Institute and Faculty of Actuaries

IFoA Life Conference

C5: Can we make conversational AI good enough to put in front of your customers?

Matt Gosden – Engage Smarter AI

Can we make conversational AI good enough to put in front of your customers?

Agenda

1. Why we need customer-facing conversational AI?
2. Intuition on Language AI and its key failure modes today
3. Some techniques for creating reliable Language AI systems

Agenda

1

Why we need customer-facing conversational AI?

2

Intuition on Language AI and its key failure modes today

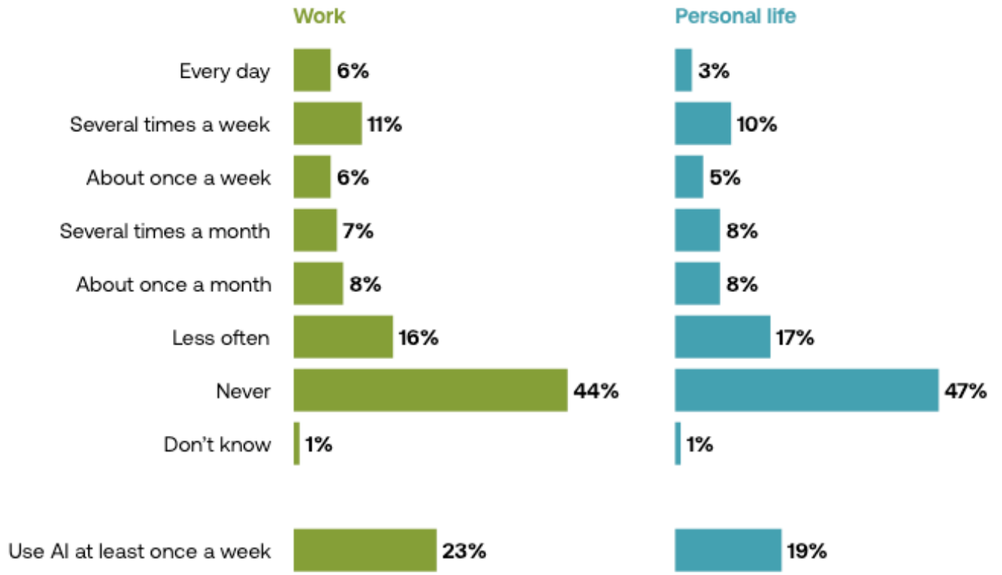
3

Some techniques for creating reliable Language AI systems

AI tool adoption is already very significant

1 in 4

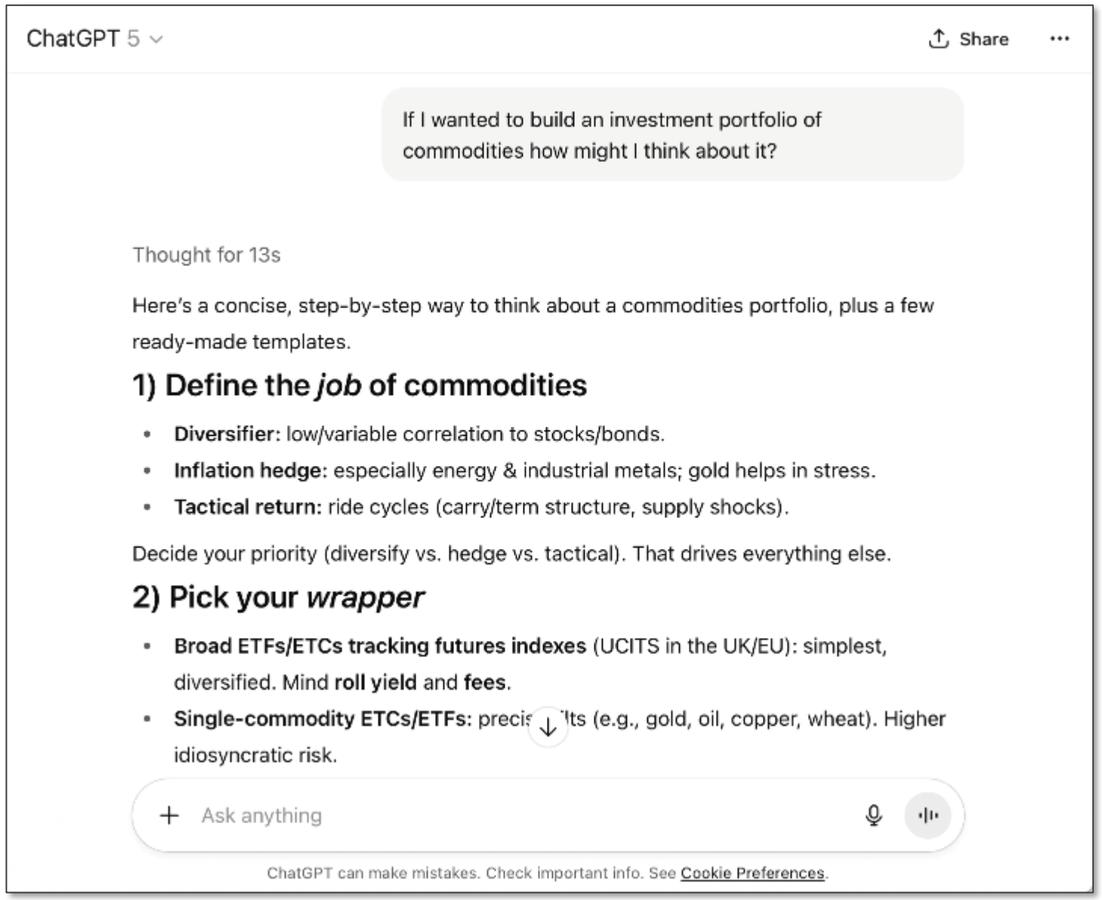
UK adults use GenAI tools at least weekly



Q: AI tools like ChatGPT, DALL-E, Midjourney and others are now used to generate various types of content, including text, images, videos and films. In the past 12 months, how often, if at all, have you used such generative AI tools in your work / personal life?

Source: Tony Blair Institute - What the UK Thinks About AI: Building Public Trust to Accelerate Adoption (Sept 2025)

Which AI tools are consumers using today?



