The Chartered Institute for IT

Faculty of Health and Care

# Understanding Large Language Models in Healthcare

BCS Clinical Risk Management Specialist Interest Group

Webinar Series 1

12th June 2025

# Macro Challenges facing the NHS

## Cut Costs

**£6.6bn**

2025/26 financial deficit. **There is a need to realise efficiencies that can be sustained.**

## Staff Shortages

**6.7%**

Total Vacancy Rate as of March 2025, equating to ~100k FTE. **There is a need to supplement a stretched workforce.**

## Increase Productivity

**+1.9%**

Average Per Year Productivity increases until 2029/30 to meet demand. **This must be done by a workforce already stretched.**

> Artificial intelligence is a technology with almost unlimited power to cut waste, speed things up and save lives. And with this plan – the NHS will harness it properly.
>
> **Sir Kier Starmer, 2025**

# So why LLMs in Health?

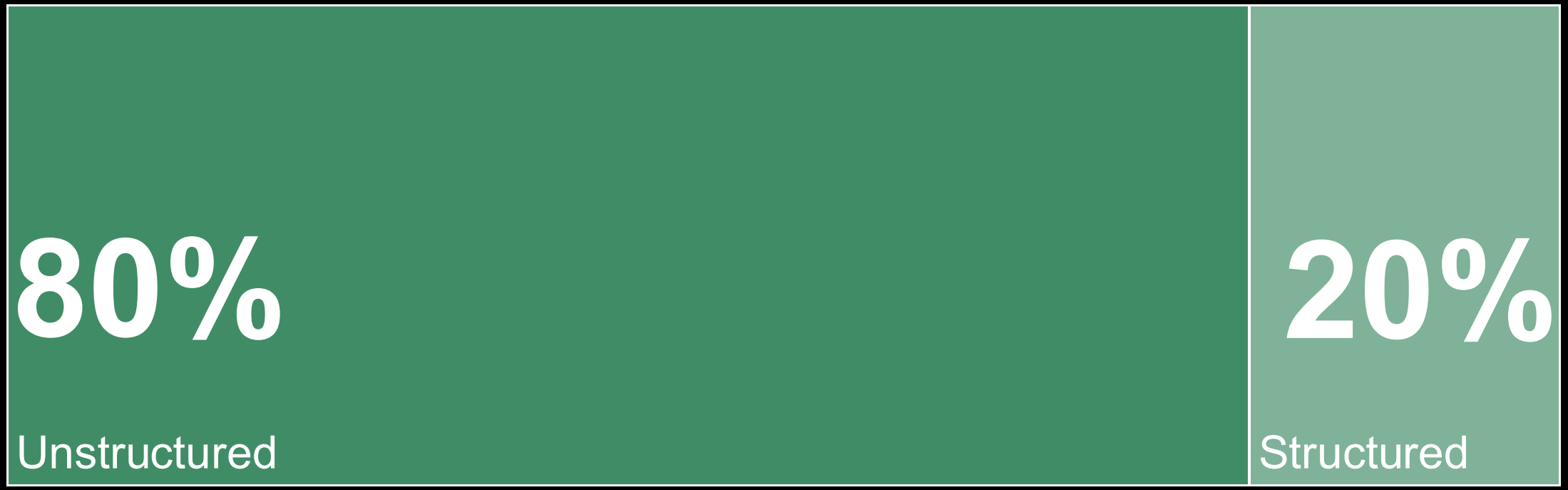
1. LLMs are suited for **any application that requires** understanding or interacting in **natural human language.**

