1 September 2025

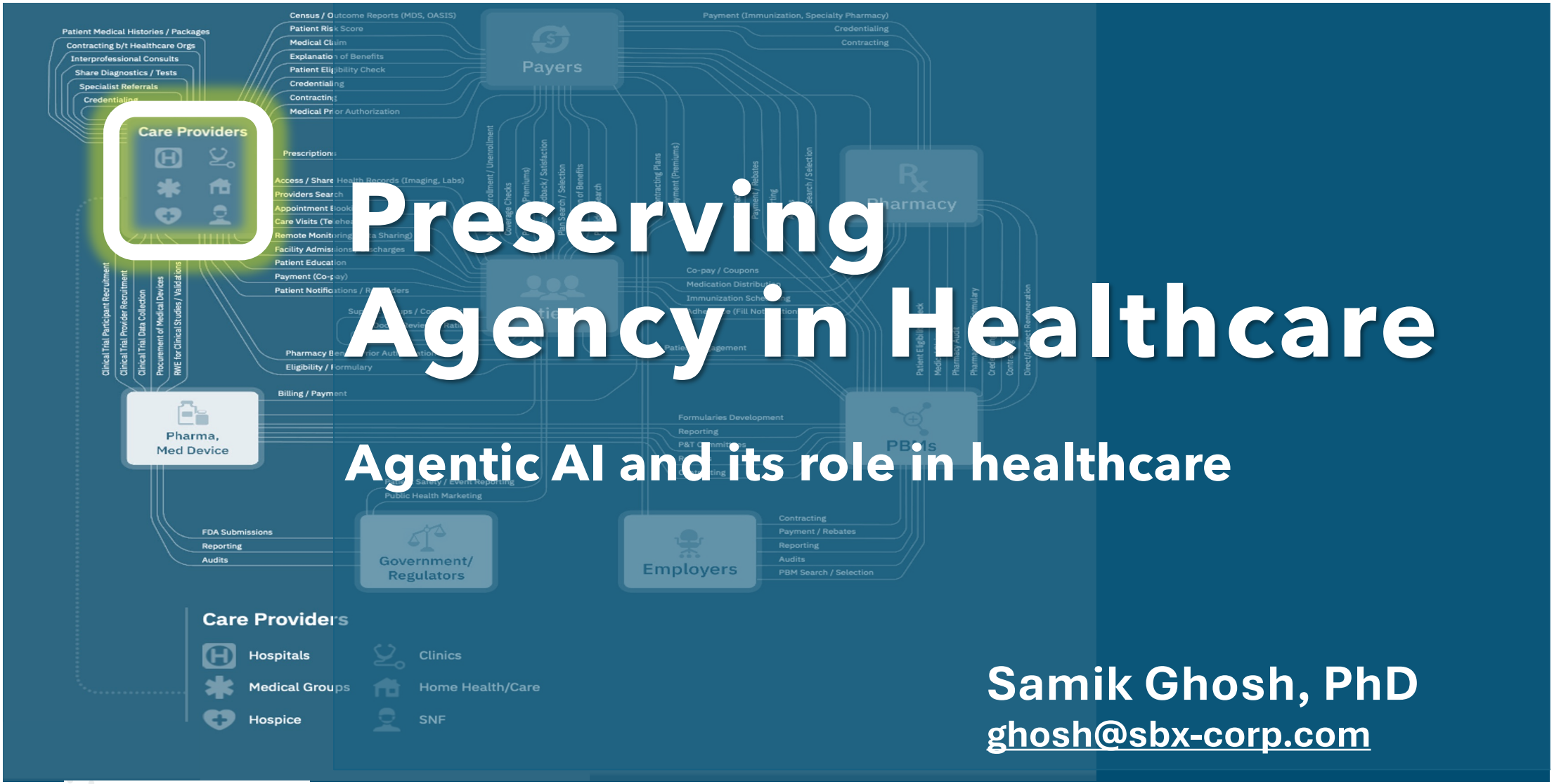
Introduction to Agentic AI and its role in Healthcare

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

Samik Ghosh, COO, SBX Corporation

Samik Ghosh's career spans science, engineering and entrepreneurship. He serves as COO at SBX Corporation Tokyo group, CEO of SBX Technologies, a SBX subsidiary company focused on healthcare and is a Senior Scientist at The Systems Biology Institute, Tokyo.

He holds a PhD in Computer Sciences from University of Texas and has diverse experience in life-science and healthcare industries in India, USA and Japan.



Preserving Agency in Healthcare

Agentic AI and its role in healthcare

Samik Ghosh, PhD
ghosh@sbx-corp.com

Disha
Platform for supply side of health

SBI The Systems Biology Institute | SBX | SBX Technologies | SBX BioSciences

sbi.jp
sbx-corp.com
sbx-tech.com
sbx-biosci.com

Agentic AI in Healthcare

Workshop

29 August

2025 A.D

A.D = Agent Dominance

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

Journey began 25 years ago in “systems thinking”

25 years of Innovation

- Over 2 decades in the business of Biology
- Category creator in “**Systems Biology**”
- Pioneer at the intersection of **biology, data & compute**
- Bootstrapped, profitable and stable organic growth over a decade

The Systems Biology Institute (SBI), Tokyo, Japan is **a non-profit private research institution established in 2000** to promote systems biology research and its application to medicine and global sustainability. SBI focuses on the rapid translation of basic research to practical outcomes for both business and clinical applications. SBI has been engaged in a series of research programs funded by government grants, both national and international as well as extensive private contracts for industrial applications of systems biology research.

SBX Corporation, Japan (SBX Group and its subsidiaries) **is a tech.bio spin-off from SBI, Tokyo, established in 2011**, focused on digital biology products and services to enable drug discovery and healthcare. Driven by deep science, powerful platforms, and precision engineering, the diverse team of SBX spread across Japan, North America, and India harnesses the power of data and intelligence to provide bespoke solutions for global companies in drug discovery, health, and wellbeing.

Global Talent Network

Building agents for craftmanship

POWERED BY GARUDA

Connectivity and automation platform

Powered by Garuda™ is a technology framework, enabled by the Garuda platform, which provides connectivity and automation of diverse data, devices and analytics to create novel pipelines, products and solutions in various science and industry verticals in platform-agnostic manner.


STUDY DESIGNS

Software for systems biology: from tools to integrated platforms

Samik Ghosh*, Yukiko Matsuoka*, Yoshiyuki Asai, Kun-Yi Hsin and Hiroaki Kitano*

Abstract | Understanding complex biological systems requires extensive support from software tools. Such tools are needed at each step of a systems biology computational workflow, which typically consists of data handling, network inference, deep curation,

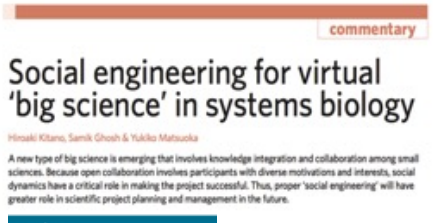
nature REVIEWS GENETICS

Nov **2011**
Ghosh et al.

POWERED BY TAXILA

Text analytics and intelligence platform

Taxila is a living system which grows and learns by forming new connections from an ever-accumulated knowledge.
Powered by Taxila™ is a technology framework, enabled by the Taxila platform, which provides automatic context-aware aggregation and search of relevant information, driven by AI-powered mining and analytics engine for driving actionable insights with an intuitive user interface to create new and own insights on the topic and resource of your interest.


commentary

Social engineering for virtual 'big science' in systems biology

Hiroaki Kitano, Samik Ghosh & Yukiko Matsuoka

A new type of big science is emerging that involves knowledge integration and collaboration among small sciences. Because open collaboration involves participants with diverse motivations and interests, social dynamics have a critical role in making the project successful. Thus, proper 'social engineering' will have greater role in scientific project planning and management in the future.

nature chemical biology

May **2011**
Kitano, Ghosh, Matsuoka

POWERED BY Gandhara

Biomedicine focused intelligence platform

Gandhara provides Wisdom of the Crowd AI approach which integrates a large number of machine learning and deep learning algorithms for big data analytics.


The Systems Biology Institute (SBI) participated in the AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge, and ranked 4th amongst 142 global submissions

The AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge
(Registration Open: 25 June 2015, Close: 21 March 2016)
https://www.synapse.org/#!Synapse:syn4231880/wiki/

Nat Commun. 2019 Jun 17;10(1):2674.
https://www.nature.com/articles/s41467-019-09799-2

The AstraZeneca-Sanger Drug Combination Prediction Challenge

Public domain machine learning challenge to predict drug synergy from multi-dimensional big data (2015-2016)

Harnessing Intelligence in various Shades & Size

Shades of AI

Rule-based Systems

Rule-based, rigid and template-driven. Focus on automation (RPA) for efficiency.
Example: Rule-based report generation and checking

Static AI Systems

Systems that employ learning models on static datasets and solve specific problems
Example: Supervised learning for classification of documents,

Dynamic AI Systems

Systems which are trained on static datasets for specific problem but with capability to augment their learning from new data in dynamic manner.
Example: Risk assessment of compounds based on existing data with the ability to learn from novel compound information

Generative AI systems

Pre-trained large-scale models (language models) trained on vast datasets with the ability to generate output based on specific prompts

Large Language Models for Pathway Curation: A Preliminary Investigation

Nikitha Karkera[1], Nikshita Karkera[1], Mahanash Kumar[1], Samik Ghosh[1,2], and Sucheendra K. Palaniappan[1,2,✉]

[1]The Systems Biology Institute, Tokyo, Japan
[2]SBX Corporation, Tokyo, Japan

Keypoints/Objectives

- Study evaluates the feasibility of using Large Language Models (LLMs) for pathway curation in scientific literature.
- Using GPT-3.5 based fine tuned models for pathway curation, we compare its performance with existing methods, focusing on precision, recall and F1 score metrics.
- Emphasize the promise and need for further research on using LLMs for pathway curation.

The pathway curation task involves analyzing scientific literature to identify and represent cellular processes as pathways. This process, often time-consuming and labor-intensive, requires significant curation efforts amidst the rapidly growing biomedical literature. Natural Language Processing (NLP) offers a promising method to automatically extract these interactions from scientific texts. Despite immense progress, there remains room for improvement in these systems. The emergence of Large Language Models (LLMs) provides a promising solution for this challenge. Our study conducts a preliminary investigation into leveraging LLMs for the pathway curation task. This paper first presents a review of the current state-of-the-art algorithms for the pathway curation task. Our objective is to check the feasibility and formulate strategies of using these LLMs to improve the accuracy of pathway curation task. Our experiments demonstrate that our GPT-3.5 based fine-tuned models outperform existing state-of-the-art methods. Specifically, our model achieved a 10 basis point improvement in overall recall and F1 score compared to the best-existing algorithms. These findings highlight the potential of LLMs in pathway curation tasks, warranting further research and substantial efforts in this direction.

Pathway Curation | LLM | gpt4 | gpt3.5 | Generative AI

Correspondence: sucheendra@sbi.jp

Introduction

rating these pathways requires integration of data from various sources, including experimental studies and scientific literature. Researchers often use these information to carefully assemble information about genes, proteins, and small molecules, and other relevant players along with their roles and relationships within the intricate web of cellular signalling and Regulation (2, 3). Significant efforts have been made in the detailed mapping of process or disease-specific pathways. Recent notable efforts include a comprehensive map of lipid mediator metabolic pathways(4), Alzheimer's disease (5), Parkinson's disease (6), and the influenza A virus replication cycle (7). Human curation, while accurate, is extremely time-consuming and cannot keep pace with the increase in the number of publications (8). There is a need to move towards automation of the extraction and interpretation which will significantly speed up the process and reduces the likelihood of human error or oversight. This not only make the process more efficient but also allows for the timely updating and expanding of pathway databases.

The goal of the Pathway Curation task is to using natural language processing (NLP) techniques to automatically extract and interpret information about these processes directly from scientific literature. The task that was defined in the BioNLP Shared Task 2013 (9), focuses on evaluating the applicability of event extraction systems to assist in the curation, evaluation, and maintenance of bio-molecular pathway models. Event extraction in our context refers to the process of identifying and characterizing biological events, such as gene expression, phosphorylation, complex formation from text. It further comprises of the following sub problems: 1) identifying relevant entities (like genes or proteins), 2) determining the relationships or interactions between these entities and 3) classifying these interactions into predefined categories. Over the years, many algorithms and platforms have been proposed to tackle the problem with varying degree of success (10)

The primary objective of this study is to investigate the effectiveness of using Large Language Models (LLMs) for the pathway curation task. Utilizing the sophisticated natural language understanding and generative capabilities of LLMs, we

SLAS Technology

Contents lists available at ScienceDirect

SLAS Technology

Journal homepage: www.elsevier.com/locate/slast

ProtoCode: Leveraging large language models (LLMs) for automated generation of machine-readable PCR protocols from scientific publications

Shuo Jiang[a,1], Daniel Evans-Yamamoto[b,1], Dennis Bersenev[a], Sucheendra K. Palaniappan[b,*], Ayako Yachie-Kinoshita[a,b,*]

ARTICLE INFO

ABSTRACT

1. Introduction

data, analysis scripts, and protocols have become common practices, the reusability of protocols remains mostly limited to human use. The extension of automated machine reusability on protocols would open the gate to exploit the rich literature for applications such as lab automation.