ChatGPT

Gemini

Copilot

perplexity

Claude

Consumers are already using AI for financial tasks

66%

who have used GenAI before say they use it to **seek financial advice**

For Gen Z and Millennials this rises to 82%

(US data)

Source: Credit Karma – Rise of FinAI survey USA (Sept 2025)

Consumers are already using AI for financial tasks

Top GenAI use cases:

#1 – Health and Wellness

#2 – Finance

35% Financial Concepts
35% Goal setting and action plans
34% Budgeting and expenses
33% Optimising savings
32% Investing and stock market

(US data)

Source: Credit Karma – Rise of FinAI survey USA (Sept 2025)

Consumers are already using AI for financial tasks

75%

say they feel GenAI lets them ask the financial questions **they'd be too embarrassed to ask anyone else**

(US data)

Source: Credit Karma – Rise of FinAI survey USA (Sept 2025)

But most AI systems make too many mistakes

52%

Of those who have acted on financial advice they received from GenAI say they made a **poor financial decision based on the information they received**.

(US data)

Source: Credit Karma – Rise of FinAI survey USA (Sept 2025)

But most AI systems make too many mistakes

UK Pensions Question Answering Accuracy Study (2024)

accuracy

98% Pensions Expert AI

80% Human staff (actual responses)

60% GPT4

Source: Engage Smarter Pensions Questions Study (2024)

How should the finance industry respond to this need?

Agenda

1

Why we need customer-facing conversational AI?

2

Intuition on Language AI and its key failure modes today

3

Some techniques for creating reliable Language AI systems

Basics of Language AI Models

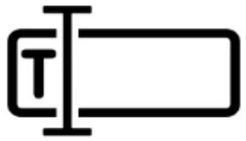
Input

Model

Output

Text Generators

Text Embeddings

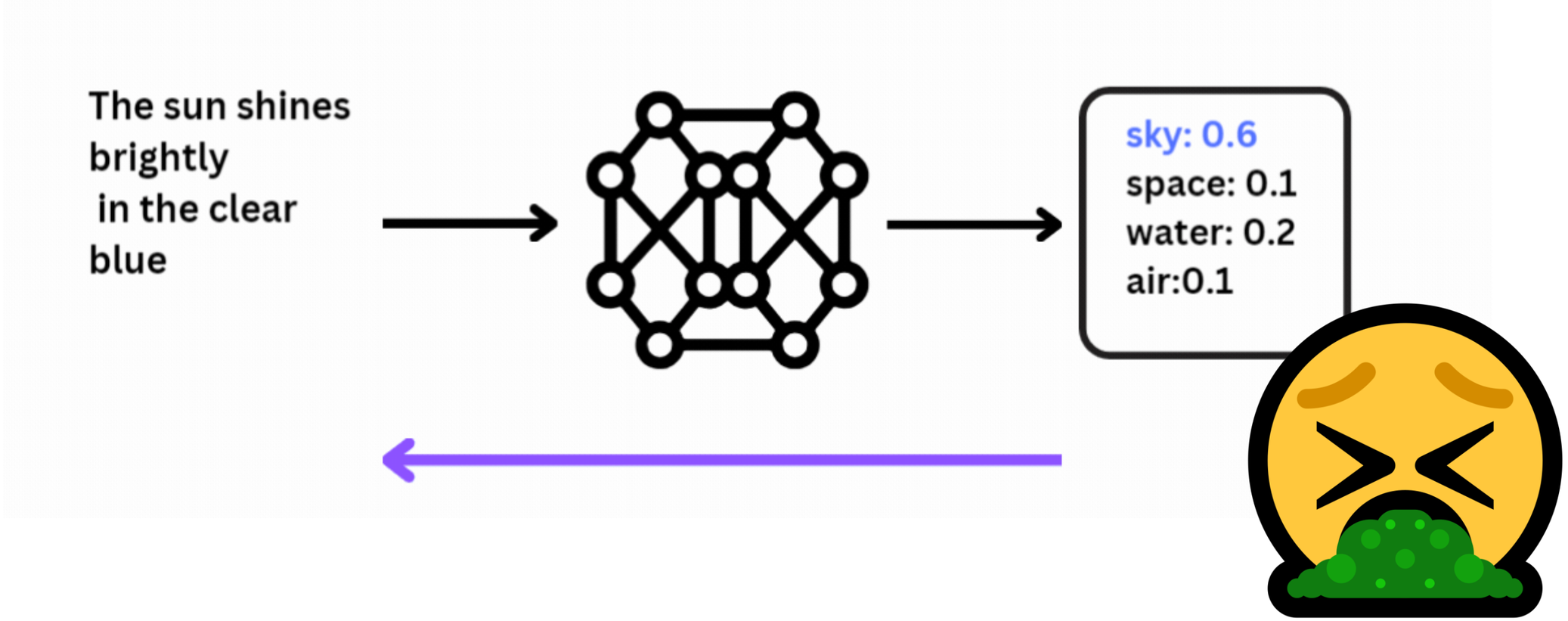


[0.234, -0.011, ... , 2.121]

Pre-Trained on vast Internet Data

Post-Trained on Human Preferences

Today's Large Language Models are auto-regressive next token generators – they are 'simulators'



"LLM" = Large Language Model
"Token" = Parts of words
"Auto-regressive" = Generate tokens one at a time

Naturally this leads to some big problems for LLMs

Pre-trained on Internet Data → Factual Errors and Bias

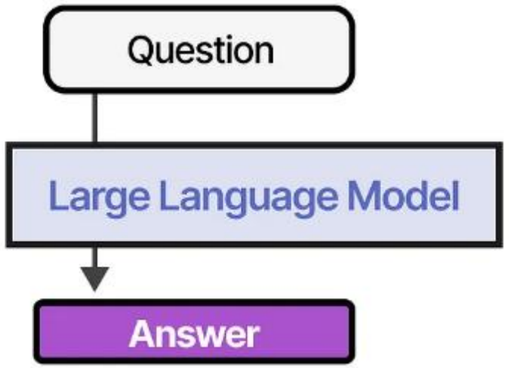
Post-trained on Human Preferences → Hallucinations and Sycophancy

