UW Medicine

January 31, 2024

Tim Dellit, M.D.
Chief Executive Officer, UW Medicine
Executive Vice President for Medical Affairs and
Dean of the School of Medicine, University of Washington
1959 NE Pacific Street, Box 356350
Seattle, Washington 98195-6350

RE: Response to Aug. 25, 2023 Large Language Models Workgroup Charge Letter

Dear Dr. Dellit:

Thank you for the opportunity to lead the Large Language Models (LLM) Workgroup. The purpose of the LLM Workgroup was to ensure UW Medicine is prepared to address the unique considerations generative AI tools, including LLMs, raise in the healthcare setting, including patient care, business operations supporting our healthcare activities and research that is integrated with the clinical environment.[1] The Workgroup has spent the past five months exploring the landscape of generative AI in the healthcare setting and assessing how to position UW Medicine to leverage this new technology responsibly to advance our mission to improve the health of the public.

The Workgroup began by developing foundational principles to guide our approach to generative AI. These recommended guiding principles include the following proposed "North Star":

> We strive to innovate and implement generative AI in a responsible manner to advance UW Medicine's mission.
>
> - Generative AI presents significant opportunity and should be considered for use to accelerate our strategic, financial and operational goals.
> - Research involving generative AI has significant potential to advance this emerging technology (particularly for application in the healthcare setting), which can result in accelerated translation of tools to improve the delivery of healthcare for the benefit of our patients.
> - Any use of generative AI tools must be responsible, compliant (with relevant laws, regulations and policy), ethical and balance potential benefit with potential risks.

The recommended guiding principles are attached as Exhibit A.

The Workgroup tackled its charge by creating three subgroups: 1) Use & Applications, 2) Risk Identification and 3) Governance. Membership of the sub-groups was driven by subject matter expertise

[1] The scope of our work was on generative AI, with a primary focus on large language models. For simplicity, this report will use the term "generative AI," which is inclusive of LLMs.

and interest. Each subgroup prepared a detailed report of its work, inclusive of feedback from the full workgroup *(incorporated as Exhibits B-D)*. The following summary report highlights the key findings of each subgroup, as well as insights from the collective Workgroup.

WHAT ARE GENERATIVE AI AND LARGE LANGUAGE MODELS?

Generative AI models can produce synthetic text, images, video and sound. The Workgroup focused on a class of models known as Large Language Models (LLMs) as applied to generate text, because these models are at the forefront of publicly accessible AI (e.g., ChatGPT) and poised for integration into clinical workflows. These models have unprecedented abilities to process and generate language and present significant potential in the clinical context (e.g., alleviation of documentation-related burdens).

LLMs differ fundamentally from methods used in most current applications of AI in medicine, because it is possible to interact with LLMs in natural language. Questions in natural language can be asked by anyone, eliminating typical barriers to using AI (though knowledge of how best to pose these questions can lead to more acceptable and reliable answers). Responses in natural language from LLMs look convincing: they tend to be coherent and are grounded in large amounts of training data – more text than any human correspondent could possibly read. However, these responses may contain subtle inaccuracies, referred to as “confabulations” or “hallucinations,” that are only detectable with scrupulous review. For variety, some randomness is also typically built into the mechanism used to generate responses, and LLMs may respond differently to the same question asked twice. For these reasons, human interaction is typically needed to sufficiently evaluate LLMs. Evaluation also is often subjective, because there is no single correct answer to a request to draft a response to an inbox message, for example. By contrast, an AI model that predicts sepsis can be evaluated against observed outcomes. Because of their size and architecture, it also is difficult to know why an LLM responded a particular way, and developing methods to control their output is a research frontier.

Looking more deeply, LLMs are neural network models with billions of parameters trained to predict subsequent words in large amounts of text, and in some cases, to follow instructions and produce preferred responses. Only a few technology companies possess the resources required to train the largest and best-performing models, and the details of the text they are trained on are seldom revealed. For example, GPT-4 was trained by OpenAI on unspecified text using hardware provided by Microsoft, who have in turn partnered with Epic to provide an integrated infrastructure for clinical use of LLMs. LLMs can also be accessed by the public, which typically involves sending questions to a commercially hosted model (such as ChatGPT) over the internet, an approach that presents risks when sensitive information is included.

GENERATIVE AI GOVERNANCE APPROACH

At the heart of the Workgroup charge is how UW Medicine should organize itself to take advantage of generative AI in a responsible manner. We know this technology presents significant opportunity. But how do we navigate the complexities associated with use of this technology in the healthcare setting so that we can translate the hype into tangible success? Healthcare systems across the country are asking themselves this same question.