Protocols are typically described in the form of text in scientific publications. Some peer-reviewed journals require the authors to describe the resources and methods in a standardized format to ensure that all necessary information is available to reproduce the protocol [1]. Journals dedicated to protocols and repositories such as protocol.io (https://www.protocols.io/) provide a platform to describe protocols in detail with additional notes. Videos are also used to visually aid researchers to replicate specific techniques (https://www.jove.com/). While these standardization and guidelines increase reproducibility and reliability of the described protocol, there remains some limitations. The first limitation is that the current models/guidelines have large variability that even the same procedures in two protocols can be described

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.08.588614; this version posted April 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Link Prediction for Hypothesis Generation: An Active Curriculum Learning Infused Temporal Graph-Based Approach

Uchenna Akujuobi[1*†], Priyadarshini Kumari[2*†], Jihun A. Choi[1], Samy Badreddine[3], Kana Maruyama[1], Sucheendra K. Palaniappan[4], Tarek R. Besold[3]

[1]Sony AI, Tokyo, Japan.
[2]Sony AI, Cupertino, USA.
[3]Sony AI, Barcelona, Spain.
[4]The Systems Biology Institute, Tokyo, Japan.

*Corresponding author(s). E-mail(s): uchenna.akujuobi@sony.com; priyadarshini.kumari@sony.com;
Contributing authors: jihun.a.choi@sony.com; samy.badreddine@sony.com; kana.maruyama@sony.com; sucheendra@sbi.jp; tarek.besold@sony.com;
†These authors contributed equally to this work.

Abstract

Over the last few years Literature-based Discovery (LBD) has regained popularity as a means to enhance the scientific research process. The resurgent interest has spurred the development of supervised and semi-supervised machine learning models aimed at making previously implicit connections between scientific concepts/entities explicit based on often extensive repositories of published literature. Understanding the temporally evolving interactions between these entities can provide valuable information for predicting the future development of entity relationships. However, existing methods often underutilize the latent information embedded in the temporal aspects of interaction data.
In this context, motivated by applications in the food domain—where we aim to connect nutritional information with health-related benefits—we address the

Replicating a High-Impact Scientific Publication Using Systems of Large Language Models

Dennis Bersenev[1], Ayako Yachie-Kinoshita[1,2], Sucheendra K. Palaniappan[1,2,*]

[1]SBX BioSciences, Inc., 1600 - 925 West Georgia Street, Vancouver, V6C 3L2, BC, Canada.
[2]The Systems Biology Institute, Saisei Ikedayama Bldg., 5-10-25, Higashi Gotanda, Shinagawa-ku, 141-0022, Tokyo, Japan.
*Corresponding author.

Contributing authors: bersenevdennis@sbx-bioci.com; yachie@sbx-bioci.com; sucheendra@sbi.jp;

Abstract

Publications focused on scientific discoveries derived from analyzing large biological datasets typically follow the cycle of hypothesis generation, experimentation, and data interpretation. The reproduction of findings from such papers is crucial for confirming the validity of the scientific, statistical, and computational methods employed in the study, and it also facilitates the foundation for new research. By employing a multi-agent system composed of Large Language Models (LLMs), including both text and code generation agents built on OpenAI's platform, our study attempts to reproduce the methodology and findings of a high-impact publication that investigated the expression of viral-entry-associated genes using single-cell RNA sequencing (scRNA-seq). The LLM system was critically evaluated against the analysis results from the original study, highlighting the system's ability to perform simple statistical analysis tasks and literature reviews to establish the purpose of the analyses. However, we also identified significant challenges in the system, such as nondeterminism in code generation, difficulties in data procurement, and the limitations presented by context length and bias from the model's inherent training data. By addressing these challenges and expanding on the system's capabilities, we intend to contribute to the goal of automating scientific research for efficiency, reproducibility, and transparency, and to drive the discussion on the role of AI in scientific discovery.

npj | systems biology and applications

www.nature.com/npjsba

TECHNOLOGY FEATURE OPEN

Asking the right questions for mutagenicity prediction from BioMedical text

Sathwik Acharya[1], Nicolas K. Shinada[1,2], Naoki Koyama[3], Megumi Ikemori[4], Tomoki Nishioka[5], Seiji Hitaoka[5], Atsushi Hakura[5], Shoji Asakura[5], Yukiko Matsuoka[1,2] and Sucheendra K. Palaniappan[1,2]

Assessing the mutagenicity of chemicals is an essential task in the drug development process. Usually, databases and other structured sources for AMES mutagenicity exist, which have been carefully and laboriously curated from scientific publications. As knowledge accumulates over time, updating these databases is always an overhead and impractical. In this paper, we first propose the problem of predicting the mutagenicity of chemicals from textual information in scientific publications. More simply, given a chemical and evidence in the natural language from publications where the mutagenicity of the chemical is described, the goal of the model/algorithm is to predict if it is potentially mutagenic or not. For this, we first construct a golden standard data set and then propose MusaPreBERT, a prediction model fine-tuned on BioLinkBERT based on a question-answering formulation of the problem. We leverage transfer learning and use the help of large transformer-based models to achieve a Macro F1 score of >0.88 even with relatively small data for fine-tuning. Our work establishes the utility of large language models for the construction of structured sources of knowledge bases directly from scientific publications.

npj Systems Biology and Applications (2023)9(1) ; https://doi.org/10.1038/s41540-023-00324-2

INTRODUCTION

Mutagenicity assessment of chemicals (drug products) is an important step in drug development process. Usually, the mutagenicity potential of chemical substances is assessed using in vitro gene mutation assays such as Ames test[1] and the mouse lymphoma TK mutation assay (MLA). Specifically, the Ames test is a bacterial reverse mutation test that uses various chemical strains (usually strains of four different species of bacteria (Salmonella typhimurium) such as TA98, TA100, TA102, etc., and Escherichia coli WP2 uvrA and WP2 uvrA/pKM101, and examining if they result in changes in the DNA of the organism (base-pair substitutions, frameshifts, insertions, and deletions). A positive test indicates that a chemical is mutagenic. Usually, regulatory guidelines such as those by OECD, and International Conference on Harmonization (ICH) guidelines are followed for genotoxicity testing methods that help evaluate the effect of chemical substances on human health. Also, ICH M7 guidelines for the assessment and control of mutagenic impurities in pharmaceuticals require mutagenicity tests to be an important prerequisite for regulatory approvals.

Owing to the ever-expanding expertise of chemicals/virtual libraries and practical issues in keeping pace with their growth combined with the difficulty of isolating pure forms of chemicals (without impurities), regulatory authorities/researchers often turn to in silico methods such as Quantitative Structure-Activity Relationship (QSAR) [2] for mutagenicity assessment of chemicals. While these methods do not rely on conducting (the often time-consuming and expensive) experimental tests such as Ames, they rely on previous Ames mutagenicity results and chemical structure properties of the chemical for predicting the mutagenic potential of chemicals. Computational and machine learning models have played a key role in advancing this discipline[3–6] by focusing on understanding the chemical, structural, and functional properties of chemicals that contribute to mutagenicity.

Most of the QSAR based tools often look up previously known mutagenicity results of chemicals which in turn are obtained from previously published scientific literature. Often, scientific publications in this field discuss the mutagenicity of chemical(s) in Ames assay in natural language. Consequently, an expert would comprehend the publications and compile them into data repositories that are used in the QSAR softwares. Keeping such information updated as more publications get added is a challenge in addition to manual effort in comprehending and making such structured data repositories. Machine learning models, especially from the disciple of natural language processing (NLP) can help in this regard. In fact NLP has seen several breakthroughs in the last few years and can be optimally adapted and leveraged for applications in bio-medicine[7,8].

With this background, we first formulate the problem of predicting the mutagenicity of a chemical given the text that describes its mutagenicity as an NLP task. Given the lack of prior training data/ benchmarks, we first start by creating a golden standard data set (GSD). We then propose a machine learning pipeline for training a model that can predict the mutagenicity of a chemical given the text which describes it. To do so, we have proposed a question-answering model based on transfer learning on large language models. These large language models are often based on transformers pretrained on large-scale natural text. We tried multiple language models before getting the best results on the model trained on the Stanford BioLinkBERTlsoc model[9], a model pre-trained on large-scale scientific text from sources such as PubMed. Our final trained model, named MusaPreBERT can achieve an accuracy of >88 percent with a macro average F1 score of >0.88.

Our final goal with this study is for machine learning models to be able to continuously update the mutagenicity results from publications into high-quality data sources. We also hope that fellow researchers would leverage on the generated data set and

Published in partnership with the Systems Biology Institute

The Journal of Toxicological Sciences (J. Toxicol. Sci.) Vol.49, No.3, 105–115, 2024

105

Original Article

DTox: A deep neural network-based *in visio* lens for large scale toxicogenomics data

Takeshi Hase[1,2,3,4,5], Samik Ghosh[1], Ken-ichi Aisaki[6], Satoshi Kitajima[6], Jun Kanno[1,6,7], Hiroaki Kitano[1,2] and Ayako Yachie[1,2]

[1]The Systems Biology Institute, Saisei Ikedayama Bldg., 5-10-25 Higashi Gotanda Shinagawa, Tokyo, 141-0022, Japan
[2]SBX BioSciences, Inc, 1600 - 925 West Georgia Street, Vancouver, BC V6C 3L2, Canada
[3]Institute of Education, Tokyo Medical and Dental University, 2HC, M&D Tower, 1-5-45 Yushima, Bunkyo-ku, Tokyo, 113-8510, Japan
[4]Faculty of Pharmacy, Keio University, 1-5-30 Shibakoen, Minato-ku, Tokyo, 105-8512, Japan
[5]Center for Mathematical Modelling and Data Science, Osaka University, 1-3, Machikaneyama-cho, Toyonaka-shi, Osaka 560-8531, Japan
[6]Division of Cellular and Molecular Toxicology, Center for Biological Safety and Research (CBSR), National Institute of Health Sciences (NIHS), 3-25-26 Tonomachi, Kawasaki-ku, Kawasaki-shi, Kanagawa-ken, 210-9501, Japan
[7]Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba City, Ibaraki-ken, 305-8575, Japan
[8]Integrated Open Systems Unit, Okinawa Institute of Science and Technology (OIST), Onna-son, Okinawa, Japan

[Contributed by Jun Kanno]

(Received December 8, 2023; Accepted December 9, 2023)

ABSTRACT — With the advancement of large-scale omics technologies, particularly transcriptomics data sets on drug and treatment response repositories available in public domain, toxicogenomics has emerged as a key field in safety pharmacology and chemical risk assessment. Traditional statistics-based bioinformatics analysis poses challenges in its application across multidimensional toxicogenomic data, including administration time, dosage, and gene expression levels. Motivated by the visual inspection workflow of field experts to augment their efficiency of screening significant genes to derive meaningful insights, together with the ability of deep neural architectures to learn the image signals, we developed *DTox*, a deep neural network-based *in visio* approach. Using the Percellome toxicogenomics database, instead of utilizing the numerical gene expression values of the transcripts (gene probes of the microarray) for dose-time combinations, *DTox* learned the image representation of 3D surface plots of distinct time and dosage data points to train the classifier on the experts' labels of gene probe significance. *DTox* outperformed statistical threshold-based bioinformatics and machine learning approaches based on numerical expression values. This result shows the ability of image-driven neural networks to overcome the limitations of classical numeric value-based approaches. Further, by augmenting the model with explainability modules, our study showed the potential to reveal the visual analysis process of human experts in toxicogenomics through the model weights. While the current work demonstrates the application of the *DTox* model in toxicogenomic studies, it can be further generalized as an *in visio* approach for multi-dimensional numeric data with applications in various fields in medical data sciences.