2. LLMs are essential for **generating** natural, **human-like text.**

3. They are **well suited to Health Data**.

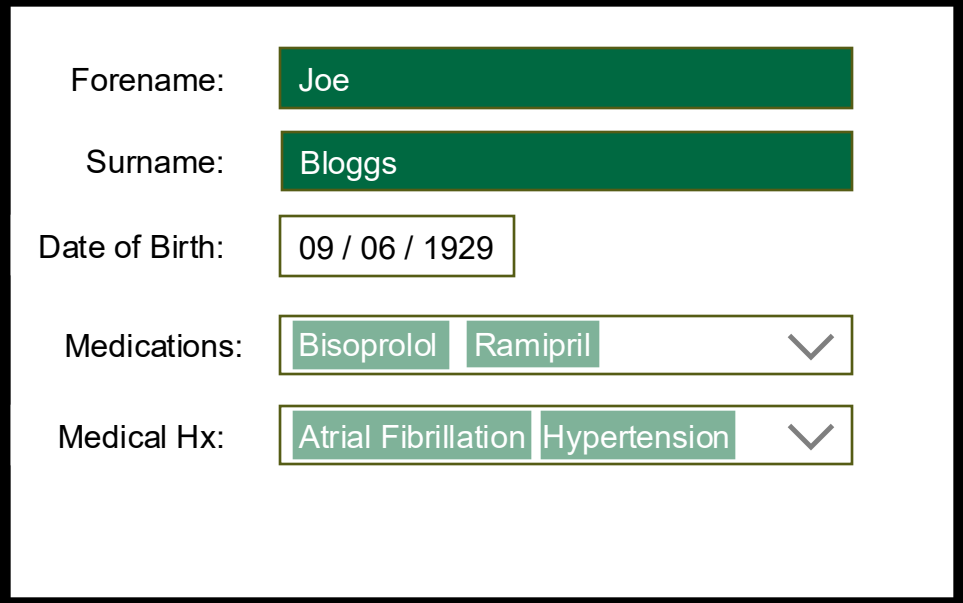
# Health Data Landscape

**All Healthcare Data estimate split***




***npj Digital Medicine, 2024**

# Structured Data




| Feature | Structured Data |
|---|---|
| **Format** | Predefined, organised (like a spreadsheet) |
| **Storage** | Stored in organised tables (databases like Excel, SQL) |
| **Ease of Analysis** | Easy to search, filter, and analyse |
| **Examples** | Patient names, age, blood pressure (numbers, codes) entered in set fields |
| **Data Type** | Text, Numbers, Dates, Categories |
| **Tools Needed** | Excel Spreadsheets, SQL |

# Unstructured Data

Case Note

Patient reports increasing fatigue and shortness of breath over the past 3 days.

Suspected heart failure exacerbation.

| Feature | Unstructured Data |
|---|---|
| **Format** | Free-form, could be handwritten. |
| **Storage** | Stored in files (PDF documents, videos, images) |
| **Ease of Analysis** | Harder to search, needs special tools |
| **Examples** | Clinician notes, Discharge letters, X-rays, audio recordings |
| **Data Type** | Text, images, audio, video |
| **Tools Needed** | Natural Language Processing, Optical Character Recognition, Computer Vision |

**The power of a Large Language Model is it can process this unstructured data, derive context and generate an output.**

# So, what are LLMs?

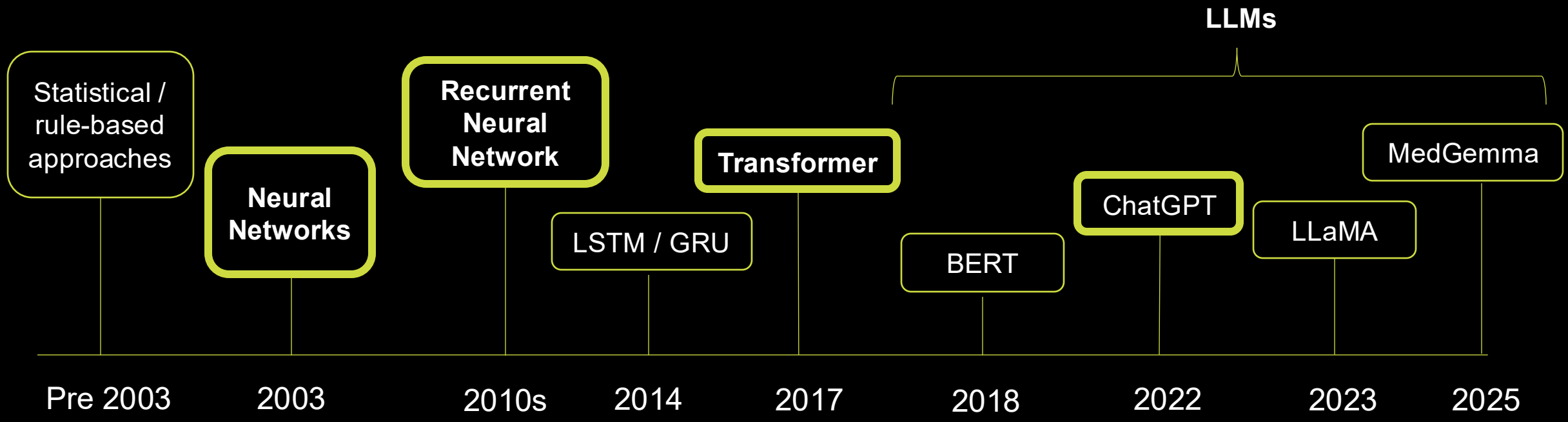
1. What does **LLM** mean
2. The history of **Language Modelling**
3. Foundation models
4. Opensource
5. Tailoring LLMs to the **medical domain**

# What is a Language Model?

Language models are **mathematical approaches** to **understanding, generating and predicting** human language.