How UW Medicine approaches generative AI governance will determine in large part whether we can efficiently and effectively leverage this technology in the healthcare setting. But what does effective governance look like in this space? The Workgroup identified several key principles required for successful governance:

- Making the right thing to do the easy thing to do
- Taking a proactive (versus reactive) approach
- Implementing formal/transparent operational processes
- Developing policies that are not overly restrictive and that are monitored and enforced
- Ensuring clarity of scope (e.g., generative AI vs. other forms of AI, clinical vs. research)
- Identifying accountable leadership, trusted and empowered by the organization
- Designing a scalable approach
- Ensuring governance is efficient, integrated and aligned with other governance structures
- Framing our institutional approach positively (here's what you can do and how to do it versus here's what you cannot do)
- Ensuring broad stakeholder involvement, internally and in alignment with its affiliated institutions.
- Appropriately resourcing the work to support our approved institutional approach

These principles led to a broad concept of governance, inclusive of the following key areas: 1) strategy, 2) policy, 3) governance structure, 4) operational workflows, 5) risk assessment, 6) communication and education and 7) leadership and accountability. A successful generative AI governance approach should address each of these areas. *See Figure 1.*

In its charge, the Workgroup was asked to develop a proposed committee or governance structure for developing policies, addressing issues and overseeing UW Medicine's institutional approach to the use and training of LLMs. To design and advance the governance needed, the Workgroup recommends a phased approach.

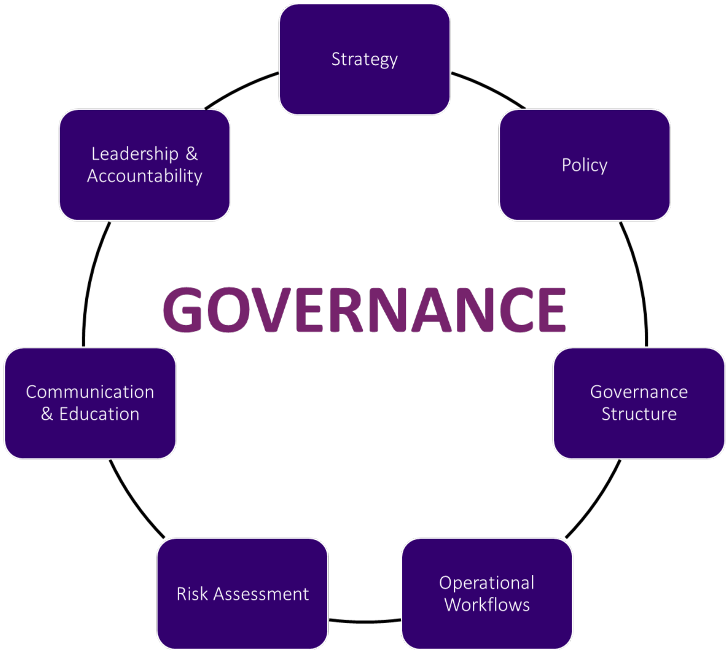


Figure 1: Governance

- **Phase 1 – Develop the Foundation**: The LLM Workgroup has completed Phase I, exploring the landscape of uses, applications and risks, as well as establishing guiding principles and proposing a broad go-forward governance approach.
- **Phase II – Build the Infrastructure:** Focus on building the infrastructure needed to support Phase III, including development of a defined and comprehensive business approach to generative AI in the healthcare setting, including key concrete deliverables.

- **Phase III -- Steady State**: Launch (and operated pursuant to) the permanent structure.

As the work of the initial Workgroup (Phase I) concludes, we recommend **establishing a short-term Generative AI Taskforce** to take the key findings from Phase I and to develop and recommend a business approach, including the deliverables described herein, needed to move UW Medicine to a Phase III steady state. The recommended Taskforce deliverables align with the key areas of governance and are described in detail in the subsequent sections of this report. If resourced appropriately, Phase II could be completed in approximately nine (9) months.

STRATEGY

Establishing an institutional strategy to guide work, including allocation of finite resources to accomplish key objectives, is critical for any initiative. It is particularly important in the generative AI space, because the landscape of generative AI use and applications is vast and continues to expand and evolve rapidly.