Auto-regressive Generation → Errors in long responses

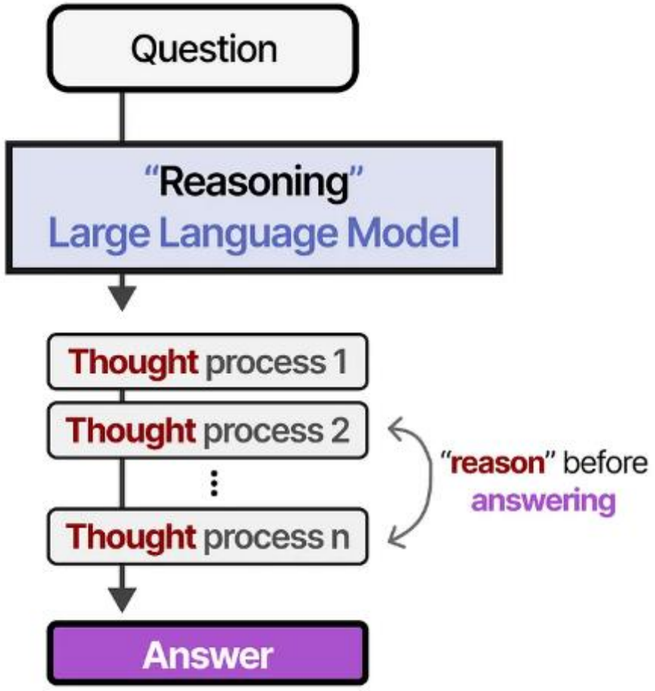
Input Context Window → Compounding Errors and Jail-breaking

Recent Improvements – Reasoning Models

Regular LLMs

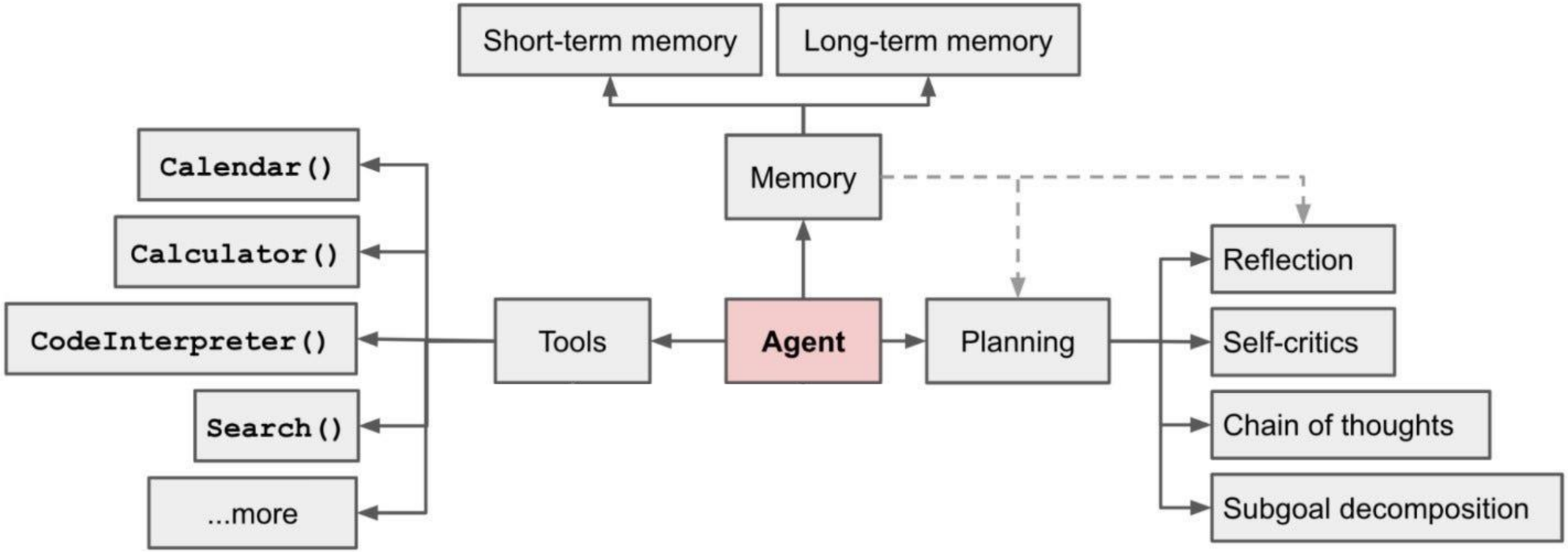


Reasoning LLMs



What if we used more compute to explore the answer before responding?