They support us in tasks like translation, text generation and determining sentiment.

# How has Language Modelling changed over time?




*LSTM = Long short-term memory
*GRU = Gated Recurrent Unit
*BERT = Bidirectional Encoder Representations for Transformer
*GPT = Generative Pre-trained Transformer

# How has Language Modelling changed over time?

## NEURAL NETWORK

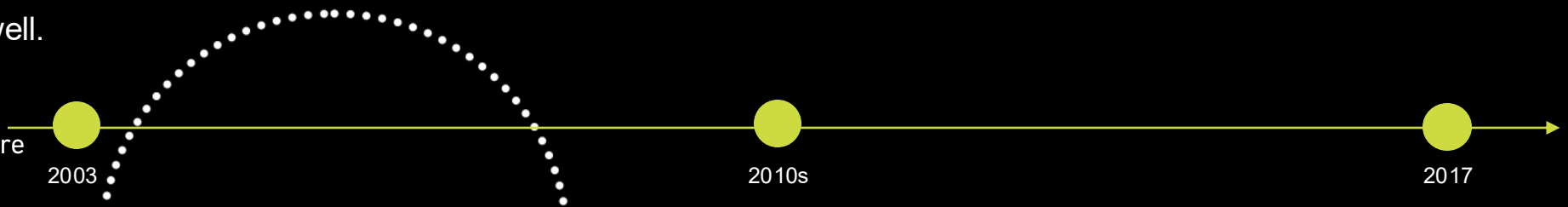
Introduced complex computation to language modelling. However, were unable to process context well.

## RECURRENT NEURAL NETWORKS

Were able to handle context, however, were slow to process information.

## TRANSFORMER

Brought about a new architecture for processing information that allowed a rich contextual understanding coupled with fast processing. This means we can train models built on this architecture with more data than ever before.



# How has Language Modelling changed over time?

## Neural Network

Dr enters consultation room

Dr has no memory, or awareness of what came before

Dr sees a one sentence sticky note of the patient's history

Dr makes a decision

2003

## Recurrent Neural Network

Dr enters consultation room

Dr hears one word at a time

Dr is transcribing as they go

Dr tries to remember what was said earlier

By the time the patient gets to the important information, they may have forgotten how it began

2010s

## Transformer

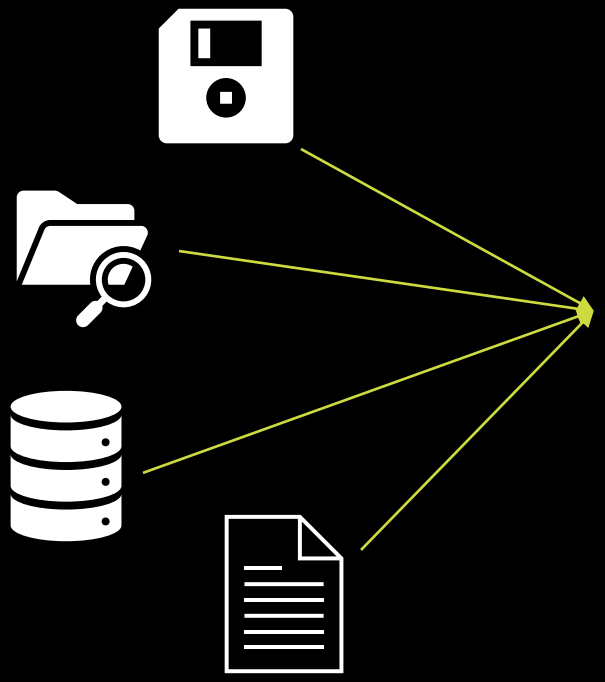
Dr enters the consultation room

Dr sees the entire patient history laid out on a whiteboard in front of them

Dr decides on the course of action with all patient information available and easily referenceable

2017

# From Transformer to LLM



**Transformer** → **Foundational LLM**

# Foundational LLMs

**Large models trained on broad data that can be adapted to a wide range of downstream tasks. Examples include:**

Company Name:
**OpenAI**
Model Family:
**GPT Series**
Notable implementations:
**ChatGPT**

Google

Company Name:
**Google**
Model Family:
**BERT, Gemini**
Notable implementations:
**Gemini**

Meta

Company Name:
**Meta (Facebook)**
Model Family:
**LLaMA series**
Notable implementations:
**LLaMA 2, LLaMA 3**