It would be easy to get bogged down in a sea of generative AI pilots, applications and vendors. To ensure we are focused and working toward a defined vision, the Workgroup recommends the Generative AI Taskforce **develop a UW Medicine strategy for generative AI in the healthcare setting.** This work should include, without limitation, an assessment of how generative AI can be leveraged to advance the UW Medicine clinical strategic plan, a prioritization framework to guide what tools/functionality we pilot, exploration of partnership opportunities (e.g., vendor partnerships, consortia/collaboratives), learnings from and alignment with peer institutions, consideration of leveraging external generative AI tools versus building internal capability to develop generative AI solutions, and potential philanthropic or other funding to support implementation. Finally, the strategy should address the balance between encouraging innovation and engaging with generative AI in a safe, ethical and responsible manner in the clinical environment.

To inform this next phase of work, the Workgroup surveyed existing and potential applications for LLMs that UW Medicine might contemplate using. The Workgroup investigated 27 different generative AI healthcare-related use cases. These 27 use cases fall into four categories or “method groups”:

1. **Drafting text and ambient listening note generation** – Clinician-facing use cases that take speech and/or text input and generate text output, such as messages or documentation. The resulting text should be reviewed for correctness, to mitigate the risk of model-derived inaccuracies. Applications to support the use cases explored in this category are all available now, are in flight or are under development by Epic, as well as other third-party vendors (UW Medicine has not yet implemented any of these use cases). They aim to decrease provider burnout by automating portions of workflows. They may also increase patient satisfaction by reducing provider response times. Some initial reports from early adopters indicate increased provider efficiency, satisfaction and work-life balance. These tools could potentially expand access to care at UW Medicine by increasing capacity for clinical work.
2. **Search, synthesis and analysis** – Applications for analyzing and synthesizing large and diverse data at scale. There is a strong incentive to provide HIPAA-compliant “sandbox” environments to support the development of local expertise and mitigate the risk of UW employees exploring

LLMs through less secure channels. In addition, currently available tools to support human analysts, such as Epic's SlicerDicer SideKick and UW Medicine's Leaf AI, have the potential to enhance the productivity of UW Medicine analytics staff, and commercial products are available that could support search and summarization of policy and procedure documents. As they are neither patient nor provider facing, these use cases offer benefits with minimal risk, while providing opportunities to develop institutional expertise. The Workgroup also identified several aspirational opportunities based on our understanding of LLM capabilities (e.g., no current product exists). Successful implementation of locally developed solutions for these proposals would require development of institutional skills, staff and infrastructure and would involve substantial effort for development, implementation and validation. That said, these proposals provide an important context for considering our long-term institutional strategy for development and adoption of generative AI, including LLMs**.**

3. **Translation** – Summarization, abstraction, and other transformations of language from one format or audience to another. Use cases include translation from English to non-English languages to reduce burden on translators, and summarization of information, from a patient's medical chart for example, for a non-medical audience (e.g., generating patient-facing care plan or lab report summaries, with provider review, translating informed consent forms). The use cases provide opportunities to improve patient-centered care and revenue cycle. LLMs can fundamentally change the way we conduct research and offer care for patients by meeting patients "where they are" in a language and vocabulary that matches their needs and preferences. Some of these uses cases have applications that are ready for use today, while others would require further feasibility studies. Epic has also expressed interest in supporting some of the translation use cases, so more information may be available in the near future.
4. **Augmentation/Automation and scheduling** – Large-scale augmentation/automation of repetitive tasks typically requiring specialized training or knowledge. Most of the augmentation/automation and scheduling use cases lean more aspirational with limited studies and/or commercial products available. While the use cases have the potential to be impactful financially, they are still largely theoretical. The one exception is the ability to call patients for appointment scheduling or research data collection—there is a tool available today that performs the tasks necessary effectively. Piloting lower risk, high impact use cases such as automated phone calls for research may be a reasonable option for gaining institutional knowledge.

The chart on the following page outlines all the use cases explored in each method group. For an in-depth analysis of each use case, including potential impact and feasibility, please see *Exhibit B, Use & Applications Subgroup Report*.

Method Group	Use Cases
Drafting Text and Ambient Listening Note Generation	Ambient listening note documentation
	Drafting in-basket replies
	Drafting Discharge or Interim Summaries
	Order Composer for inpatient and ambulatory settings
	Drafting Prior Authorization requests
Search, Synthesis and Analysis	Augmented/automated determination of patients eligible for clinical trials or studies using clinical data
	Compliance Surveillance and Coding of CPT coding for E&M and Procedural Services; ICD10-CM coding
	Clinical Decision Support including interpretation of patient symptoms and signs; Physical examination findings, Imaging, lab and other diagnostic studies
	Data integration from multiple sources
	Problem List Cleanup
	Analytics query development
	Semantic search and knowledge extraction for policies, procedures, and job aids
	General purpose HIPAA-compatible generative AI sandbox
Translation	Chart summarization
	Patient-facing care or plan summarization
	Point-of-Care language translation
	Chart Abstraction
	Interactive patient intake
	Patient-centered pathology reports
	Revenue Cycle – Customer Service inquires via MyChart
	Revenue Cycle – Denial Appeals
	Revenue Cycle – Coding
Augmentation/Automation & Scheduling	Calling patients for appointment scheduling or research data collection
	Telemedicine/nurse triage call in
	Improving utilization of clinic appointments and OR block time
	Improving processes to reduce the need to capture Medicare ABNs and commercial insurers' waivers (to reduce non-coverage determinations) and to improve ABN/waiver utilization when required
	Training and Education of patients and providers