Key words: Artificial intelligence, Deep learning, Toxicogenomics, Percellome

INTRODUCTION

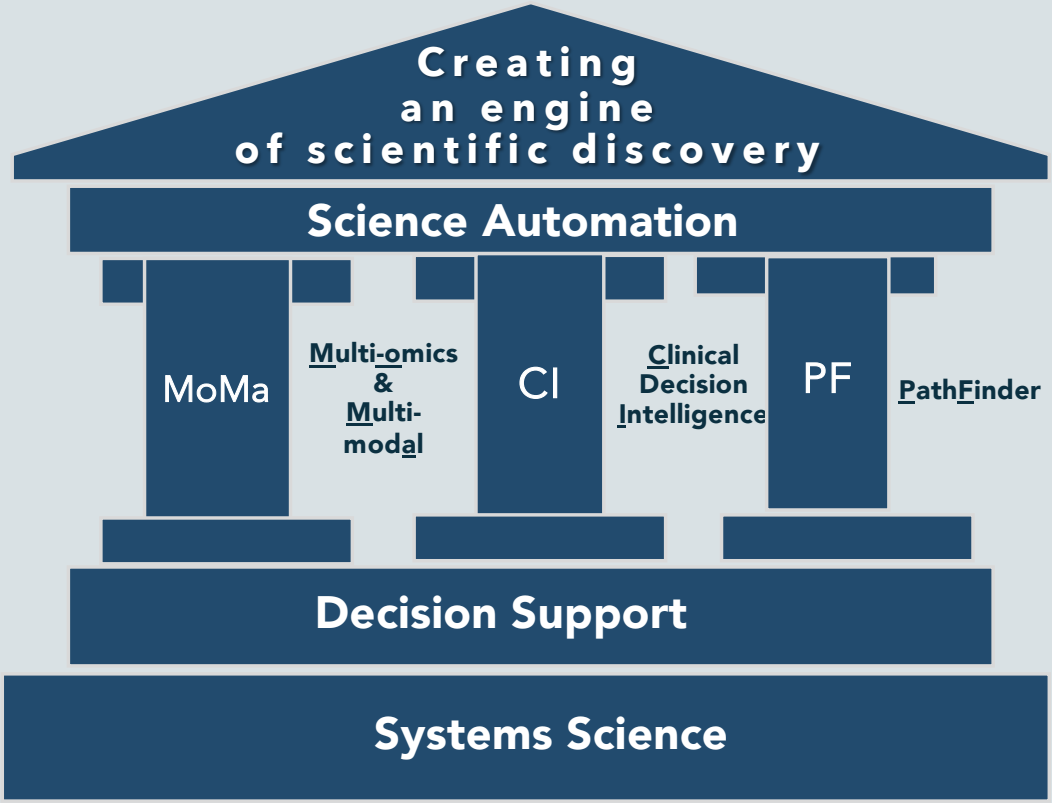
Toxicogenomics is the study of molecular events in the living system induced by the chemical agents entering the system. Its goal is to comprehensively elucidate the molecular mechanisms that lead to toxicological adverse events, and subsequently improve the diagnostics, treatment for known toxicants, and prediction of toxicity for

Correspondence: Ayako Yachie-Kinoshita (E-mail: yachie@sbi.jp)

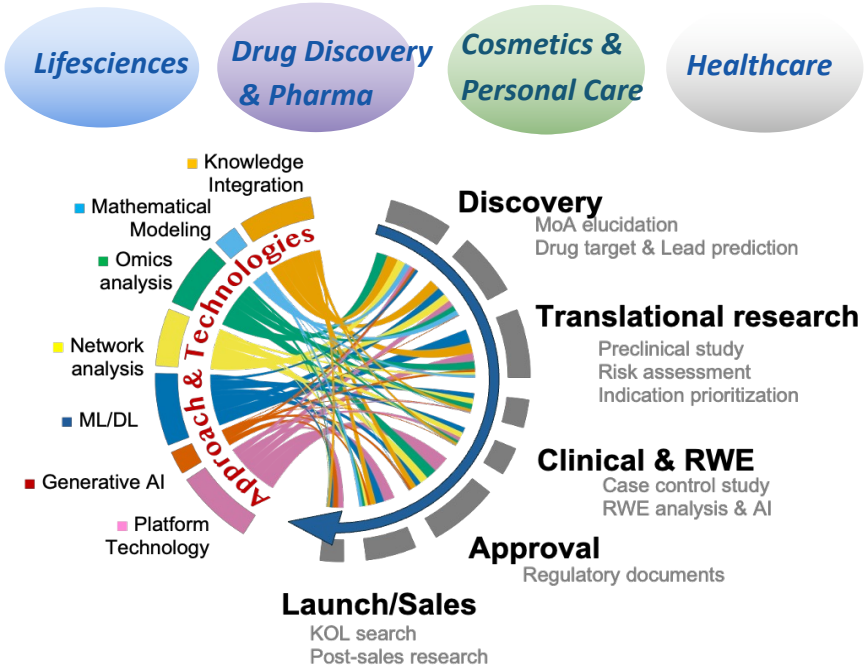
Vol. 49 No. 3

https://www.sbi.jp/publications

Bridging the gap from “cure” to “care”



Outcomes *not* Output



Over 35 Partners across across academia, pharma, healthcare and medical technologies

Top 3 Pharma in Japan

Fortune 500 Global companies including consumer health players

Complexity of Care

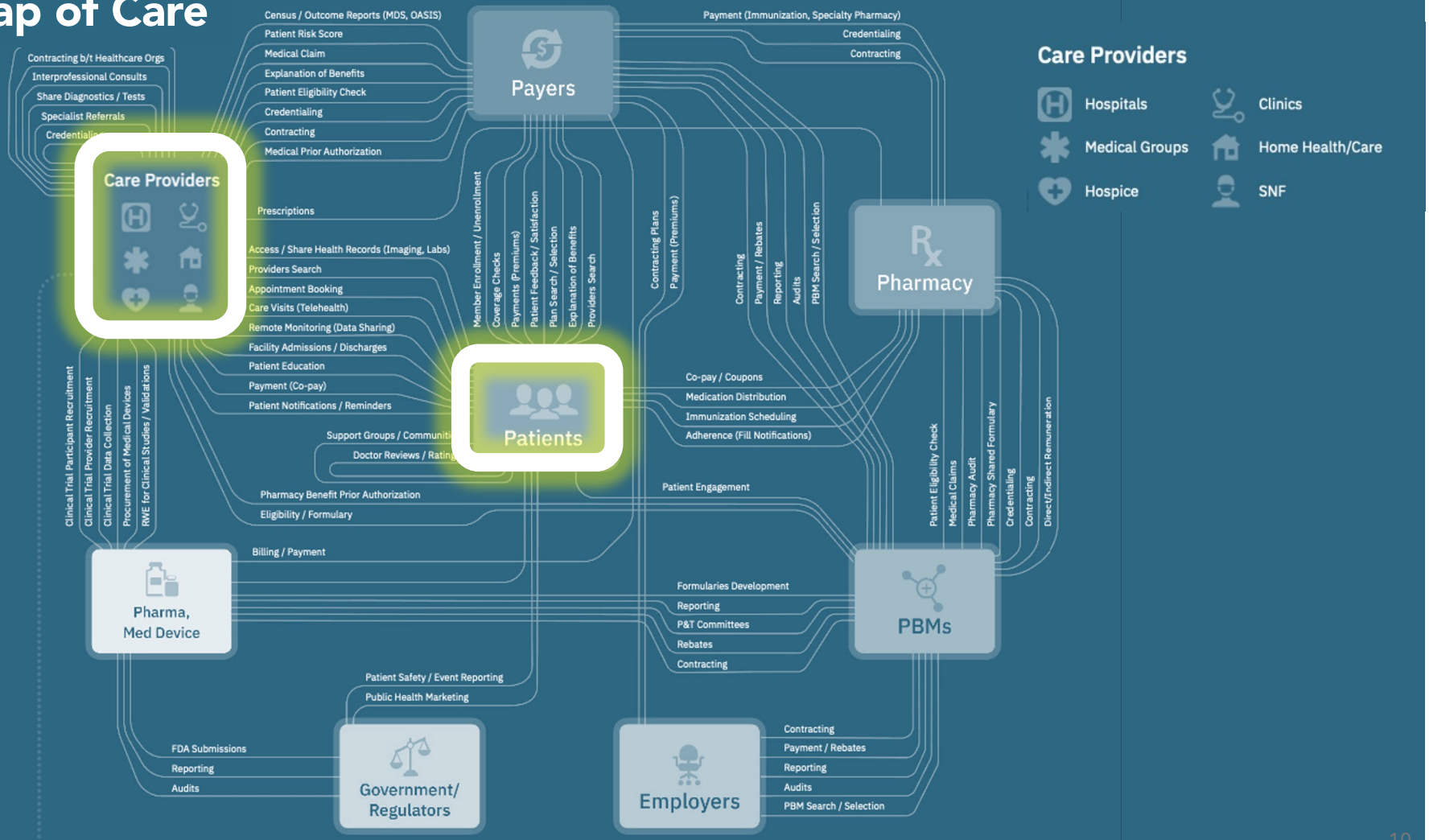
"The hospital [is] altogether the most complex human organization ever devised."

-- Peter Drucker

Peter F. Drucker. They're Not Employees, They're People,HBR Magazine 2002 Article
https://hbr.org/2002/02/theyre-not-employees-theyre-people

We need a “compass” & “map” to navigate

The Subway Map of Care



SUMMIT

Care is DUST(y) - Difficult, Uncertain, Scary & Tedious

Diagnostic laboratories

Analyze and provide accurate and identifiable data on samples to clinic

Clinic operations & support staff

Research Laboratory/ Pharma trials

- Conduct research on new formulations and drugs
- Perform clinical trials
- Collect and analyze data

Doctors

Serve patients better
Serve more patients

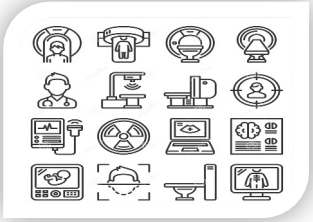
Patients

Get better and continuous advise and consultation on health and well being

Clinic

- Provide better healthcare outcomes to patients
- Improve operational efficiency and productivity
- Increase sales/ value to patients