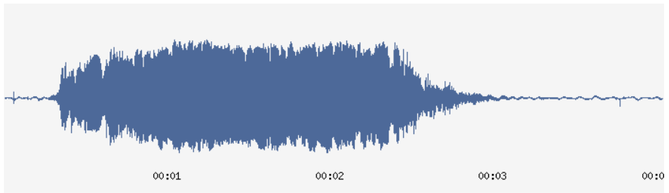
Recent Improvements – Tool Use

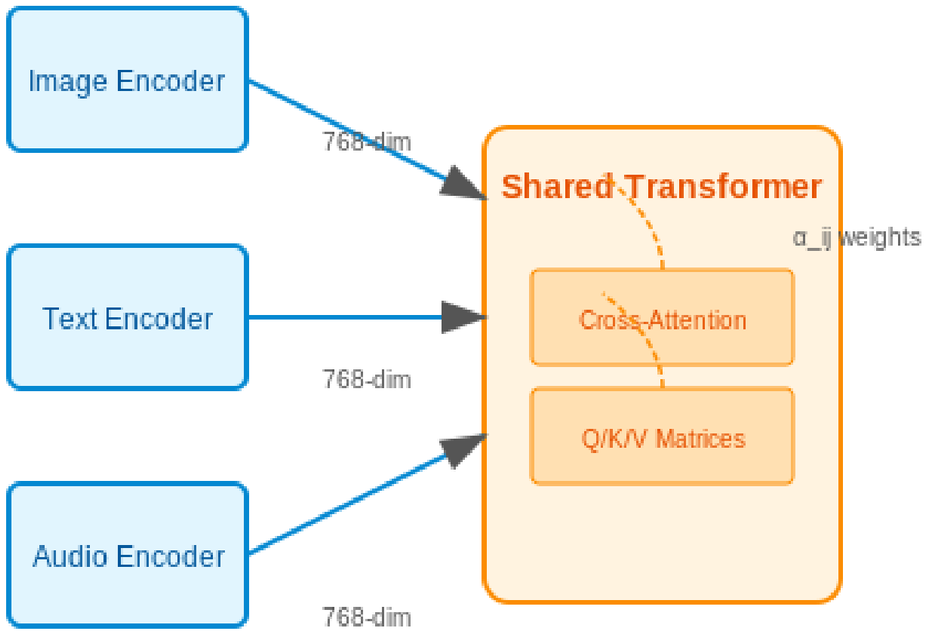


"Tool" = something the LLM (Agent) can call to perform a specific sub-task

Recent Improvements – Multi-modality

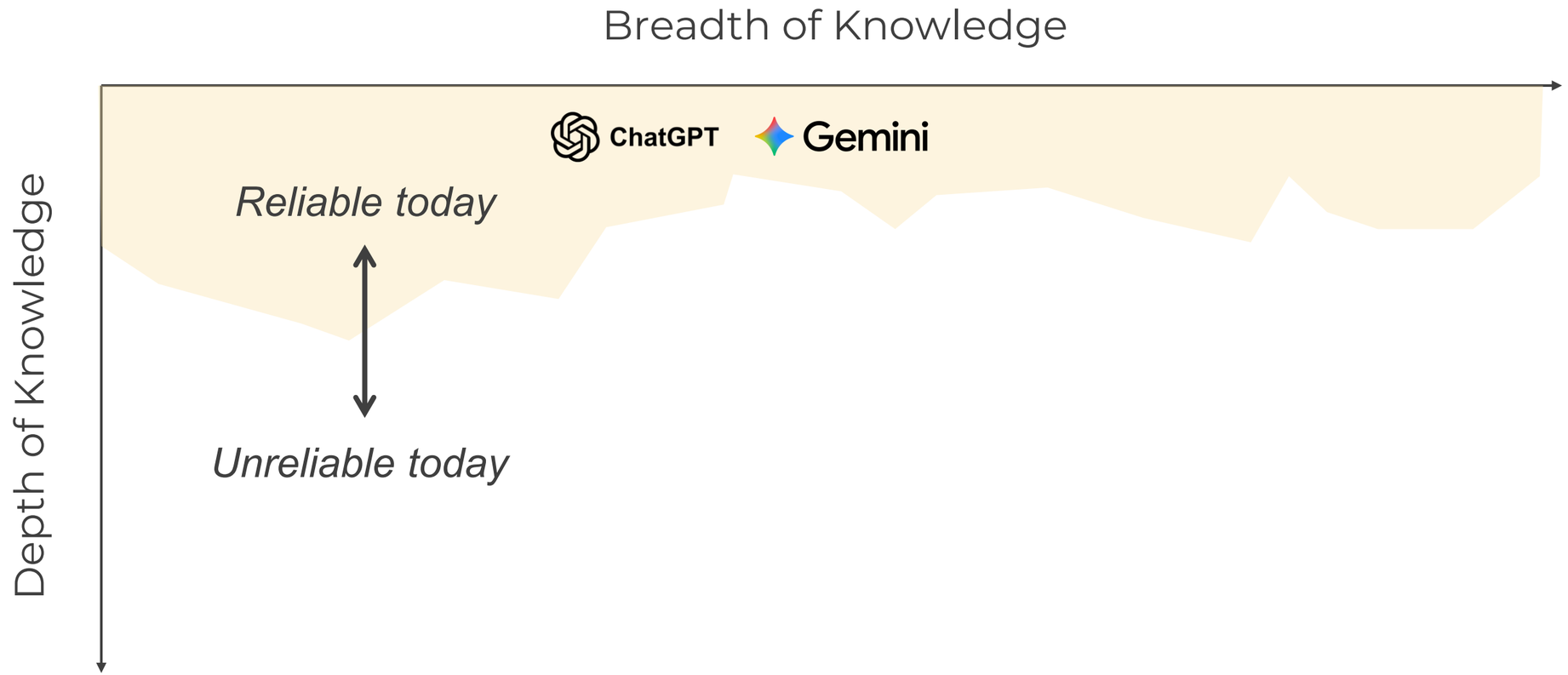
"The peasants say it is the Hound of the Baskervilles calling for its prey ..."



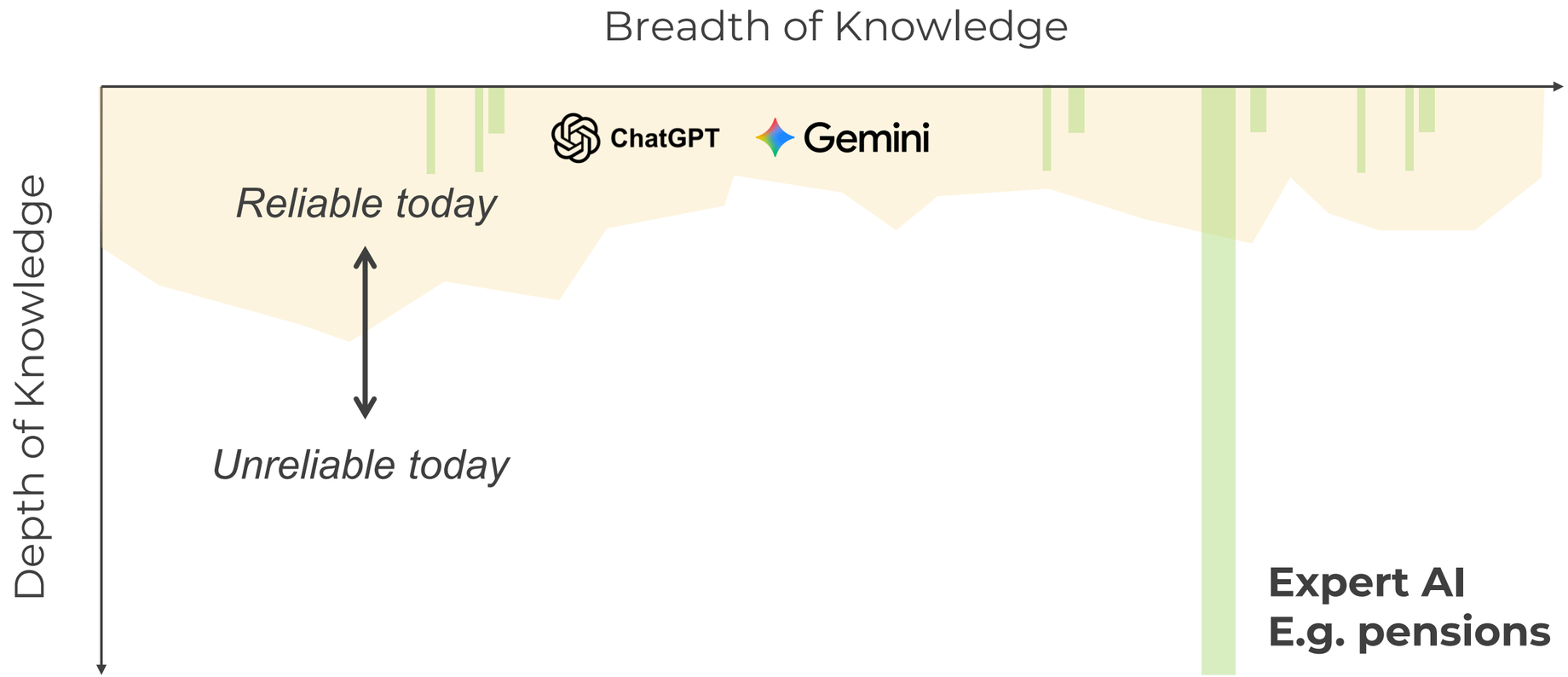


"multi-modality" = understand different types of information

The large AI labs are focused on developing general knowledge and capability



We should be able to build better Narrow Expert AI



Agenda

1

Why we need customer-facing conversational AI?