The Workgroup anticipates that the list of potential use cases for these tools and functionality in the healthcare setting will continue to grow and evolve. An important concern across all these use cases is the need to develop institutional resources for evaluation and risk assessment.

POLICY

Policies provide guidance, consistency, accountability and clarity on what is and is not acceptable within an organization. Generative AI is creating a lot of attention; people are excited about its potential and want to start utilizing this new technology, whether through publicly available tools, functionality integrated with existing software (e.g., Epic or Microsoft Office) or by developing tools in-house. Without clear guardrails, people may use generative AI tools without understanding the risks and potential consequences.

The Workgroup recommends the Generative AI Taskforce **develop and propose an institutional policy to govern the use of generative AI in the healthcare setting.** The policy (or policies) should address use of these tools/functionality for patient care, business operations supporting the healthcare enterprise and use of publicly available tools in the course of everyday work. The policies also should address any unique requirements associated with clinical research involving generative AI or the use of clinical data to support research involving generative AI.

A clear institutional policy will help mitigate certain risks associated with the use of generative AI. When developing the institutional policy, the Workgroup recommends the Taskforce consider the following areas:

- Use, access and disclosure of protected health information (PHI), personally identifiable information (PII) and other sensitive data;
- Institutional risk tolerance regarding the protections required to share large de-identified data sets;
- Allowable level of uncertainty regarding training data accuracy given the unknowns of external models;
- Allowable use of generative AI in clinical care, including a definitive statement that no care should be provided without human clinical judgement and decision-making;
- Prohibition of the sale or other "commercialization" of PHI or PII; and
- Clear decision-making authority and approval pathways.

The above is not an exhaustive list, but rather a starting place for further work. In addition, there are many existing policies (e.g., privacy and security policies) that the Taskforce should review to determine whether updates are needed to ensure our administrative controls are in alignment across UW Medicine.

GOVERNANCE STRUCTURE

The Workgroup envisions the steady state governance structure (Phase III) will be responsible for overseeing and managing UW Medicine's institutional approach to generative AI going forward. The Workgroup recommends the Generative AI Taskforce **design a proposed long-term governance structure (including draft charters and proposed membership)** to oversee UW Medicine's institutional approach to generative AI in the healthcare setting. In creating the structure, the Taskforce should consider existing governance structures and how this new structure might fit within, replace and/or complement what exists today, taking care to avoid duplication, confusion and overloading those with key subject matter expertise. When proposing membership, the Taskforce should consider the scope of expertise needed on the governance structure versus ad hoc participants available for consultation. The Workgroup encourages consideration of a broad spectrum of expertise, including technical experts, clinical operations, School of Medicine departmental representation, legal, risk management, human resources, compliance, healthcare equity, patient safety, patient experience, quality, finance and marketing and communications, along with representatives of affiliated institutions with interests closely aligned with UW Medicine.

OPERATIONAL WORKFLOWS

Hand in hand with the governance structure are the operational workflows that determine and support how generative AI opportunities are brought forward and evaluated. Since generative AI in the healthcare environment covers a broad array of uses for both clinical support and research integrated with the clinical environment, a single workflow likely will not be sufficient.

The Workgroup recommends the Generative AI Taskforce **develop (or define) operational workflows** to ensure that there are processes (whether net new or existing) to support, at a minimum, intake, assessment and approval for the following three categories:

1. Tools/functionality to support clinical activity or activity in support of the clinical enterprise;
2. Clinical research involving generative AI; and
3. Sharing of clinical data for research involving generative AI.

Just like the governance structure, the workflows should consider existing structures and processes and whether this work can be absorbed into those structures with modification or requires something new.

RISK ASSESSMENT

Any discussion of generative AI must include the potential risks associated with both implementation and use. The news continues to highlight AI mishaps, spurring federal and state regulatory initiatives. In the healthcare environment, the risks are heightened given the nature of our data, the complexity of the healthcare legal and regulatory environment, and the sacrosanct relationship between provider and patient. The Workgroup was asked to provide an overview of the legal, regulatory, ethical and mission-related risks associated with the use and training/fine-tuning of generative AI. The Workgroup identified 14 risks within eight risk areas. The risks were ranked based on the potential likelihood that the risk could occur and the potential negative legal, reputation or financial impacts if the risk occurred.