Medical equipments

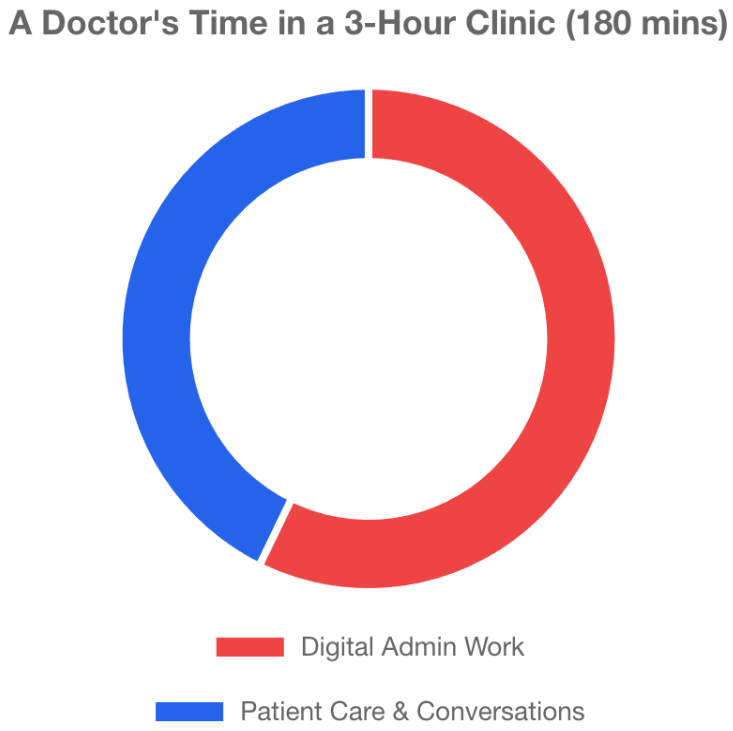


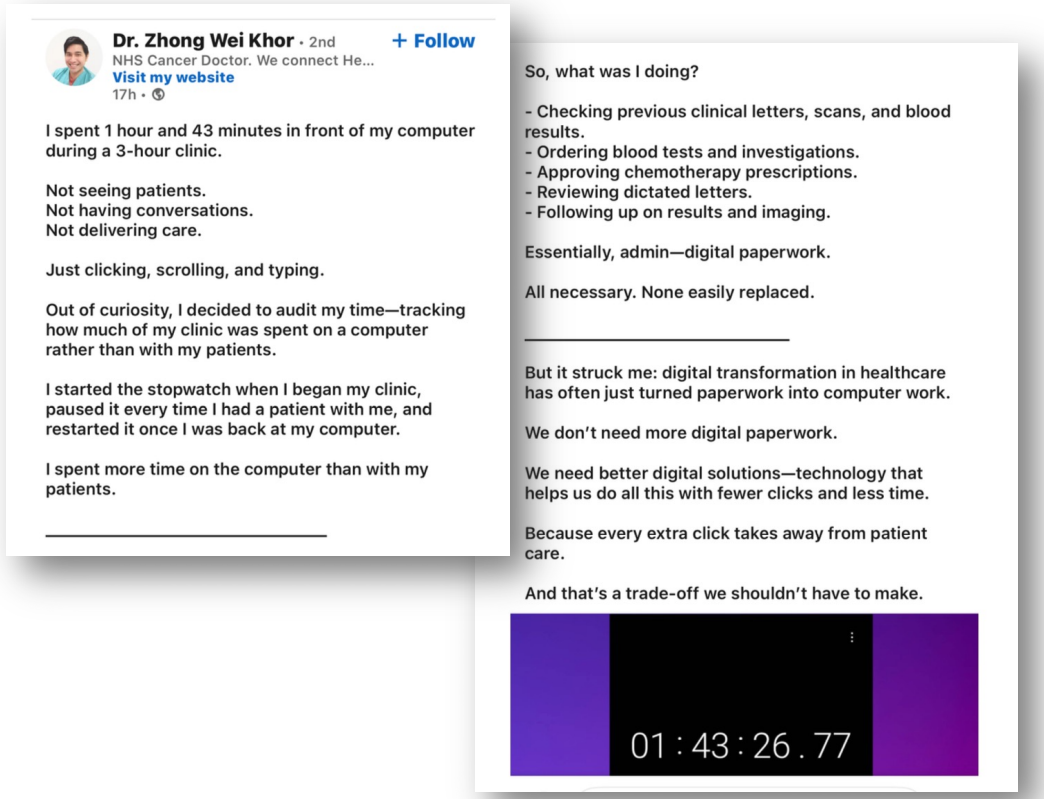
Happy Providers make Happy Patients

Death by a thousand “clicks” and “cuts” (of paper)

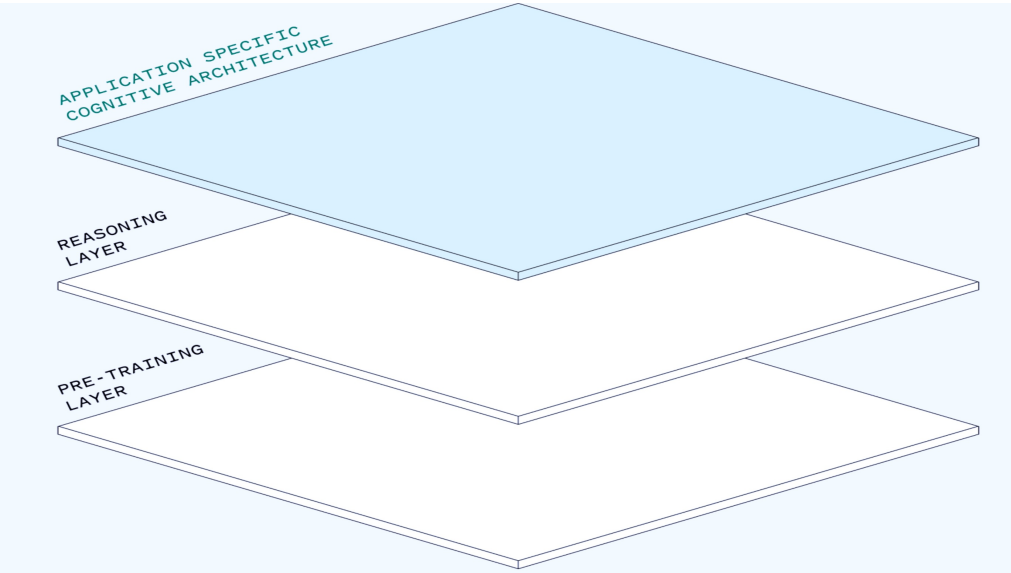
The Silent Cost of Clicks

In a typical 3-hour clinic session, a significant portion of a doctor's time is consumed by digital administrative tasks. This chart shows the breakdown, highlighting the time lost to clicks, scrolling, and typing instead of patient care.



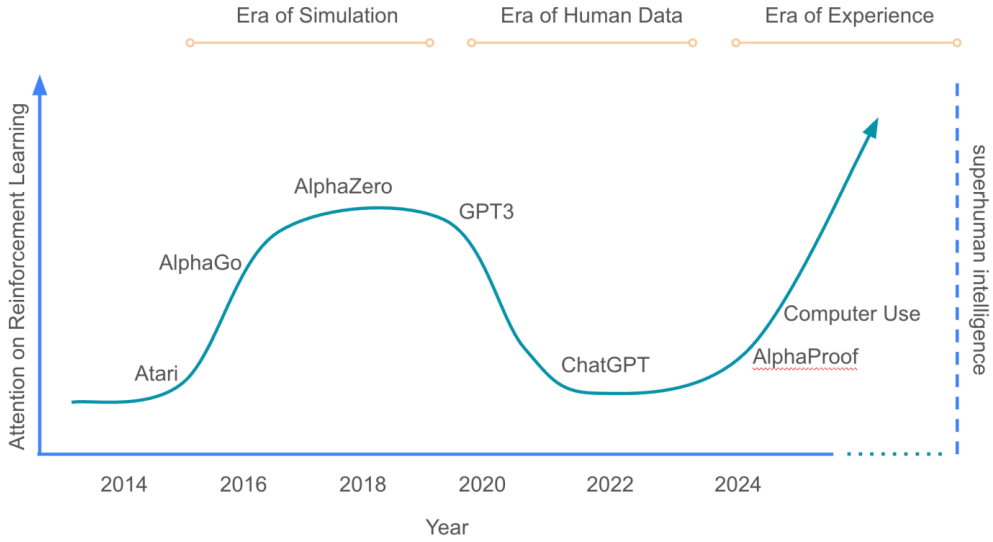


Standing on the threshold of a new Era



System 1 vs System 2 Thinking

- This leap from pre-trained instinctual responses (“System 1”) to deeper, deliberate reasoning (“System 2”) is the next frontier for AI.
- **It’s not enough for models to simply know things—they need to pause, evaluate and reason through decisions in real time.**



Welcome to the Era of Experience

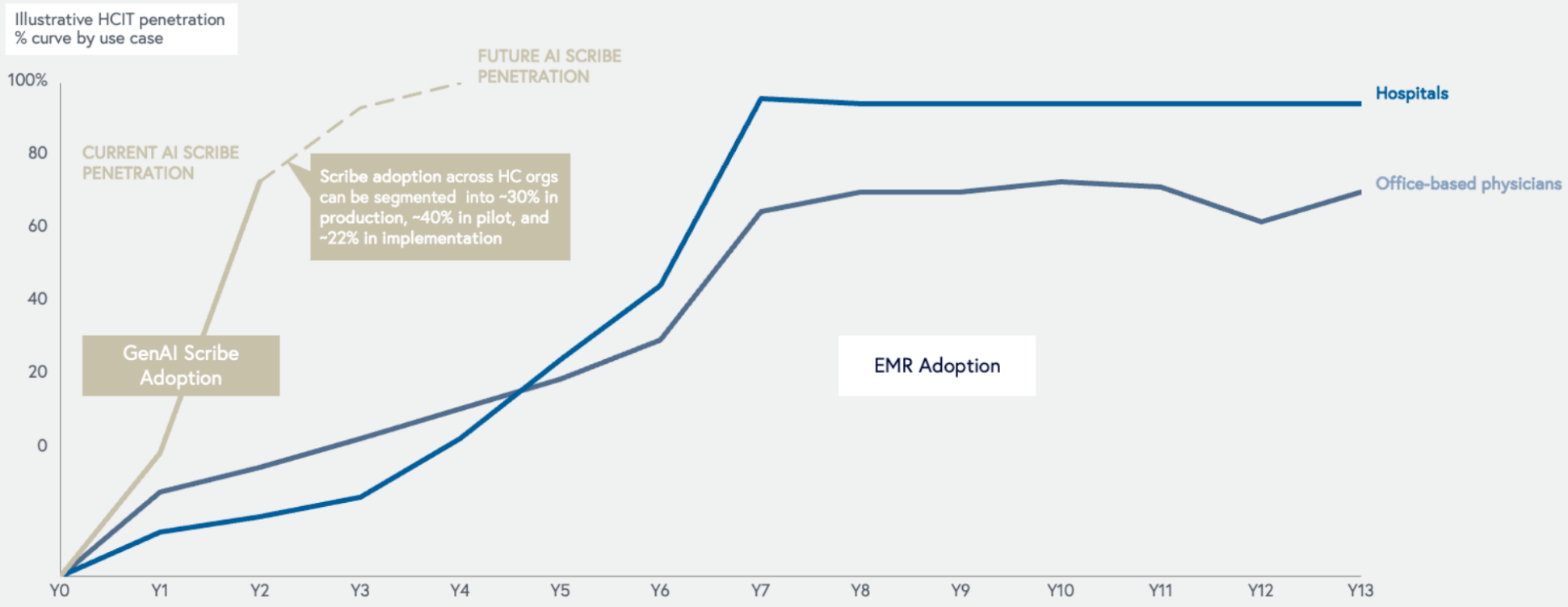
David Silver, Richard S. Sutton*

- We stand on the threshold of a new era in artificial intelligence that promises to achieve an unprecedented level of ability.
- A new generation of agents will acquire superhuman capabilities by learning predominantly from experience.

https://www.sequoiacap.com/article/generative-ais-act-o1/
https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf

With new opportunities in healthcare

How this AI wave differs from the prior software revolution



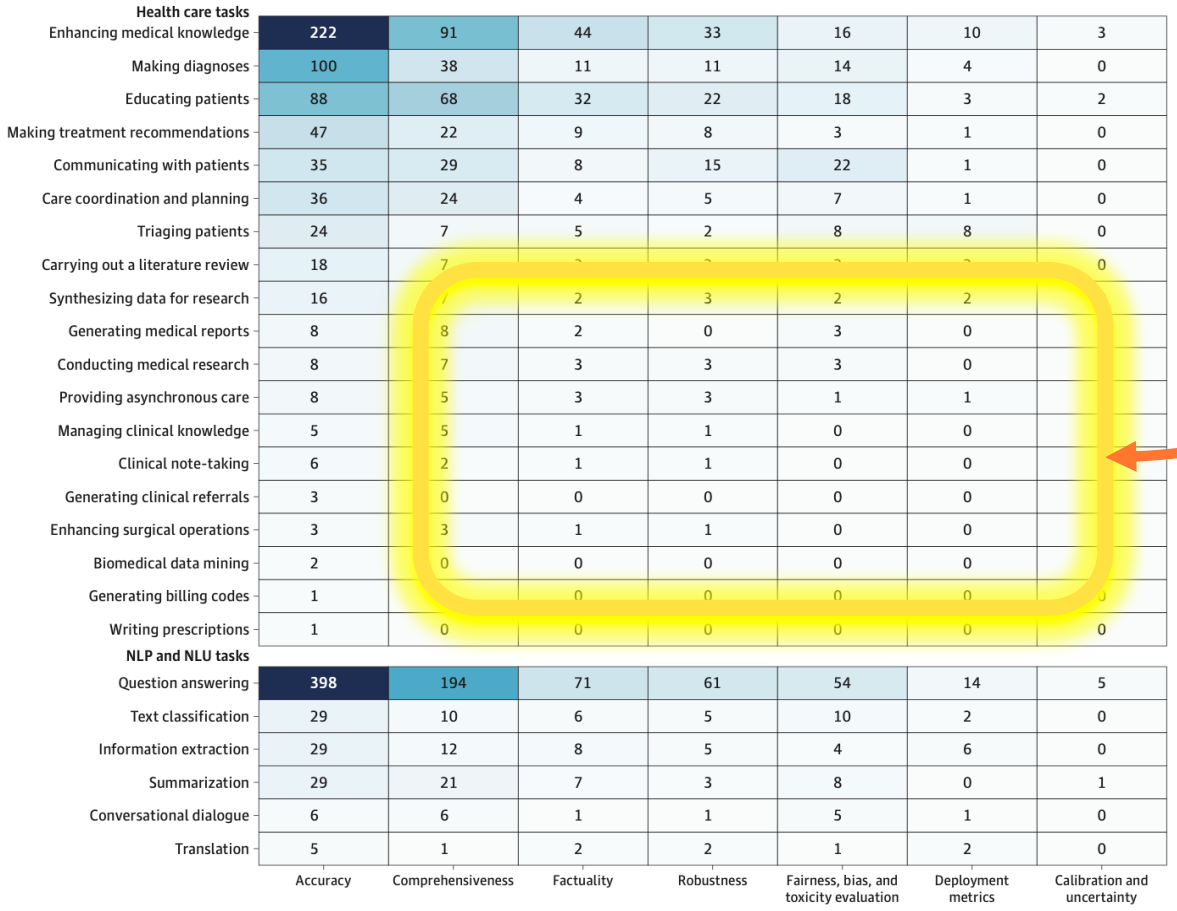
CURRENT WAVE: Unprecedented interest to try and buy solutions