# Open and Closed Source LLMs

**Open source models are those with publicly available and accessible source code that anyone can modify and share. Not all foundation models are opensource, different organisations take different approaches:**

Company Name:
**OpenAI**
Model Family:
**GPT Series**
Notable implementations:
**ChatGPT**
Open Source:
**No**

Company Name:
**Google**
Model Family:
**BERT, Gemini**
Notable implementations:
**Gemini**
Open Source:
**Model dependent**

Company Name:
**Meta (Facebook)**
Model Family:
**LLaMA series**
Notable implementations:
**LLaMA 2, LLaMA 3**
Open Source:
**Yes**

# Medical LLMs: Overview

**We can tailor LLMs to ensure they perform more robustly in a medical setting.**

## Prompt Engineering

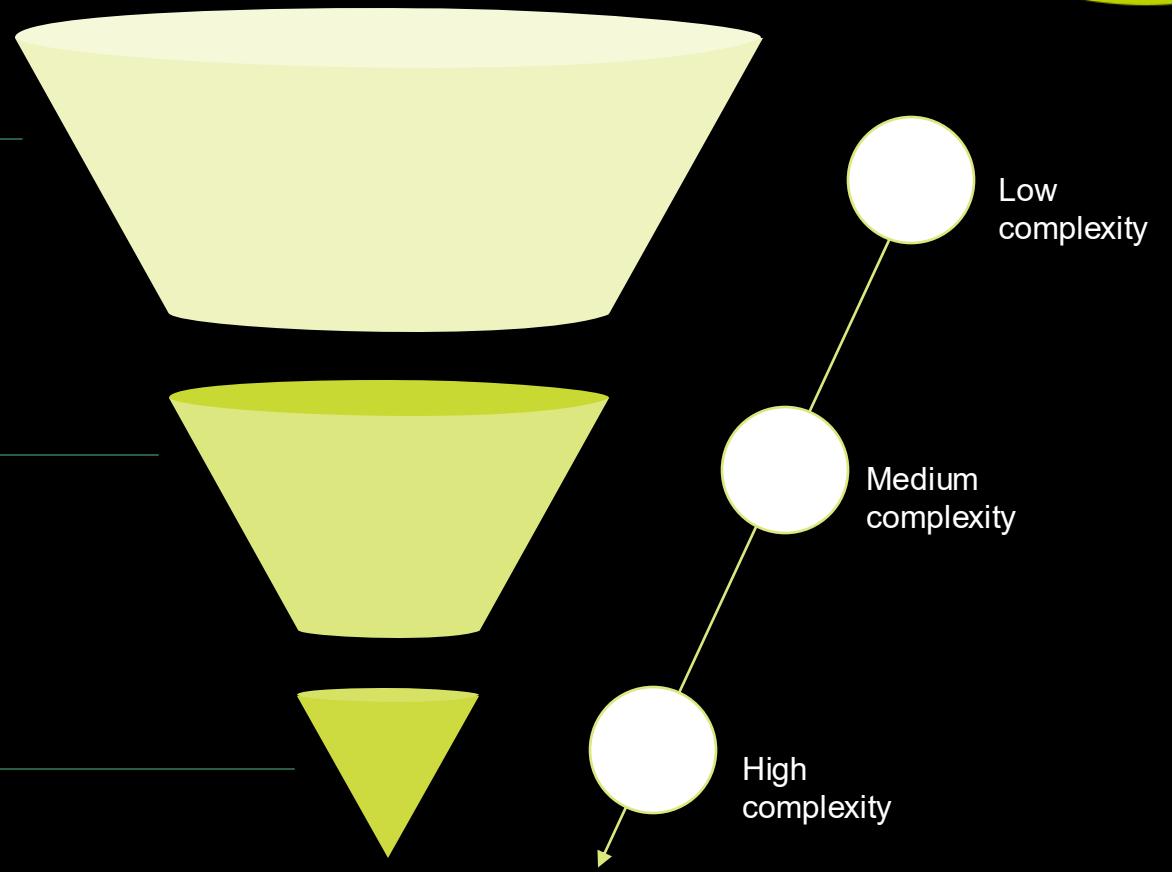
Use structured instructions to guide the model's behaviour without changing it.

## Prompt Tuning

Train domain specific vectors to be applied to the beginning of a prompt.