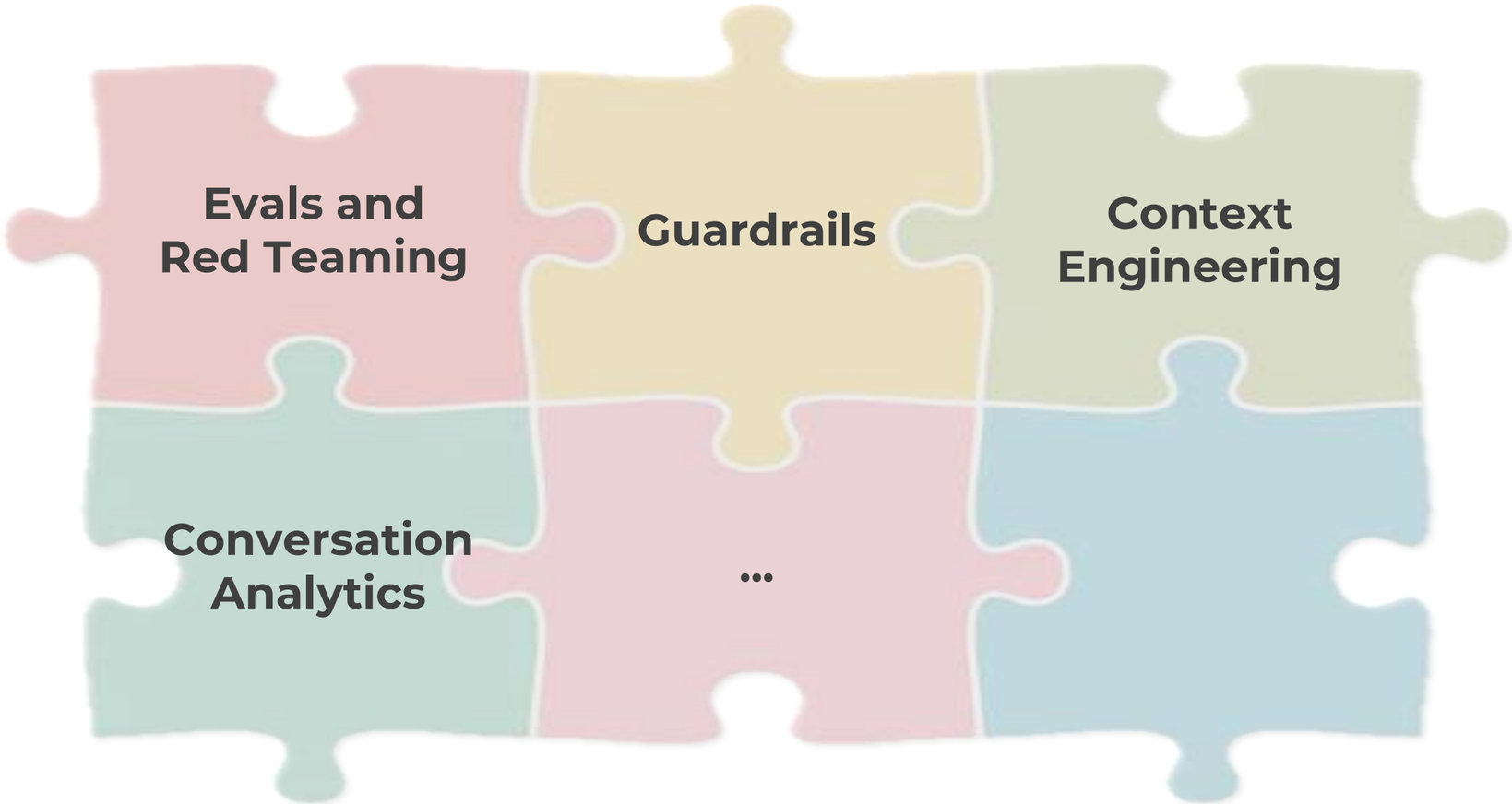
2

Intuition on Language AI and its key failure modes today

3

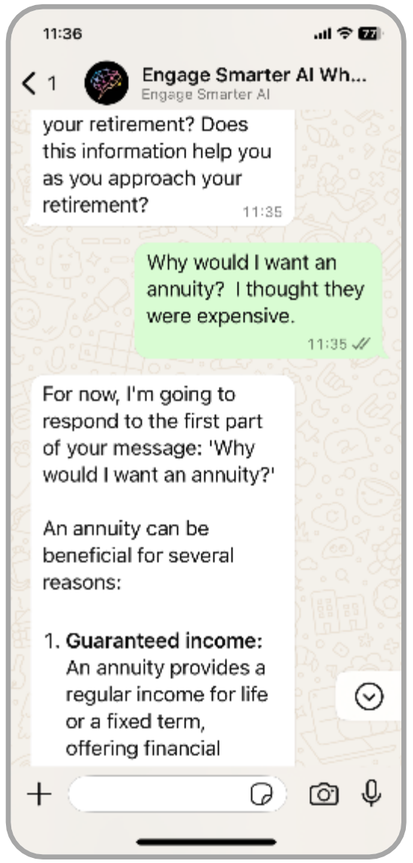
Some techniques for creating reliable Language AI systems