Ranking	Risk Area	Risk
1	Legal	Legal/regulatory landscape
2	Privacy	Data breach
3	Accuracy and integrity	Model outputs

Ranking	Risk Area	Risk
4	Security	Data use and storage
5	Other	Concerns re: LLM as a new initiative
6	Legal	Contracts
7	Model Bias	Discrimination
8	Medical/patient care	Clinical care
9	Medical/patient care	Malpractice risk
10	Privacy	Sale of PHI
11	Human Resources	Human Resources
12	Other	Brand/reputation
13	Medical/patient care	Patient experience
14	Legal	Intellectual Property

For a detailed look at each risk, see Exhibit C, Risk Identification Subgroup Report.

Some risks may be mitigated by UW Medicine's governance approach, such having a clear institutional policy on generative AI and updating existing policies as needed to align, communicating and educating across multiple audiences, including stakeholders from a variety of offices in the governance structure, and ensuring adequate resourcing.

However, as outlined above, there is a vast array of generative AI use cases, and while the uses can be categorized into useful groups, each use case has its own unique considerations. The risks associated with use of generative AI or use of clinical data to train/fine-tune generative AI will depend on a variety of factors and will require case-by-case analysis. To this end, the Workgroup recommends the Generative AI Taskforce **define processes to minimize risk that can be built into the operational protocols and systems, to the extent possible.** These processes should include, at a minimum: a) a framework or rubric that can be used operationally to assess risk on a case-by-case basis and enable tailored risk assessment; b) plans to monitor, audit and/or decommission tools/functionality; and c) identified pathway(s) to address unintended consequences, as appropriate.

The Workgroup has called out risk assessment as a separate key area of governance given how crucial due diligence and adequate mitigation measures are in this space. However, we envision the risk assessment of opportunities would be part of the operational workflows in a steady-state environment. That said, as generative AI continues to evolve, so will the risks, so it will be important to continue to reassess the risk landscape regularly.

COMMUNICATION AND EDUCATION

One of the most consistent themes throughout the Workgroup discussions was the need for a thoughtful approach to communication and education around our institutional approach to use of generative AI. Most people have heard of ChatGPT and other generative AI models and uses, but many do not have a firm understanding of how these tools work or their potential risks. As we move toward a UW Medicine institutional approach to generative AI in the healthcare setting, we need easily accessible and transparent communication to both internal and external audiences. In addition, education will be crucial (both internally to our staff, faculty and trainees and externally to our patient population, for example). Such education could include topics such as appropriate uses of generative AI tools, privacy and security risks, clinical care expectations, bias and discrimination risks, and ethical principles of generative AI use.

The Workgroup recommends the Generative AI Taskforce **create a robust education, communication and engagement strategy** targeting a variety of audiences (e.g., faculty, staff, trainees, patients, policymakers and labor unions). The strategy should propose the types of materials needed to support it and outline the risks of failure to implement a comprehensive approach to education, communication and engagement around the use of generative AI in the healthcare setting.

LEADERSHIP AND ACCOUNTABILITY

A future state that enables us to implement these tools in the healthcare setting in a responsible manner consistent with the Workgroup's guiding principles will require clear and accountable leadership.

The Workgroup recommends the Generative AI Taskforce be charged with **evaluating options and proposing an operational leadership structure accountable for our success** in building a program consistent with our institutional approach to generative AI. This work should include, at a minimum:

- Surveying organizational structures at peer institutions;
- Assessing the bandwidth necessary to tackle the work, as outlined by Generative AI Taskforce;
- Creating clarity around scope of the proposed accountability; and
- Establishing a vision for the collaboration required across UW Medicine and the University (e.g., Institute for Medical Data Science).

OTHER CONSIDERATIONS

In support of the recommendations outlined above, the Workgroup recommends the Generative AI Taskforce be charged with **developing financial, resource and other recommendations**, as appropriate, to support the proposed business approach in the next one (1) to three (3) years.

To be successful, financial investment will be critical and the level of resourcing dedicated to this work needs to be sufficient to execute on our to-be-approved institutional approach.

Insufficient investment will create risk, operational bottlenecks and ultimately, could put us at a competitive disadvantage vis-à-vis our competitors (both in the provision of accessible, high value care and as an employer of choice). Given the current financial situation of UW Medicine, the Generative AI Taskforce should propose an approach to financial resourcing that balances our need to be well-positioned in this space but stewards our funds in a responsible manner.