## Fine Tuning

Retrain a model using curated clinical datasets



# Medical LLMs: Overview

**A list of Medical specific LLMs* and their strengths**

| Name | Builders | Primary Task | Approach to creation |
|---|---|---|---|
| Med-PaLM 2 | Google | Medical Q&A | Instruction prompt tuning |
| ClinicalBERT | Harvard, Princeton, NYU | Readmission prediction | Fine tuning |
| BioGPT | Microsoft | Relation extraction, Q&A, document classification | Pretraining + Fine tuning |
| LLaMA 3 (medical variants) | *Many developers across the open source community* | Training specific | Fine tuning/prompt tuning |
| MedGemma | Google | Multimodal clinical reasoning | Fine tuning |



*Non-Exhaustive

# Use Cases for LLMs in Healthcare are rapidly evolving

Business Priorities

| Optimising the Back Office | Reduce Clinician burden & shortage impact | Prevention & population health | Improving the Patient Experience | Research & Pharma Innovation |
|---|---|---|---|---|
| **Waitlist Validation & Clinical Appt Optimisation** | **Ambient Voice Technology** | Population Risk Stratification | **Customer Service Chatbots** | Synthetic Patient Data Generation |
| **Back Office Co-Pilots** | **Discharge Document Generation** | **Population Health Forecast & Interventions** | Self-Care Assistant | Clinical Trial Support Assist |
| Clinical Coding | Early Interventions | Discharge Document Generation | Mental Health Support | Research Model Build Co-Pilot |
| Supply Chain Optimisation | Clinical Decision Support | Public Health Policy Development | Personalised Treatment Plans | **Automated Literature Review** |

Use Cases*

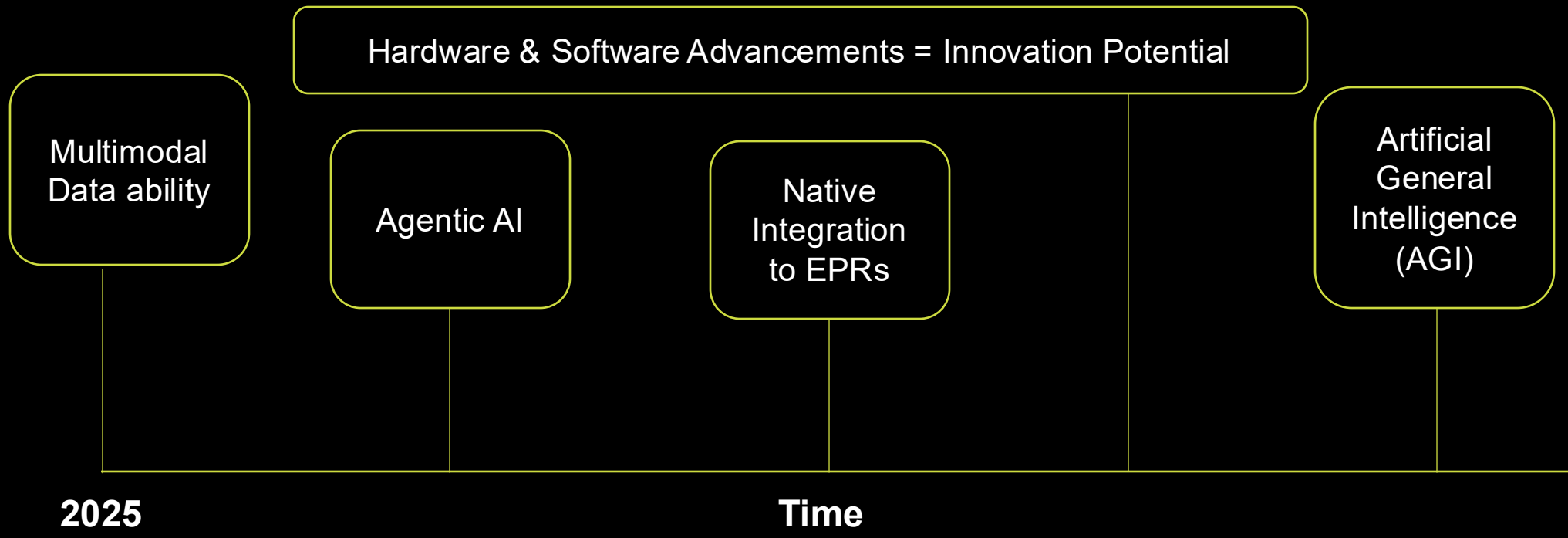
*Non-Exhaustive

# The Art of the Possible: Virtual Care powered by LLMs

Video shown during presentation:

Hippocratic AI Generative AI Healthcare Agents - YouTube

# What does the future hold?




**As Use Cases for LLMs proliferate, Safety will become increasingly important.**

# LLM-Driven Technologies

### M365 Copilot

AI assistant designed to enhance productivity in the M365 product suite.

