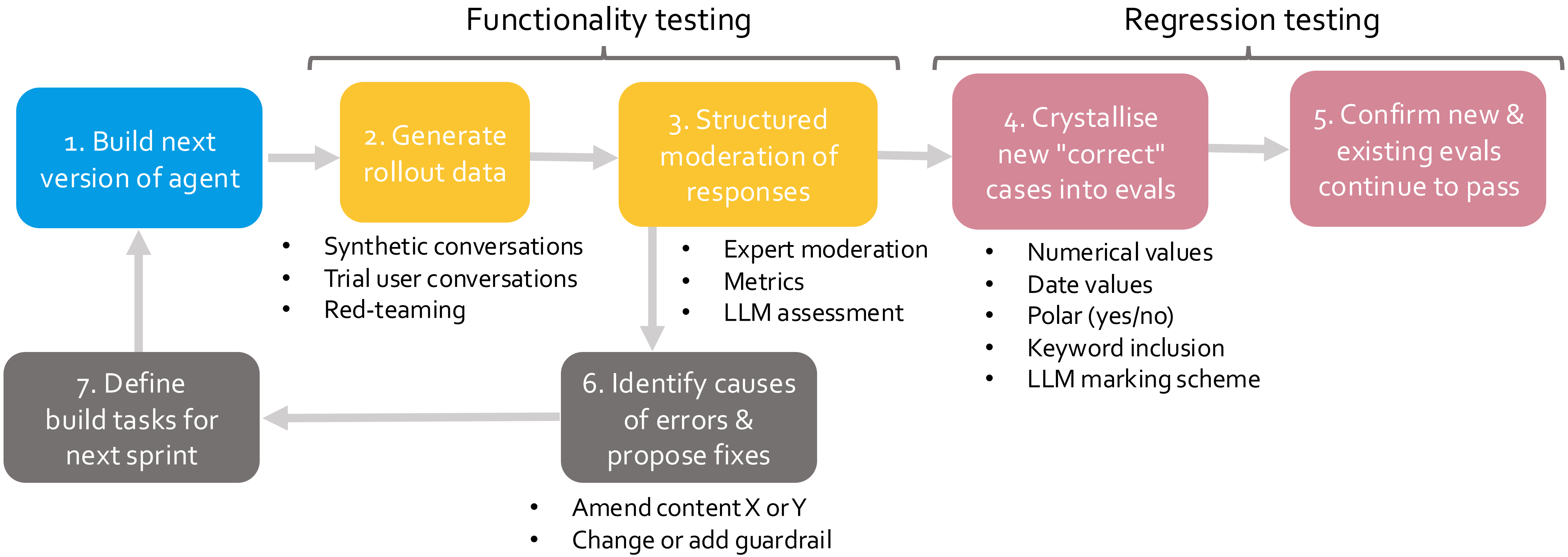
Provides out-of-date information

☑ Provides inaccurate information
☑ Provides out-of-date information
☐ Provides incomplete information
☐ Provides irrelevant information
☐ Provides confusing information
☐ Performs calculation incorrectly
☐ Uses incorrect logic
☐ Answers on-topic beyond content or calculators

DIMENSION 4

Iteration process improve safety

Each sprint is highly test-driven – testing new functionalities and ensuring that existing desired functionalities remain



Functionality testing

Regression testing

1. Build next version of agent

2. Generate rollout data
- Synthetic conversations
- Trial user conversations
- Red-teaming

3. Structured moderation of responses
- Expert moderation
- Metrics
- LLM assessment

4. Crystallise new "correct" cases into evals
- Numerical values
- Date values
- Polar (yes/no)
- Keyword inclusion
- LLM marking scheme

5. Confirm new & existing evals continue to pass

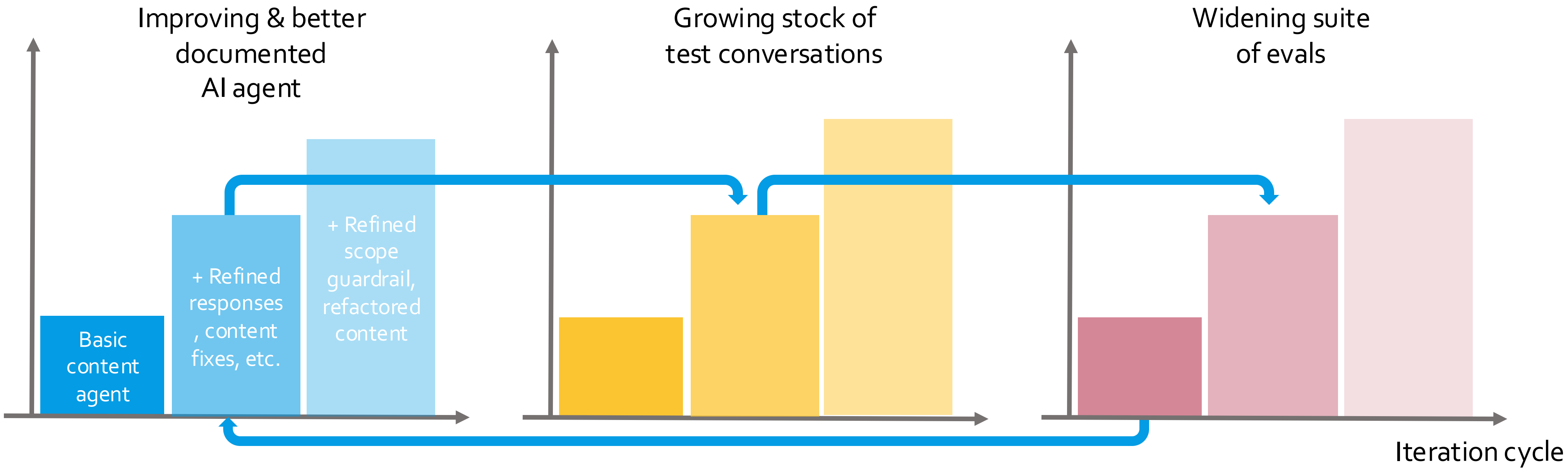
6. Identify causes of errors & propose fixes
- Amend content X or Y
- Change or add guardrail

7. Define build tasks for next sprint

DIMENSION 4

Improvement with each iteration

With each cycle, the AI Agent's behaviour is further refined, more extensively explored, and further pinned down


Improving & better documented AI agent
Basic content agent
+ Refined responses , content fixes, etc.
+ Refined scope guardrail, refactored content
Growing stock of test conversations
Widening suite of evals
Iteration cycle

DIMENSION 5

Monitoring and analytics

Once in production, the system needs to be carefully monitored

- User feedback & usage statistics
- AI-based fact checking – online or offline
- Continuing with sampling for Expert Review

Performance benchmarking important when experimenting with changes

- Use LLMOps / MLOps to track experiments and get visibility on internals – MLflow or LangSmith or similar

How Financial Services impacts the solution

- **No Financial Advice**	Domain-specific guardrails
- **Vulnerable customers**	Hand-off to human
- **Tax and pension rules vary each tax year** (or Government!)	Data becomes stale quickly
- **Lots of unfamiliar acronyms and defined terms**	Off-the-shelf LLMs easily misunderstand

Thanks for listening

Tom Dorrington Ward, CTO & Co-founder

tom@engagesmarter.ai

ENGAGE SMARTER AI

Expert AI for Financial Guidance.

Empowering customers through smarter financial decisions.