FINAL THOUGHTS

To use an analogy that resonated with the full Workgroup: for purposes of our clinical operations, we would like to be at the front of the “herd” as this revolution in technology changes the way we deliver high-quality, equitable care, conduct our business and support the well-being of our staff, faculty and trainees. UW Medicine is well positioned to lead in this space, with institutional subject matter expertise, extended experience with the deployment and evaluation of clinical decision support tools, and proximity to key commercial players.

An overarching theme (and risk) that the Workgroup identified is the rapid speed at which the generative AI space is evolving. We must be nimble and continue to move forward with our work quickly, while ensuring the necessary due diligence and operational infrastructure to support the work is in place. Our approach must balance innovation and opportunity with our obligation to use this technology in a safe, ethical and responsible way. Only by striking the right balance will we be able to use these incredible new tools in the healthcare setting to advance our mission to improve the health of the public.

We’d like to thank the members of the Workgroup for their expertise and significant contributions over the past few months. We would also like to give special thanks to our subgroup leads Nic Dobbins (Use and Applications), Beth DeLair and Grace Lin (Risk Identification) and to Lauren Fischer and Nadege Mohr, who helped support this work from start to finish. This Workgroup’s energy and engagement over the last several months is an indicator of the significant opportunity that lies ahead.

We look forward to your review and discussion of next steps. Please let us know if you have any questions. Thank you.

Sincerely,

DocuSigned by:
Ana Anderson
7ECC95E1C55D43F...
Ana Anderson, Co-Chair
Senior Director
Clinical Business Affairs
UW Medicine

DocuSigned by:
TCohen
9DBA9640B281435...
Trevor Cohen, Co-Chair
Professor
Department of Biomedical Informatics & Medical Education
University of Washington

DocuSigned by:
Eric J. Neil
D282961B9090446...
Eric Neil, Executive Sponsor
Chief Information Officer
UW Medicine

DocuSigned by:
Margaret Peyton
31840B5BE4E9438...
Margaret Peyton, Executive Sponsor
Clinical Business Affairs & Regulatory Officer
Clinical Business Affairs
UW Medicine

And on behalf of the Workgroup members:

Hasan Ahmad

Sally Beahan

Todd Burstain

Nathan Cross

Augie D'Agostino

Beth DeLair

Robert Doerning

Nic Dobbins

Keith Eaton (Fred Hutch)

Lauren Fischer (PM)

Malia Fullerton

Marcia Gonzalez (Fred Hutch)

Lisa Hammel

Brad Henley

Noah Hoffman

Theresa Kim

Margaret Lane

Jeff Leek (Fred Hutch)

Jesse Levin

Grace Lin

Kristal Mauritz-Miller

Nadege Mohr (PM)

Sean Mooney

Leo Morales

Shawntá Mosley-App (Fred Hutch)

Adina Mueller (Fred Hutch)

Aimée Olivier (Fred Hutch)

Adam Parcher

Kelly Patrick (Fred Hutch)

Gerianne Sands

Anneliese Schleyer

Angad Singh

Paul Sutton

Peter Tarczy-Hornoch

Drew von Eschenbach

Meliha Yetisgen

Exhibits

- Exhibit A: Recommended Generative AI Guiding Principles
- Exhibit B: Use & Applications Subgroup Report
- Exhibit C: Risk Identification Subgroup Report
- Exhibit D: Governance Subgroup Report

<u>Exhibit A</u>

Recommended Generative AI Guiding Principles

North Star:

We strive to innovate and implement generative AI in a responsible manner to advance our mission.

- Generative AI presents significant opportunity and should be considered for use to accelerate our strategic, financial and operational goals.
- Research involving generative AI has significant potential to advance this emerging technology (particularly for application in the healthcare setting), which can result in accelerated translation of tools to improve the delivery of healthcare for the benefit of our patients.
- Any use of generative AI tools must be responsible, compliant (with relevant laws, regulations, and policy), ethical, and balance potential benefit with potential risks.