### Ambient Scribe

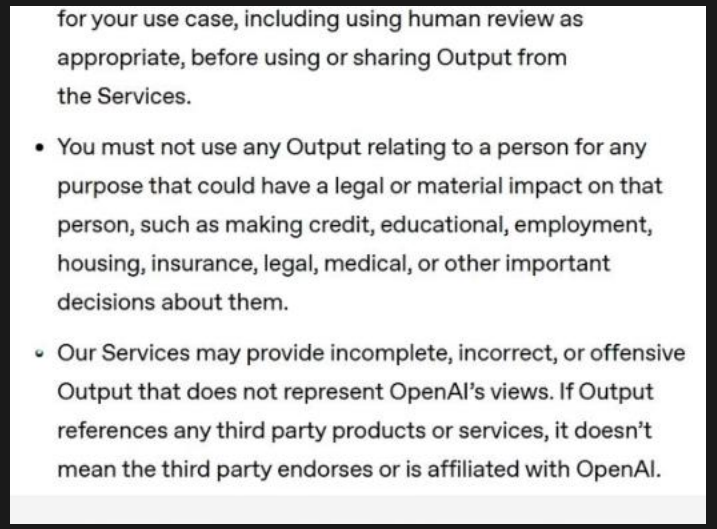
LLM-based tool designed to support medical documentation and clinical workflows.

---

Examples of foundation models underpinning AI applications



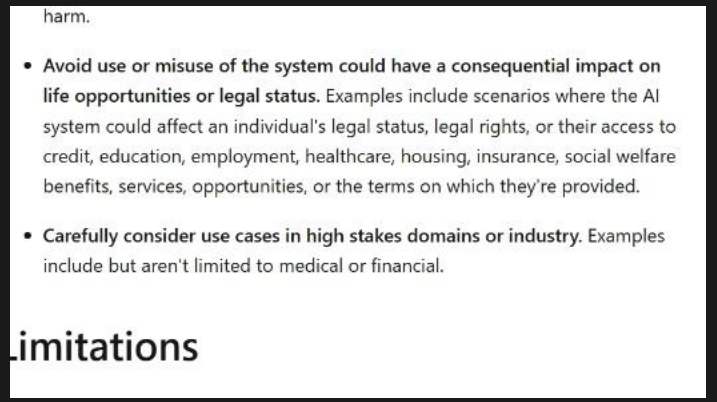

# Responsible AI: Transparency and Usage


for your use case, including using human review as appropriate, before using or sharing Output from the Services.

- You must not use any Output relating to a person for any purpose that could have a legal or material impact on that person, such as making credit, educational, employment, housing, insurance, legal, medical, or other important decisions about them.
- Our Services may provide incomplete, incorrect, or offensive Output that does not represent OpenAI's views. If Output references any third party products or services, it doesn't mean the third party endorses or is affiliated with OpenAI.

## Open AI Terms of Use

Foundation Model Disclaimer

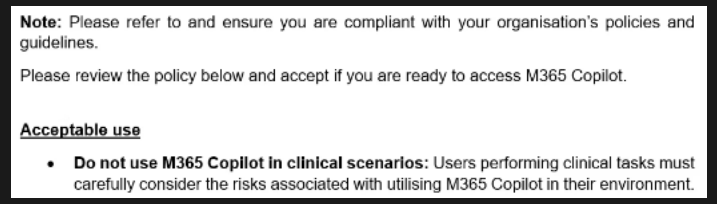

harm.

- **Avoid use or misuse of the system could have a consequential impact on life opportunities or legal status.** Examples include scenarios where the AI system could affect an individual's legal status, legal rights, or their access to credit, education, employment, healthcare, housing, insurance, social welfare benefits, services, opportunities, or the terms on which they're provided.
- **Carefully consider use cases in high stakes domains or industry.** Examples include but aren't limited to medical or financial.

Limitations

## M365 Copilot Transparency Note

Product Disclaimer


**Note:** Please refer to and ensure you are compliant with your organisation's policies and guidelines.

Please review the policy below and accept if you are ready to access M365 Copilot.

**Acceptable use**

- **Do not use M365 Copilot in clinical scenarios:** Users performing clinical tasks must carefully consider the risks associated with utilising M365 Copilot in their environment.