PREVIOUS WAVE: Transition from pen-and-paper to software, requiring cultural shifts and regulatory catalyst

Source: Bain IP; Office of the National Coordinator for Health IT; Market participant interviews; Bain GenAI Survey (N = 408)

To color the white space of operations

Heat Map of Health Care Tasks, Natural Language Processing (NLP) and Natural Language Understanding (NLU) Tasks, and Dimensions of Evaluation Across 519 Studies



White space opportunity to build agency

Original Investigation | AI in Medicine

October 15, 2024

Testing and Evaluation of Health Care Applications of Large Language Models

A Systematic Review

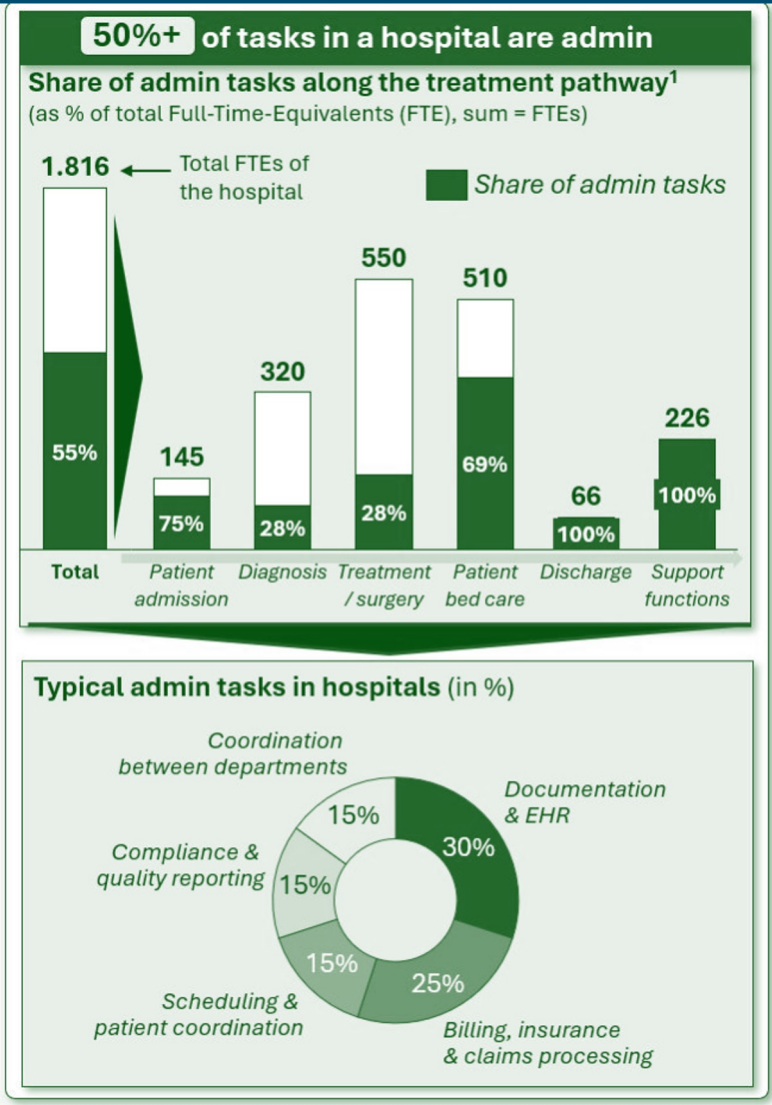
Suhana Bedi, BA[1]; Yutong Liu, MS[2]; Lucy Orr-Ewing, BA[3]; et al

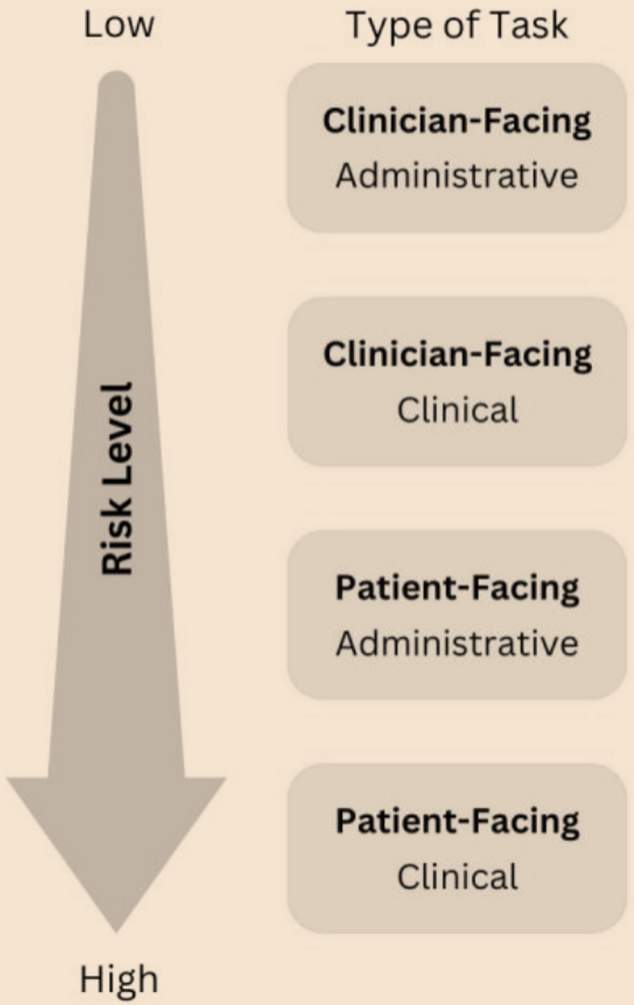
Author Affiliations | Article Information

JAMA. Published online October 15, 2024. doi:10.1001/jama.2024.21700

https://jamanetwork.com/journals/jama/fullarticle/2825147

Thoughtful balance of risk & rewards





Essence of Agents

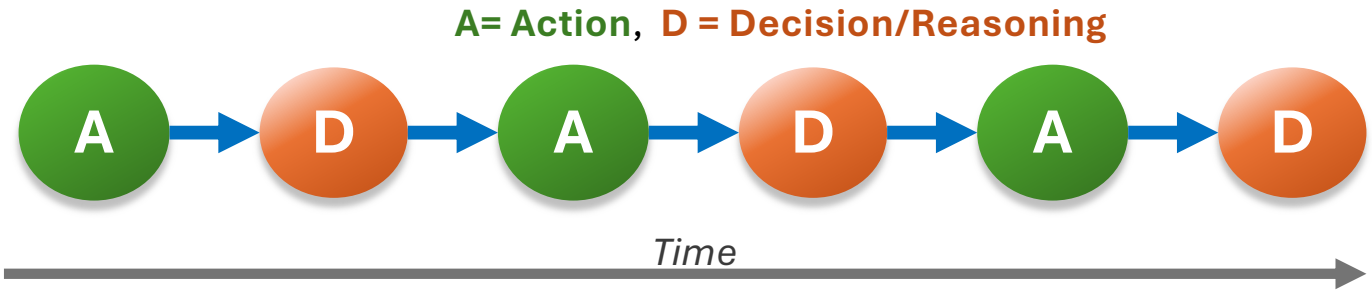
"**Agent**" can be defined in several ways

- As fully **autonomous systems** that operate independently over extended periods, using various tools to accomplish complex tasks.
- As **workflows** that describe more prescriptive implementations where LLMs and tools are orchestrated through predefined paths

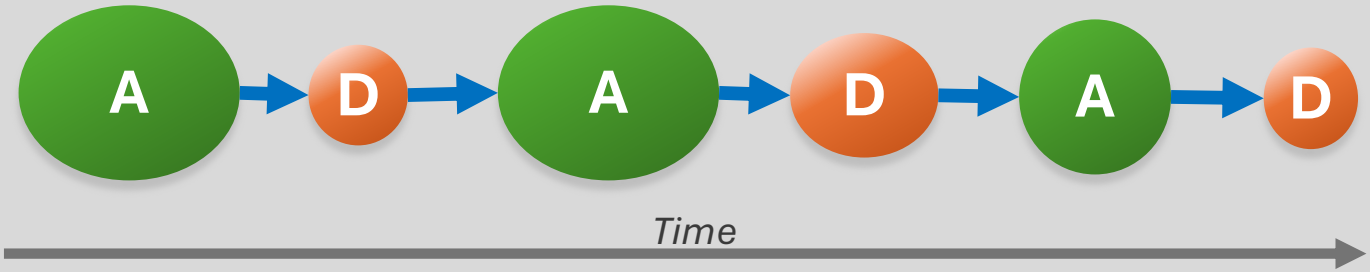
Agents are systems which provide and preserve agency to providers

Building effective agents, Engineering at Anthropic: Inside the team building reliable AI systems

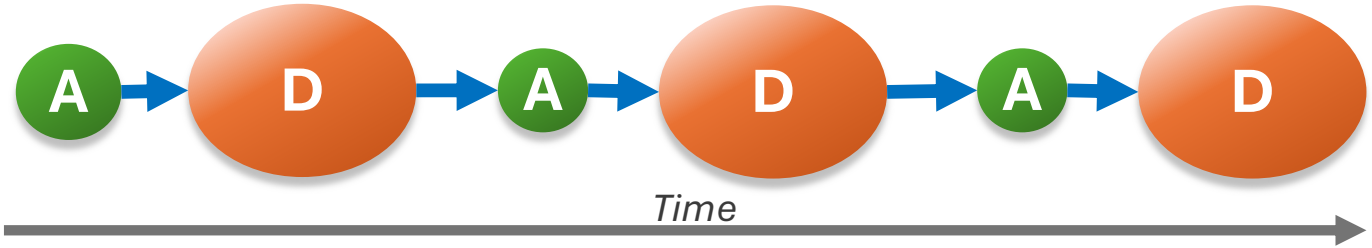
Anatomy of Agency



Most pathways (clinical, operational) are a sequence of actions (A) and decisions (D)



Digital tools tend to increase A and reduce D



Agentic systems should decrease A and increase D

Time to Think (Tik)
Time to Care (ToC)

What should agents NOT do to undermine Agency?

Don’t add WIMPs (Windows, Icons, Menu, Pointers)

- ✓ Agents are not apps
- ✓ No new screens, log-ins, forms and dashboards to navigate
- ✓ Interactions are designed to be natural and considerate, reducing digital friction.

Don’t break provider pathways

- ✓ Don’t add new actions in existing workflows
- ✓ Work in background mostly and surface when needed integrate into existing

Don’t increase the risk posture of care

- ✓ Only automate actions which are automatable
- ✓ Work in sand-boxed, tightly scoped functions

What should agents DO to augment Agency?

Clarity & Consistency

- ✓ Work on domain-limited, real-world problems
- ✓ Provide local intelligence
- ✓ Low resistance, high impact

Transparent Reasoning

- ✓ Feed-back and notifications
- ✓ Work in background mostly and surface when needed integrate into existing

Controlled Autonomy

- ✓ Define intent and reveal action
- ✓ Manage risk (autonomy slider)

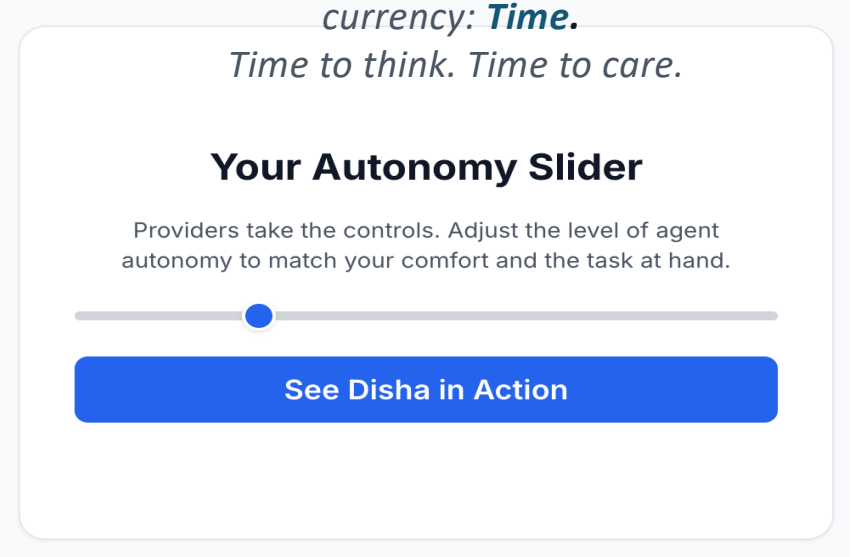
ゆとり

(Yutori)

[a Japanese concept of "giving space" or "breathing room."]