Guiding Principles:

- There must be oversight (including governance structure, policies, approval authority and ongoing operational accountability) and support (computational infrastructure, training, resources, documentation) regarding use of generative AI, including use of clinical data to train/fine-tune generative AI models.
- The legal/regulatory, ethical and mission-related risks associated with use of generative AI or use of clinical data to train/fine-tune generative AI will depend on a variety of factors and will require a case-by-case analysis.
- Whether or not to use a particular generative AI tool for a particular purpose must be assessed to ensure risks associated with use are identified, sufficiently mitigated and monitored, and that the potential benefits outweigh the potential risks.
- As an innovative and trusted healthcare delivery system, use, evaluation, and benchmarking of generative AI in practice should be both encouraged and transparent (both internally and externally) with appropriate guardrails and oversight.
- Clinical data is both valuable and highly sensitive. Commercial generative AI models should be reviewed for how this data may be used for learning or quality control to balance the benefits and risks. Any learning from this data should ensure a firewall that prevents the exposure of sensitive or private data.
- Institutional systems, structures and processes must be nimble to enable us to navigate this rapidly evolving environment efficiently.
- Use of generative AI should not change the current practices of provider professional autonomy and responsibility and patient autonomy in healthcare.
- We will strive to ensure that use of generative AI is fair to all populations of patients. For example, given a proposed use case we may audit LLM output for biased and potentially harmful language, or assess differences in accuracy of interpretation of language from different patient groups.

Exhibit B

Use & Applications Subgroup Report

Introduction 15

Summary & Key Insights 16

Drafting Text & Ambient Listening Note Generation 17

Ambient Listening Note Documentation 17

Summary 17

Key Insights 17

Use Cases 18

Search, Synthesis & Analytics 26

Summary 26

Key Insights 27

Use Cases 27

Translation 47

Summary 47

Key Insights 47

Use Cases 48

Augmenation/Automation & Scheduling 58

Summary 58

Key Insights 58

Use Cases 59

Introduction

This report summarizes early insights and findings from the Large Language Models (LLM) Workgroup Use & Applications subgroup. Our subgroup recorded, reviewed, and investigated over twenty different use cases related to the use of LLMs in healthcare. To do so, we first categorized LLM-related use cases by their primary medium of interaction or purpose, into four "method groups":

5. **Drafting text & ambient listening note generation** – Given a data context (e.g., an outpatient visit), generation of text such as messages or documentation.
6. **Search, synthesis & analysis** – Applications for analyzing and synthesizing large and heterogeneous data at scale.
7. **Translation** – Summarization, abstraction, and other transformations of data from one format or to one audience to another.
8. **Augmentation/Automation & scheduling** – Large-scale augmentation/automation of repetitive tasks typically requiring specialized training or knowledge.

Each method group includes use cases which encompasses multiple audiences and potential kinds of users, including those who would be directly interacting with LLMs and those affected by an LLM output or new workflow. The user groups identified are:

1. **Patients**
2. **Providers**
3. **Hospital system administration**
4. **Research personnel**

This report is organized by the above method groups. Each section begins with a table listing the use cases for a given method group, followed by a method-group specific summary and key insights subsection, and finally details of each use case. We evaluated and researched use cases along the following six categories of impact:

1. **Patients** – impacts to quality of care, safety, experience, satisfaction, and diversity, equity & inclusion.
2. **Providers** – impacts to provider satisfaction, such as augmentation/automation of time-consuming or tedious tasks.
3. **Administrative** – impacts to non-clinical and professional staff, including enhanced operational capabilities and efficiencies.
4. **Research** – impacts to UW researchers, including the performing of analyses, managing large datasets and translational research.
5. **Financial** – impacts on financial effort and resources, including IT infrastructure and project implementations.
6. **Technical and Operational** – impacts on operational complexity, implementations, technical difficulties and maturity of technologies used, including internal capabilities vs. vendor-based.

Summary & Key Insights

While the purpose of this report is largely exploratory, rather than a conclusive series of recommendations or suggestions for prioritization, our subgroup has nonetheless found a number of recurring themes among use cases along with tentative ideas for consideration. One theme in particular we found is that of **build vs. buy** (developing an internal solution versus licensing from a vendor). Considerations of build vs. buy are often specific to a given use case, and using a vendor solution and evaluating in a certain way may make sense for one case but not necessarily another. One commonality across use cases, however, is the need to **measure and evaluate** possible solutions and change after implementation:

1. **Comparing multiple solutions in an apples-to-apples fashion** – it is highly probable that in the near future certain use cases will have multiple possible LLM-based solutions. For example, Epic, third party software companies, and internal UW researchers may have models or systems which are capable (or purport to be) of fulfilling a given use case. How do we empirically determine the relative tradeoffs and risks of each? Even if we are inclined to use Epic-supported models where possible, how do we evaluate where those models fall short? To answer these questions, we may need evaluation procedures, policies, and internal datasets to compare "apples-to-apples". Having such resources and organization in place would allow us to efficiently and rigorously test these solutions, and ultimately enable us to make better decisions.