## Acceptable Use Policy

Client Disclaimer

# Risks of ChatGPT in Clinical Practice

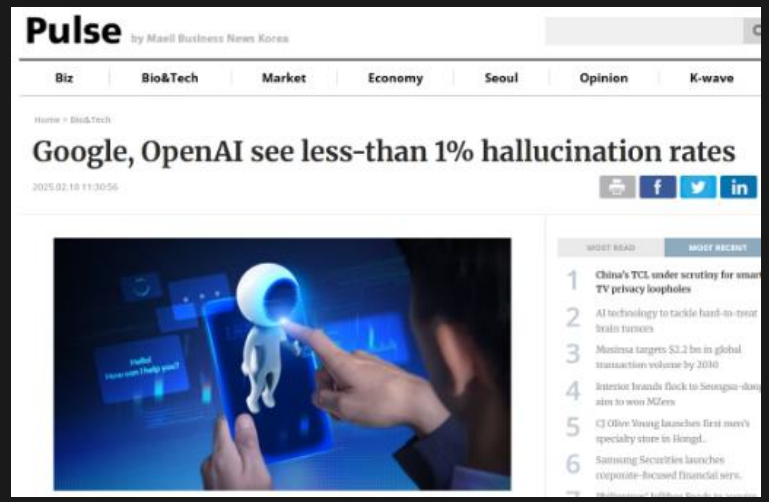
One in five GPs use AI such as ChatGPT for daily tasks, survey finds

Doctors are using the technology for activities such as suggesting diagnoses and writing letters, according to BMA

AI Hallucination Leaderboard

| Model | Hallucination rate | Factual consistency rate |
|---|---|---|
| Google Gemini-2.0-Flash-001 | 0.7 % | 99.3 % |
| Google Gemini-2.0-Pro-Exp | 0.8 % | 99.2 % |
| OpenAI o3-mini-high | 0.8 % | 99.2 % |
| Vectara Mockingbird-2-Echo | 0.9 % | 99.1 % |
| Google Gemini-2.5-Pro-Exp-0325 | 1.1 % | 98.9 % |
| Google Gemini-2.0-Flash-Lite-Preview | 1.2 % | 98.8 % |
| OpenAI GPT-4.5-Preview | 1.2 % | 98.8 % |
| Zhipu AI GLM-4-9B-Chat | 1.3 % | 98.7 % |
| Google Gemini-2.0-Flash-Exp | 1.3 % | 98.7 % |
| Google Gemini-2.5-Flash-Preview | 1.3 % | 98.7 % |
| OpenAI-o1-mini | 1.4 % | 98.6 % |
| OpenAI GPT-4o | 1.5 % | 98.5 % |
| Amazon Nova-Micro-V1 | 1.6 % | 98.4 % |
| OpenAI GPT-4o-mini | 1.7 % | 98.3 % |
| OpenAI GPT-4-Turbo | 1.7 % | 98.3 % |