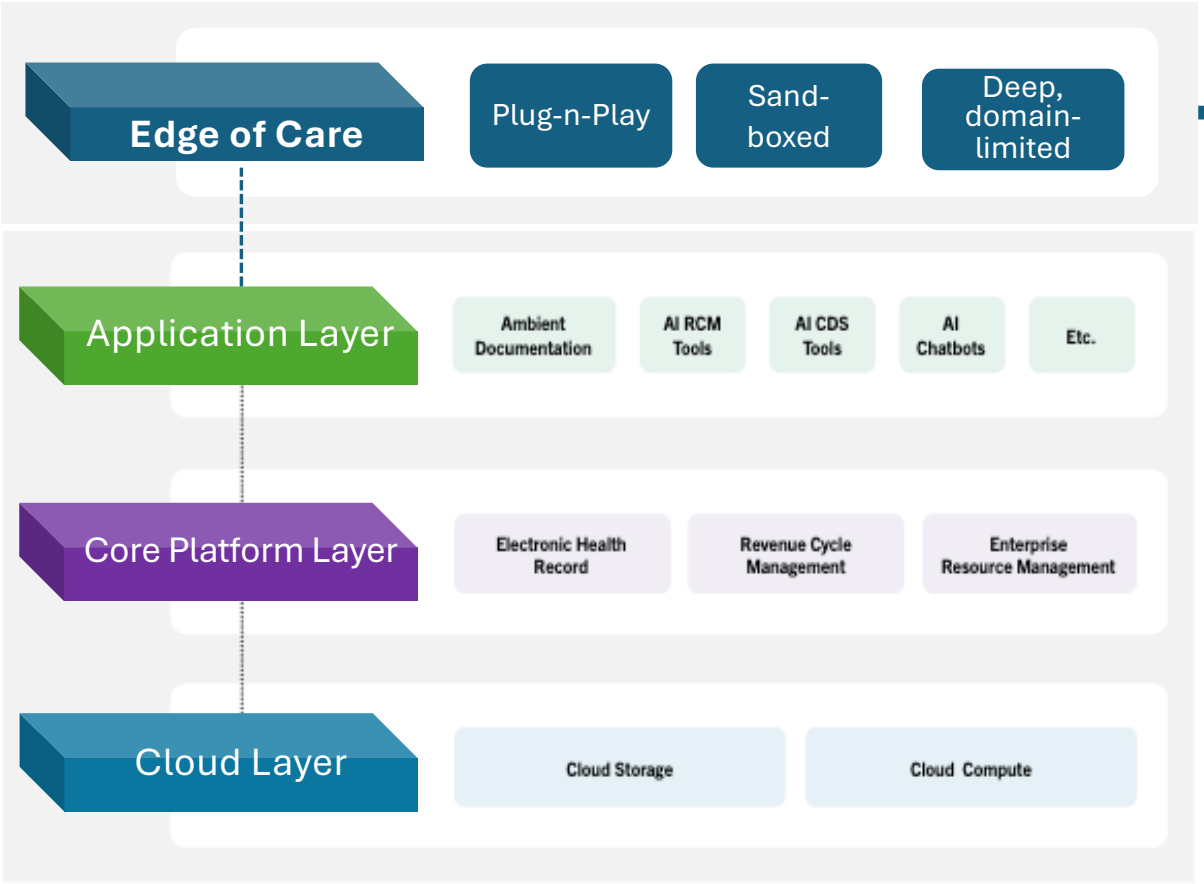
*giving provider their most valuable currency: **Time**.*

Time to think. Time to care.



How can we preserve agency – edge of care

"Mine the Gaps" at the edge of care



Advantage, Agents

- Low Latency
- Light Weight
- Orchestration
- Control, Audit & Resilience

https://phti.org/ai-adoption-early-applications-impacts/

How can we preserve agency – hyper-modal interface

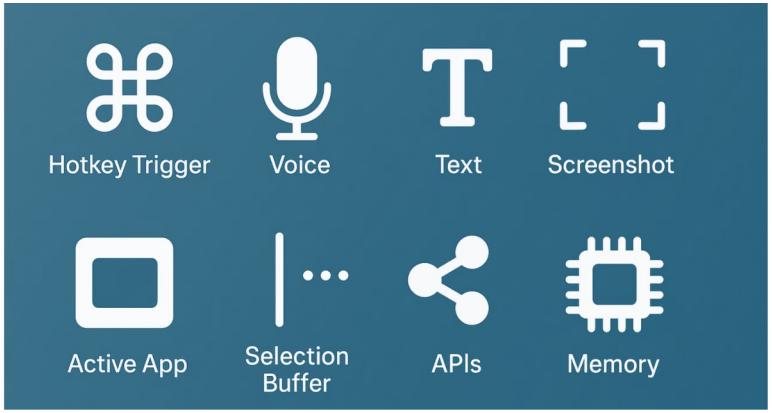
How Natural language performs on the speed dimension

Bottleneck

RECEIVE		SEND	
Read	~250wpm	Write	~60wpm
Listen	~270wpm	Speak	~150wpm

We form thoughts at 1,000-3,000 words per minute

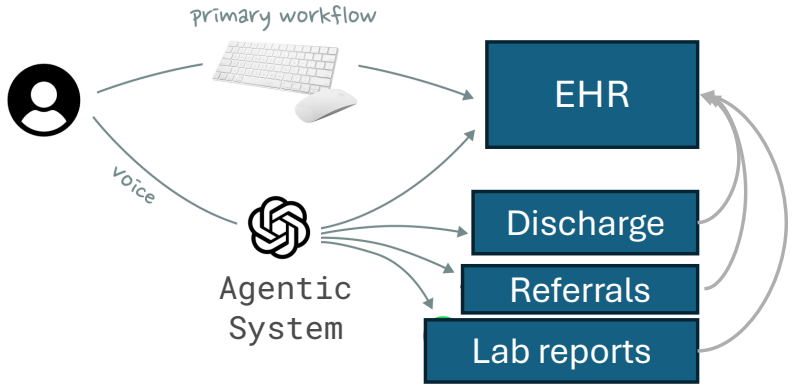
Agents need to interface on all modes



Humans use different modalities for I/O

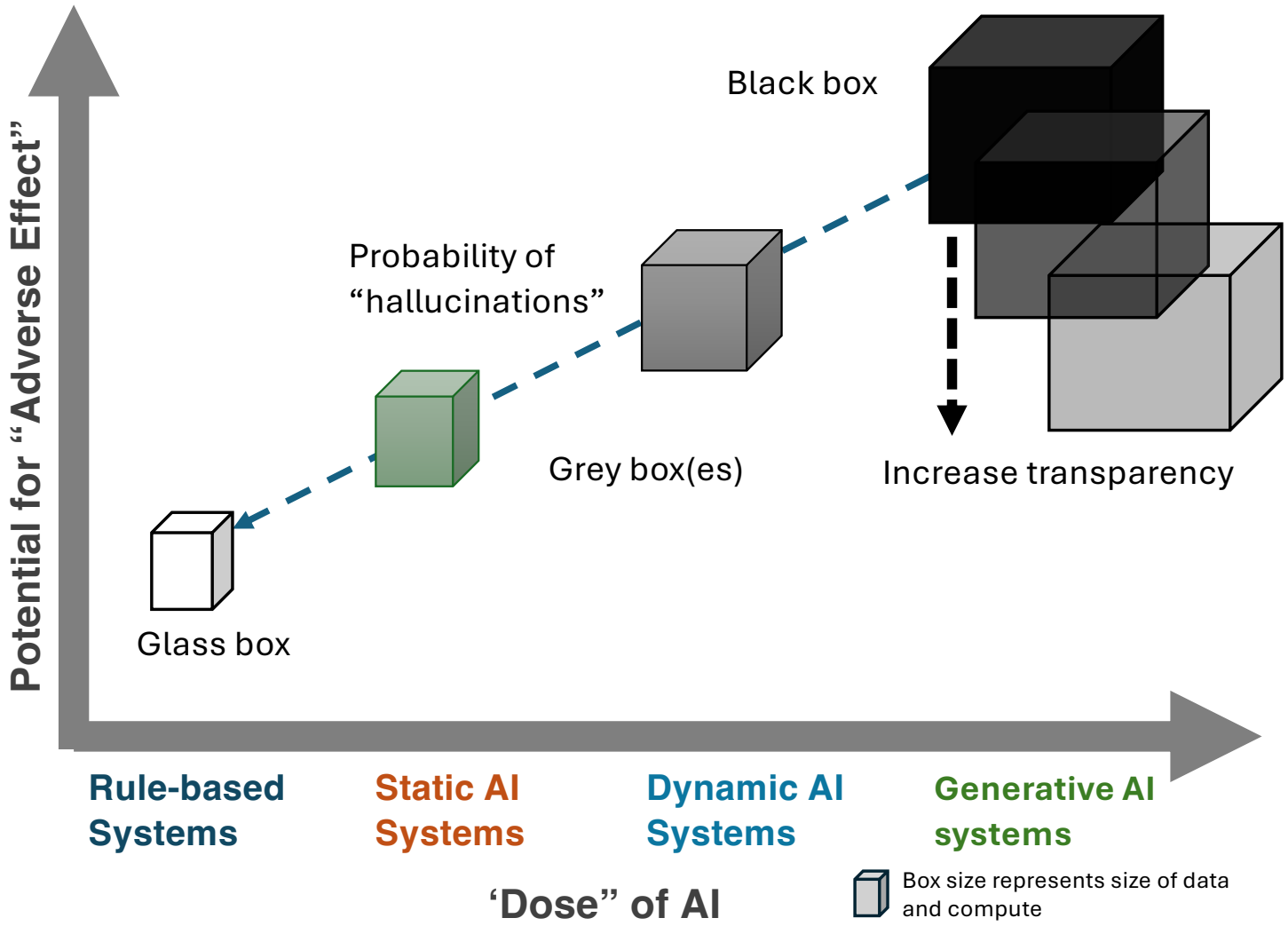
For example, instead of saying *"I think what you just said is a great idea"*, I can just give you a thumbs up. Or nod my head. Or simply smile.

And increase the bandwidth of productivity & collaboration



https://julian.digital/2025/03/27/the-case-against-conversational-interfaces/
https://roadtoartificia.com/p/the-hypermodal-interface

How can we preserve agency – Err on side of Transparency



Transparent Reasoning

Always know the "Why" behind an action.

Actions are based on reasoning and shaped by your behavior, not hard-coded rules. A "Chain of Thoughts" allows you to build intuition and trust.

arXiv:2311.09677v3 [cs.CL] 7 Jun 2024

R-Tuning: Instructing Large Language Models to Say 'I Don't Know'

Hanning Zhang[♣*], Shizhe Diao[♣*], Yong Lin[♣*], Yi R. Fung[♡], Qing Lian[♣], Xingyao Wang[♡], Yangyi Chen[♡], Heng Ji[♡], Tong Zhang[♡]

[♣]The Hong Kong University of Science and Technology

[♡]University of Illinois Urbana-Champaign

{hzhangco, sdiaoaa, ylindf, qlianab, tongzhang}@ust.hk

{yifung2, xingyao6, yangyic3, hengji}@illinois.edu

Abstract

Large language models (LLMs) have revolutionized numerous domains with their impressive performance but still face their challenges. A predominant issue is the propensity for these models to generate non-existent facts, a concern termed *hallucination*. Our research is motivated by the observation that previous instruction tuning methods force the model to complete a sentence no matter whether the model knows the knowledge or not. When the question is out of the parametric knowledge, it will try to make up something and fail to indicate when it lacks knowledge. In this paper, we present a new approach called **R**efusal-Aware Instruction **Tuning** (**R-Tuning**). This approach is formalized by first identifying the disparity in knowledge encompassed by pre-trained parameters compared to that of instruction tuning data. Then, we construct the refusal-aware data based on the knowledge intersection, to tune LLMs to refrain from responding to questions beyond its parametric knowledge. Experimental results demonstrate R-Tuning effectively improves a model's ability to answer known questions and refrain from answering unknown questions. Furthermore, when tested on out-of-domain datasets, the refusal ability was found to be a meta-skill that could be generalized to other tasks. Further analysis surprisingly finds that learning the uncertainty results in better calibration and an improved ability to estimate the uncertainty than uncertainty-based testing.[1]

Figure 1: An illustration of the parametric knowledge distribution and the instruction tuning data distribution. Pre-training embeds a large volume of parametric knowledge, while fine-tuning may involve knowledge that is not necessarily in the parametric knowledge. We explore the benefits of differentiating instruction tuning data based on parametric knowledge.