A selection of methods that help make Language AI Systems more reliable

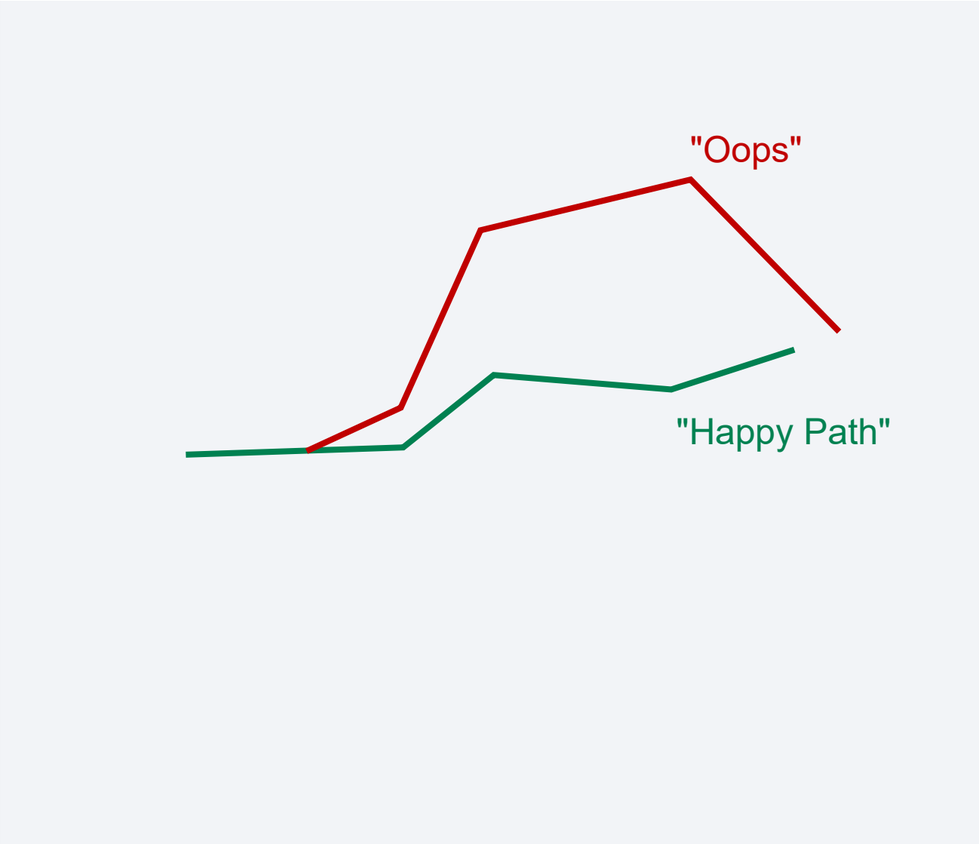


Evals enable us to understand and test accuracy

Demo Day



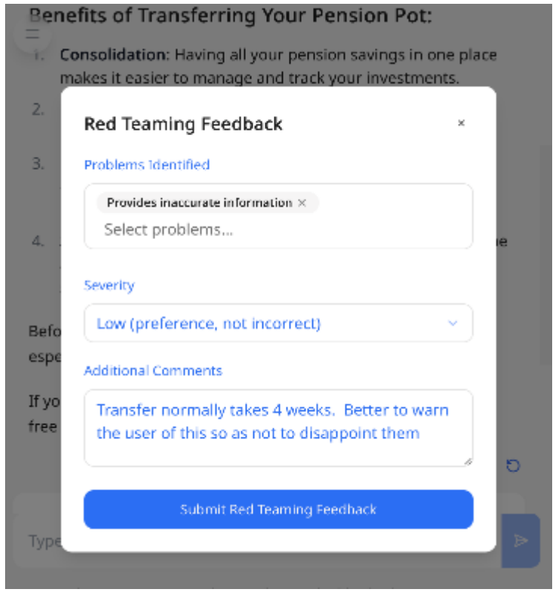
Conversation "Space"



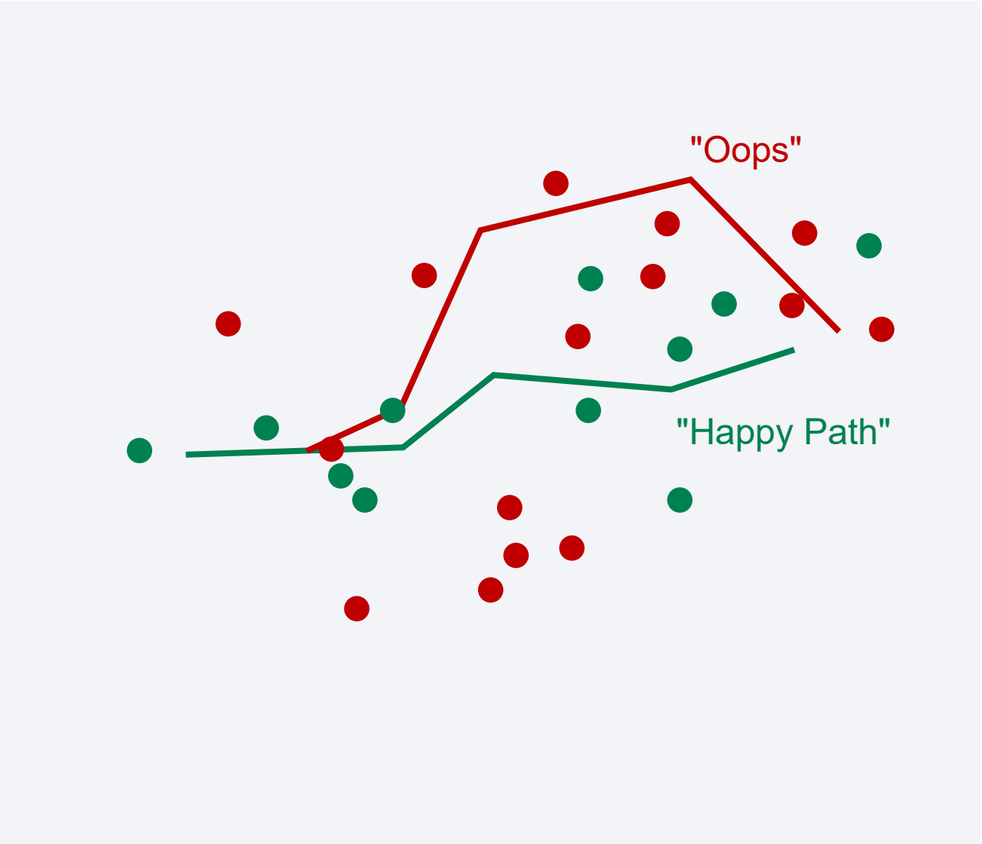
Evals enable us to understand and test accuracy

Testers

Feedback



Conversation "Space"



"Red Teaming" = The goal of breaking a system in order to identify weaknesses
"Evals" = test cases for AI systems

Guardrails allow us to cut out certain failure modes

Traditional Tech Projects

Additive process

LLM Tech Projects

Scope

Advice
Language

Numerical
Hallucination

Fact Checking

Vulnerability

...

Subtractive process

Context Engineering helps us improve the reliability of each LLM call

Scope Guardrail LLM Example

Context

- Transcript
- Instructions
- Examples

Guardrail

Is conversation in scope?