*Source: Hugging Face, as of April 29, 2025*



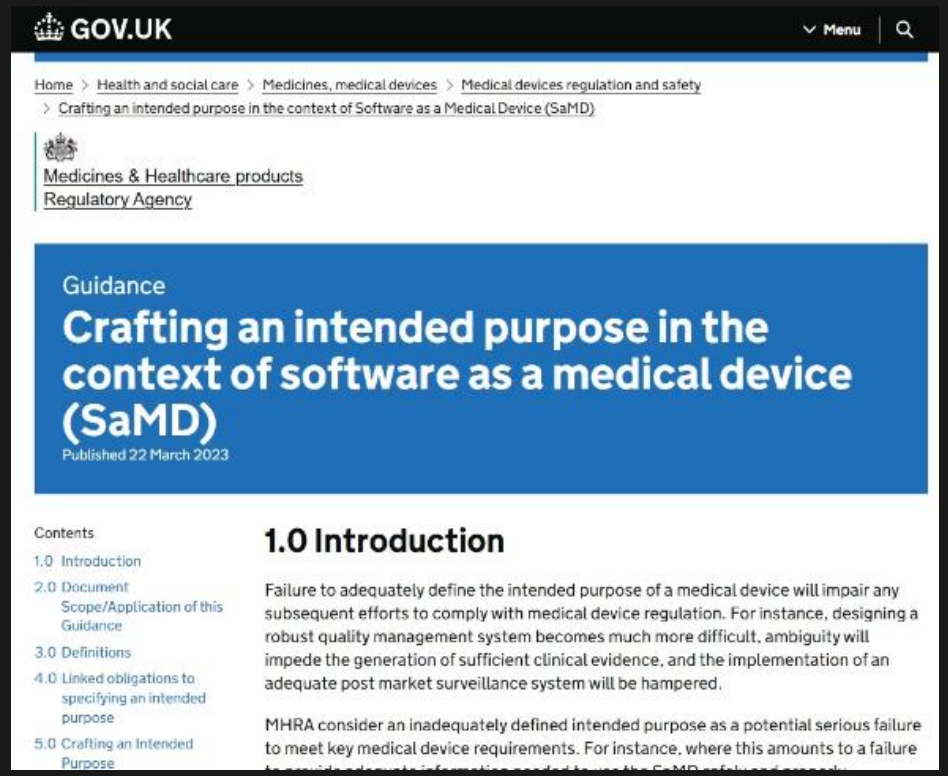
Pulse by Maeil Business News Korea

Biz | Bio&Tech | Market | Economy | Seoul | Opinion | K-wave

Google, OpenAI see less-than 1% hallucination rates

Large language models, such as ChatGPT, are unable to identify if the phrases they generate make sense or are accurate. This can sometimes lead to inaccurate results, also known as **hallucination effects,** where large language models generate plausible sounding but inaccurate text. Hallucinations can also result from biases in training datasets or the model's lack of access to up-to-date information

# Medical Device Intended Use



GOV.UK Menu

Home > Health and social care > Medicines, medical devices > Medical devices regulation and safety > Crafting an intended purpose in the context of Software as a Medical Device (SaMD)

Medicines & Healthcare products Regulatory Agency

Guidance

**Crafting an intended purpose in the context of software as a medical device (SaMD)**

Published 22 March 2023

Contents

1.0 Introduction
2.0 Document Scope/Application of this Guidance
3.0 Definitions
4.0 Linked obligations to specifying an intended purpose
5.0 Crafting an Intended Purpose

## 1.0 Introduction

Failure to adequately define the intended purpose of a medical device will impair any subsequent efforts to comply with medical device regulation. For instance, designing a robust quality management system becomes much more difficult, ambiguity will impede the generation of sufficient clinical evidence, and the implementation of an adequate post market surveillance system will be hampered.

MHRA consider an inadequately defined intended purpose as a potential serious failure to meet key medical device requirements. For instance, where this amounts to a failure

> **"the use for which the device is intended according to the data supplied by the manufacturer on the labelling, in the instructions and/or in promotional materials."**

⚠ Under the UK Medical Devices Regulations (UK MDR) 2002, what counts is not just the intended purpose declared by the manufacturer, but **also the actual claims** made about the device

# National Guidance - Ambient Scribe

| Function | Low Functioning | High Functioning |
|---|---|---|
| Speech-to-text transcription | Verbatim transcript for clinician review | - |
| Summarisation of patient consultation | - | AI-generated summary or compressed clinical note |
| Formatting outputs (e.g. SOAP notes, letters) | Manual formatting by clinician | Automatic formatting using AI or templates |
| Extraction of clinical codes (e.g. SNOMED, ICD-10) | - | Automatic coding of clinical content |
| Population of EHR fields | Manual entry by clinician | Auto-population of structured EHR fields |
| Workflow support (e.g. referral suggestions, task creation) | - | AI-generated clinical actions or recommendations |
| Interpretation of clinical content | - | Interpretation that informs diagnosis, treatment, or triage |
| Overall impact on clinical decision-making | Requires full clinician review and control | Supports or influences decisions without full human oversight |

**Increasing Regulatory Oversight**

# Governance and Compliance*

### Security

Adherence to DSPT and Cybersecurity standards, certifications and frameworks, e.g., ISO/IEC 27001, Cyber Essentials, NCSC Cyber Essentials Plus, Cyber Assessment Framework (CAF)

### SaMD & Clinical Safety Standards

Implementation of DCB0129/0160 standards and SaMD standards, e.g., ISO14971, ISO13485 and IEC62304. See also EU AI Act

### Data Protection

Compliance with UK GDPR and the Data Protection Act 2018, e.g., comprehensive Data Protection Impact Assessments (DPIAs).

### Systems Integration

Ensuring interoperability and adherence to clinical coding standards including FHIR, SNOMED CT, and ICD-10

*this is not intended to be exhaustive

# What Next?

- We will compile your questions to help in the preparation of our next webinar
- The event will be posted on the BCS site and Eventbrite

# Useful Links

1. ChatGPT
2. Introduction to Large Language Models | Machine Learning | Google for Developers
3. Prompt Engineering for AI Guide | Google Cloud
4. GPT-4.1 Prompting Guide
5. Prompt Library - Anthropic

# References

**Slide 2 - Macro Challenges facing the NHS**

- NHS Budget Constraints
- NHS Vacancy Statistics
- Productivity Mandate
- Prime Minister's remarks on the NHS: 6 January 2025 - GOV.UK

**Slide 4 – Health Data Landscape**

- Privacy-preserving large language models for structured medical information retrieval | npj Digital Medicine