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across numerous tasks; however, they are also plagued by various issues, such as the propensity of large models to fabricate non-existent facts, a phenomenon commonly referred to as *hallucination* (Maynez et al., 2020a).

Towards mitigating the hallucination, current mainstream approaches include retrieval-based methods (Peng et al., 2023; Li et al., 2023b; Luo et al., 2023), verification-based methods (Manakul et al., 2023; Elaraby et al., 2023; Cohen et al., 2023; Du et al., 2023; Gou et al., 2023), and so forth.

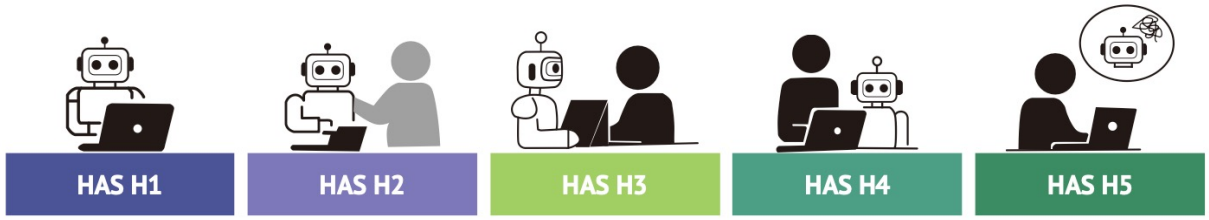
In this paper, we first identify the cause of hallucination, attributing it to the significant gap existing between the knowledge derived from human-labeled instruction tuning datasets and the parametric knowledge of LLMs. In the process of developing a large model, previous studies (Min et al., 2022; Wang et al., 2023; Zhou et al., 2023) demonstrate that almost all knowledge is acquired in the pre-training stage, while instruction tuning teaches formatting and chain-of-thought prompting guides knowledge elicitation. Consider Figure 1 as an example. During pre-training, models embed a large volume of factual knowledge, compressing it within their parameters and the fine-tuning process may include data that is out of the parametric knowledge. However, traditional fine-tuning methods force the models to complete each sentence. Even when faced with questions beyond their knowledge boundary, they venture to provide

[*]Equal Contribution.

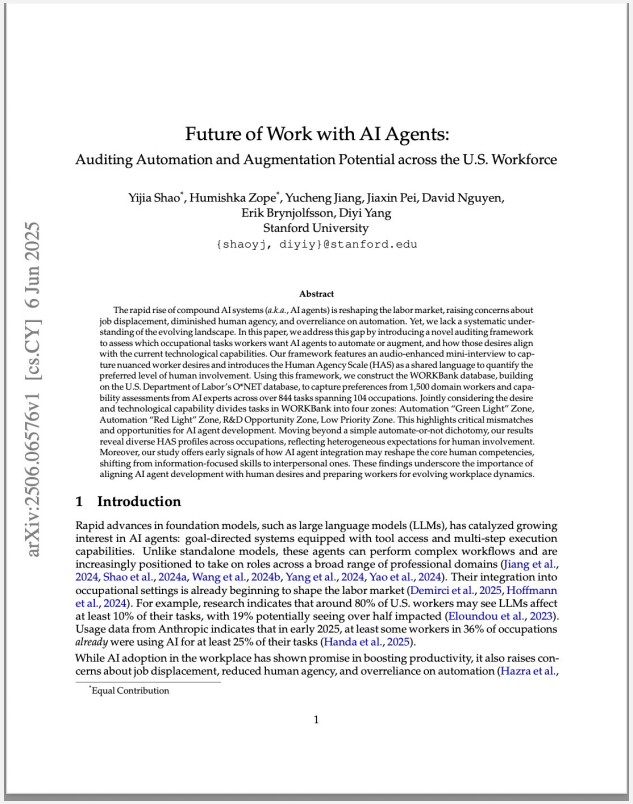
[1]Our code is available at https://github.com/shizhediao/R-Tuning.

Learning, Testing & Evaluation (LTE) is key

Human Agency Scale (HAS) to quantify the team dynamics and degree of human involvement required



	HAS H1	HAS H2	HAS H3	HAS H4	HAS H5
Team Dynamics	**AI Agent Drives Task Completion** The AI agent takes primary reponsibility for task execution with no or minimal human oversight.		**Equal Partnership** The human and the AI agent collaborate closely throughout the task.	**Human Drives Task Completion** The human takes primary responsibility for task execution with varying levels of AI assistance.	
Required Human Involvement	AI agent handles the task entirely on its own without your involvement.	AI agent needs your input at a few key points to achieve better task performance.	AI agent and you work together to outperform either alone.	AI agent needs your input to successfully complete the task.	Task completion fully relies on your involvement.
AI Role	**Automation** AI replaces human capabilities		**Augmentation** AI enhances human capabilities		
Example Tasks	• Transcribe data to worksheets and enter data into computer. • Run monthly network reports.	• Devise trading, option, or hedge strategies. • Accept payment on accounts.	• Create core game features, including storylines, role-play mechanics, etc. • Compile and analyze experimental data and adjust experimental designs as necessary.	• Coordinate and direct the financial planning, budgeting, procurement, or investment activities. • Design, plan, organize, or direct orientation and training programs.	• Participate in online forums or conferences to stay abreast of online retailing trends, techniques, or security threats.



Future of Work with AI Agents:

Auditing Automation and Augmentation Potential across the U.S. Workforce

Yijia Shao*, Humishka Zope*, Yucheng Jiang, Jiaxin Pei, David Nguyen,
Erik Brynjolfsson, Diyi Yang
Stanford University
{shaoyj, diyiy}@stanford.edu

Abstract

The rapid rise of compound AI systems (*a.k.a.*, AI agents) is reshaping the labor market, raising concerns about job displacement, diminished human agency, and overreliance on automation. Yet, we lack a systematic understanding of the evolving landscape. In this paper, we address this gap by introducing a novel auditing framework to assess which occupational tasks workers want AI agents to automate or augment, and how those desires align with the current technological capabilities. Our framework features an audio-enhanced mini-interview to capture nuanced worker desires and introduces the Human Agency Scale (HAS) as a shared language to quantify the preferred level of human involvement. Using this framework, we construct the WORKBank database, building on the U.S. Department of Labor's O*NET database, to capture preferences from 1,500 domain workers and capability assessments from AI experts across over 844 tasks spanning 104 occupations. Jointly considering the desire and technological capability divides tasks in WORKBank into four zones: Automation "Green Light" Zone, Automation "Red Light" Zone, R&D Opportunity Zone, Low Priority Zone. This highlights critical mismatches and opportunities for AI agent development. Moving beyond a simple automate-or-not dichotomy, our results reveal diverse HAS profiles across occupations, reflecting heterogeneous expectations for human involvement. Moreover, our study offers early signals of how AI agent integration may reshape the core human competencies, shifting from information-focused skills to interpersonal ones. These findings underscore the importance of aligning AI agent development with human desires and preparing workers for evolving workplace dynamics.

1 Introduction

Rapid advances in foundation models, such as large language models (LLMs), has catalyzed growing interest in AI agents: goal-directed systems equipped with tool access and multi-step execution capabilities. Unlike standalone models, these agents can perform complex workflows and are increasingly positioned to take on roles across a broad range of professional domains (Jiang et al., 2024, Shao et al., 2024a, Wang et al., 2024b, Yang et al., 2024, Yao et al., 2024). Their integration into occupational settings is already beginning to shape the labor market (Demirci et al., 2025, Hoffmann et al., 2024). For example, research indicates that around 80% of U.S. workers may see LLMs affect at least 10% of their tasks, with 19% potentially seeing over half impacted (Eloundou et al., 2023). Usage data from Anthropic indicates that in early 2025, at least some workers in 36% of occupations *already* were using AI for at least 25% of their tasks (Handa et al., 2025).

While AI adoption in the workplace has shown promise in boosting productivity, it also raises concerns about job displacement, reduced human agency, and overreliance on automation (Hazra et al.,

*Equal Contribution

1

How can we preserve agency – preserving culture & context

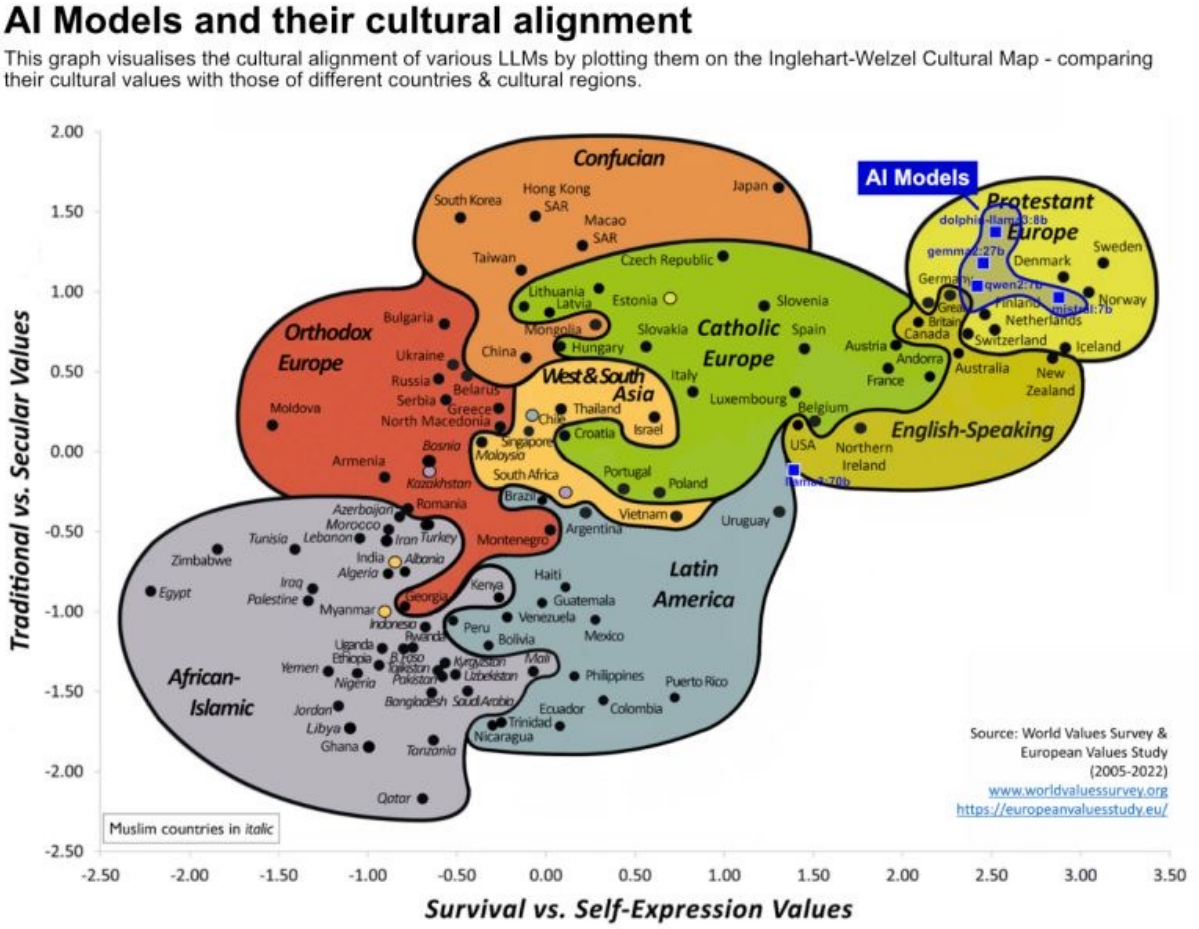
A study* tested major AI Models - the foundations powering tools millions use daily - against cultural values from 107 countries worldwide

English is the reserve currency of AI

Each one reflected the same assumptions - those of English-speaking, Western European societies. None aligned with how people in Africa, Latin America, or the Middle East actually build trust, show respect, or resolve conflicts.