Assessment

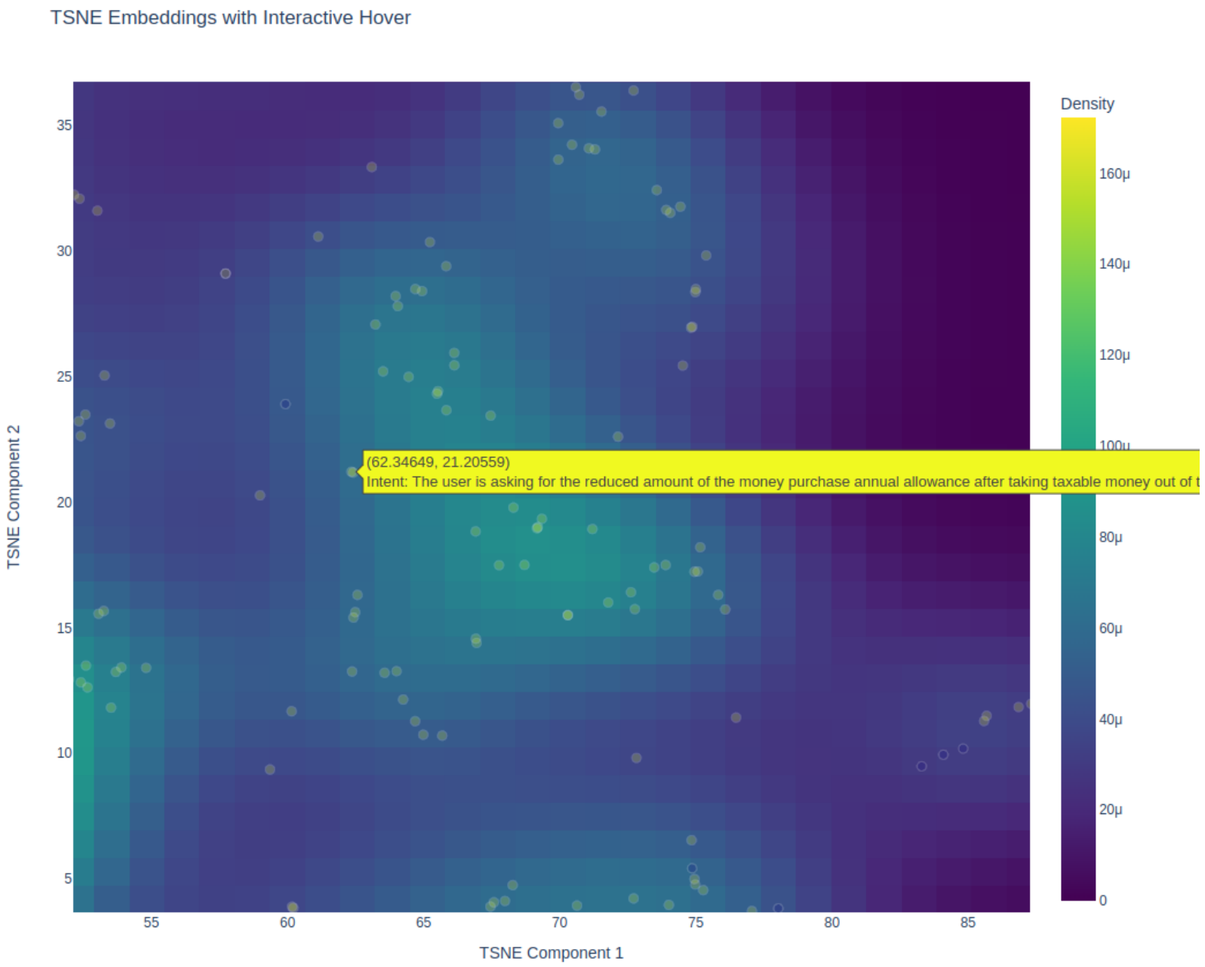
IN = in scope
OUT = out of scope
? = Unclear

What is the Guardrail response?

Ambiguous input → Should be "Unclear"

OUT OF SCOPE

Conversation Analytics allow us to focus where there are likely to be larger risks



- Is this conversation 'novel' and should be checked?
- Is this conversation close to a cluster of errors?
- Is this a common conversation we know is reliable?
- Do we have coverage of Evals in this space?

Three Takeaways

1. Consumers seem to want AI support for their finances – I think we need to be engaging with consumers directly.
2. Today's AI models are getting better but are focused on improving general intelligence, not domain accuracy.
3. We can build much more reliable domain-specific AI Systems and build strong risk management into them.

Source: Credit Karma – Rise of FinAI survey USA (Sept 2025)

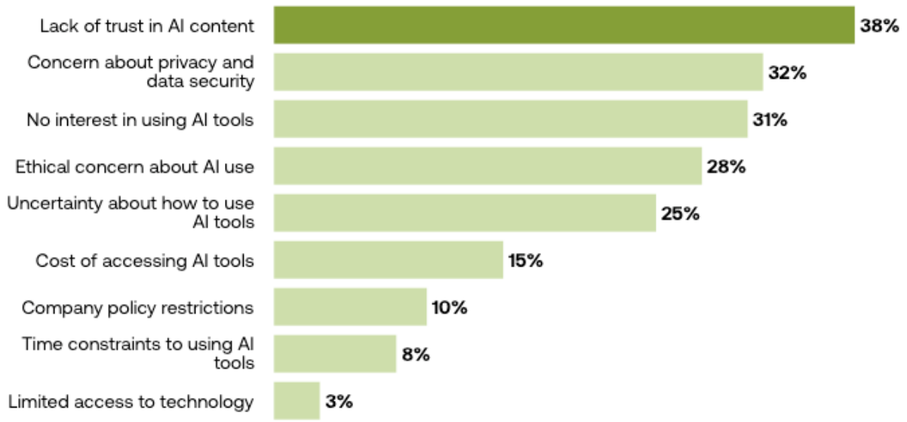
Thanks!

Q&A

Many consumers recognise the trust issue

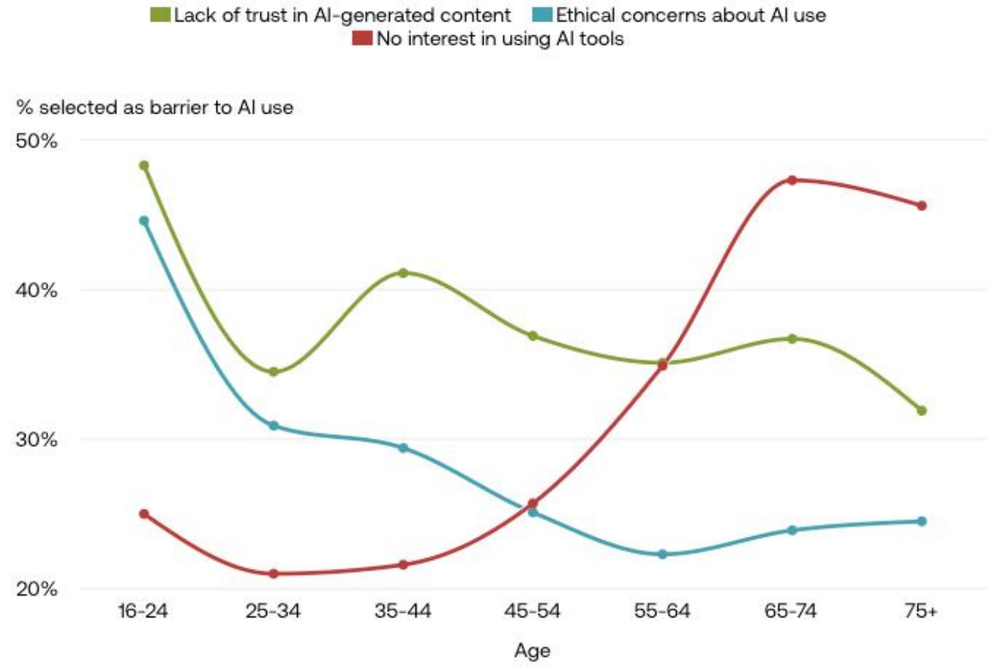
FIGURE 6

More than 30 per cent of UK adults view a lack of trust in AI-generated content and concerns about privacy and data security as the biggest barriers to adoption



Q: Which of the following barriers, if any, do you face when using or considering the use of generative AI tools (such as ChatGPT, DALL-E, or Midjourney)?

Source: Ipsos

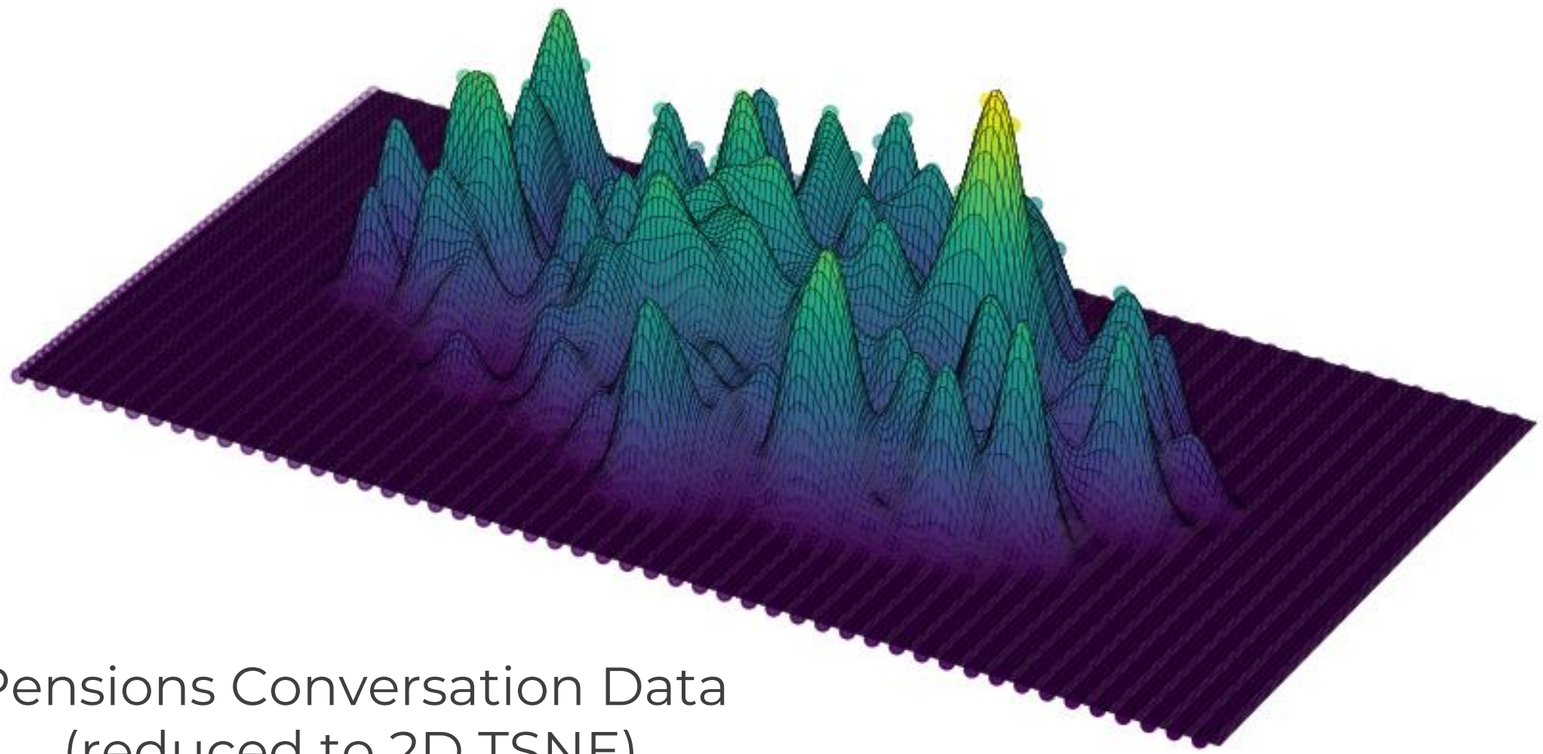


Q: Which of the following barriers, if any, do you face when using or considering the use of generative AI tools (such as ChatGPT, DALL-E, or Midjourney)?

Source: Ipsos

Source: Tony Blair Institute - What the UK Thinks About AI: Building Public Trust to Accelerate Adoption (Sept 2025)

Changing the weights (fine-tuning)



Pensions Conversation Data
(reduced to 2D TSNE)

Things we want **more of → Pull up**

Things we want **less of → Push down**

Source: Engage Smarter AI Pensions Data

Walk through a typical project ...

Answering customer queries about my product and processes (first-line support)

Probably start with a short sprint

Instructions

Identity

Inject relevant content

Product information

Process information

FAQs

Key Contacts

...