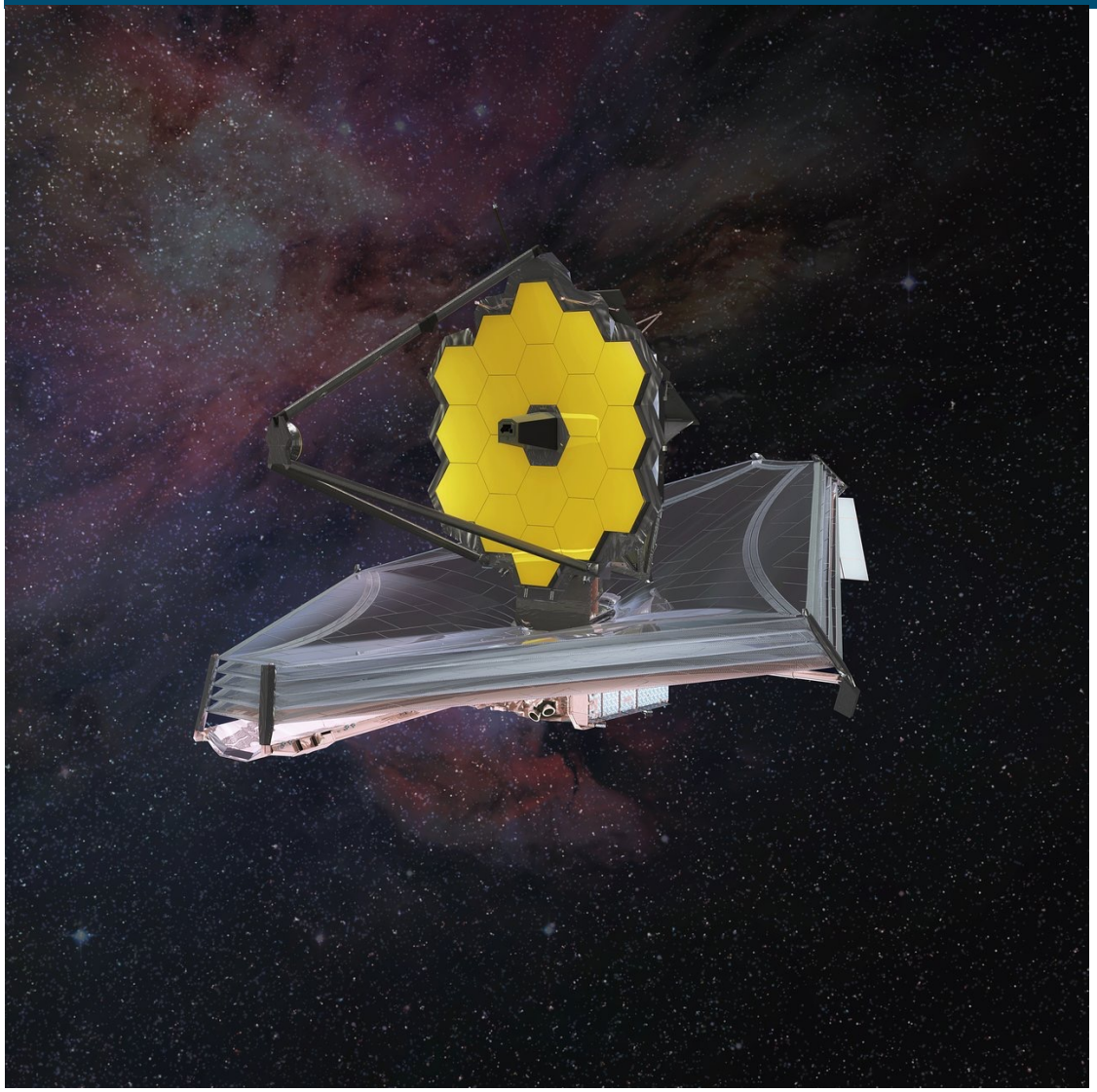
Agentic Systems, especially in healthcare need

cultural intelligence



* https://www.linkedin.com/posts/tevbannerman_theres-something-almost-nobody-is-talking-activity-7358405153139367937-4GVf/

The Healthcare Telescope



From Better Eyes to New Windows

Initially, telescopes were valued simply for magnifying what the eye could see—sharper and farther—but they evolved to detect entirely new phenomena like radio waves or gravitational signatures, redefining observation itself

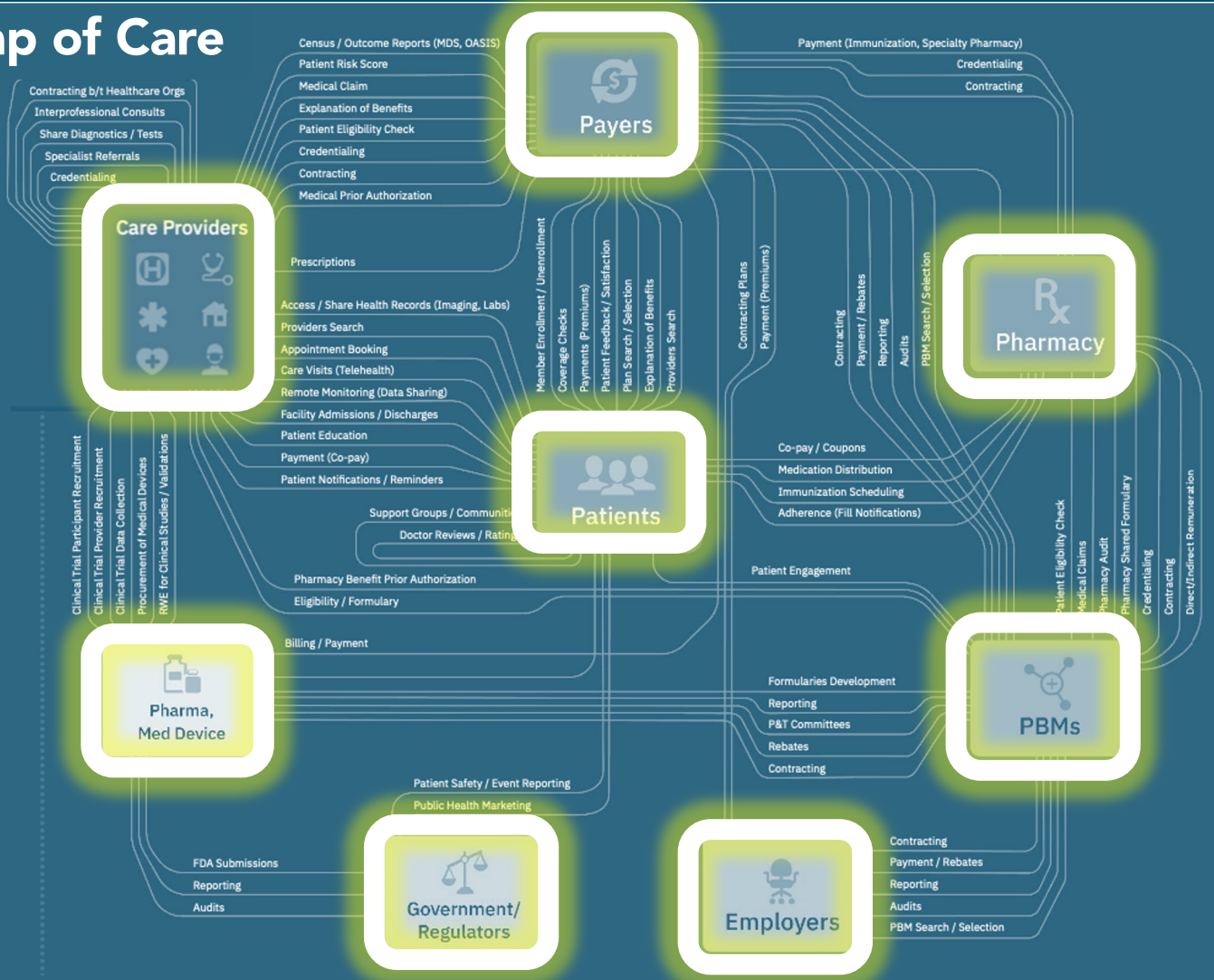
A Shared Direction

The most powerful tools do not just reflect our capabilities back to us. They help us work in new ways, see new patterns, and act with new perspective. To do that, they need to be introduced with care. Not as replacements for human roles, but as extensions of human systems

The Telescope problem

Increase the aperture of utility in care

The Subway Map of Care



SUMMIT

Disha

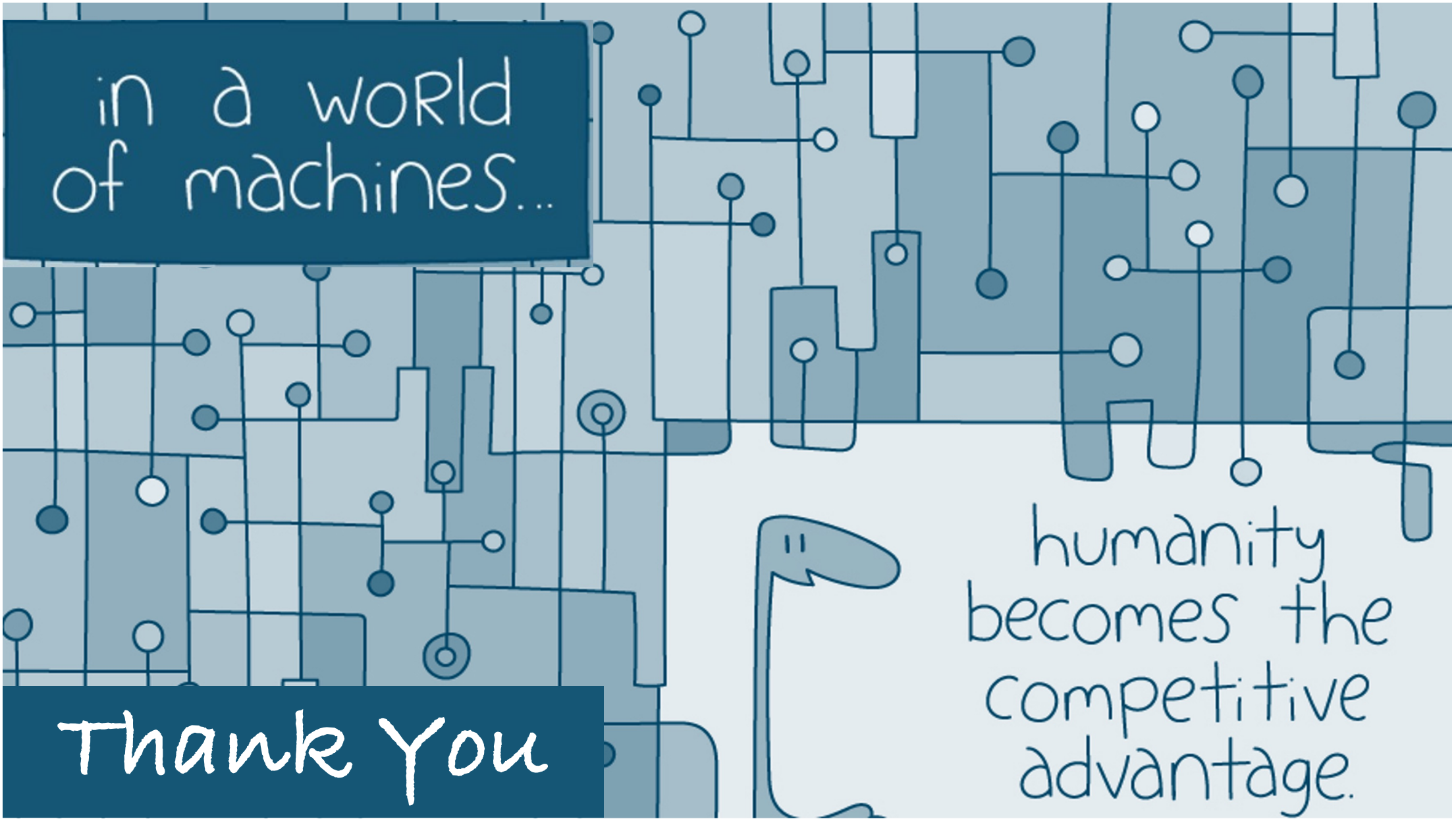
Platform for supply side of health

Disha (दिशा)
means
"direction"
and
"path"

Happy Providers make Happy Patients

Providers

Patients


in a world of machines...
humanity becomes the competitive advantage.
Thank You

THANK YOU