2. **Evaluating existing baselines and measuring change** – After implementing a given solution for a use case, how well can we measure improvement (and return on investment)? And related, how well can we measure current baseline workflows without the use of LLMs? To answer these questions, we should ensure that we can adequately measure current workflows, as well as concretely define the performance metrics we aim to improve.

Drafting Text & Ambient Listening Note Generation

Use Case	Patient	Provider	Administrative	Research
Ambient Listening Note Documentation	X	X	X	
Drafting in-basket replies	X	X		
Drafting Discharge or Interim Summaries		X		
Order Composer for inpatient and ambulatory settings		X		
Drafting Prior Authorization requests		X	X	

Summary

These clinician-facing use cases take speech and/or text input to generate text output within Epic (and other EHR systems as well) . They are all available now, are in flight, or are under development by Epic (as well as other third-party vendors). They aim to decrease provider burnout by automating portions of workflows and freeing time for providers to work at the top of their license, thereby increasing provider satisfaction. Patient satisfaction may secondarily increase by reducing response time to In-basket messages and increasing provider availability for patient care. Increased capacity for clinical work would expand access to care at UW Medicine. Published reports from early adopters indicate significantly increased provider efficiency, satisfaction, and work-life balance.

Key Insights

While the potential benefits of these Epic-native use cases are straightforward, the risk of inaccuracy of LLM output and reliance on a terminal human in the loop to review LLM output poses a potential challenge for safe implementation. Close monitoring after implementation will be crucial to determine utilization and ROI metrics.

Use Cases

Ambient listening note documentation

Clinician-Facing

Primary Author(s): Hasan Ahmad

Description: Current applications approved for use with Epic include DAX Express and Abridge. Nabla was also added to the Epic Connection Hub two weeks ago, but the level of integration with Epic and other details are to be determined.

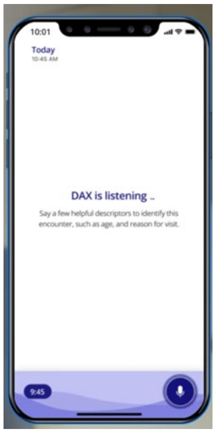
DAX Express

- Technology utilized: iPhone. Android or tablets/iPads not supported.
- Used by providers

Abridge

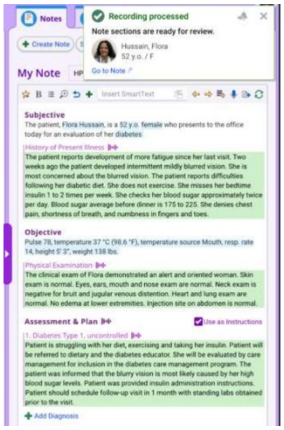
- Technology utilized: iPhone and Android mobile devices
- Currently used by providers, road map for use by all members of care team who interact with patients

Clinicians and other care team members spend considerable time writing notes instead of speaking face-to-face with patients. Alternatively, they pay for expensive virtual or in-person scribes to support documentation efforts.



DAX is listening ..

Say a few helpful descriptors to identify this encounter, such as age, and reason for visit.



Recording processed

Note sections are ready for review.

Hussain, Flora
52 y.o. / F

Subjective

The patient, Flora Hussain, is a 52 y.o. female who presents to the office today for an evaluation of her diabetes

History of Present Illness

The patient reports development of more fatigue since her last visit. Two weeks ago the patient developed intermittent mildly blurred vision. She is most concerned about the blurred vision. The patient reports difficulties following her diabetic diet. She does not exercise. She misses her bedtime insulin 1 to 2 times per week. She checks her blood sugar approximately twice per day. Blood sugar average before dinner is 175 to 225. She denies chest pain, shortness of breath, and numbness in fingers and toes.

Objective

Pulse 78, temperature 37 °C (98.6 °F), temperature source Mouth, resp. rate 14, height 5' 3", weight 138 lbs.

Physical Examination

The clinical exam of Flora demonstrated an alert and oriented woman. Skin exam is normal. Eyes, ears, mouth and nose exam are normal. Neck exam is negative for bruit and jugular venous distention. Heart and lung exam are normal. No edema at lower extremities. Injection site on abdomen is normal.

Assessment & Plan

☑ Use as Instructions

1. Diabetes Type 1, uncontrolled

Patient is struggling with her diet, exercising and taking her insulin. Patient will be referred to dietary and the diabetes educator. She will be evaluated by care management for inclusion in the diabetes care management program. The patient was informed that the blurry vision is most likely caused by her high blood sugar levels. Patient was provided insulin administration instructions. Patient should schedule follow-up visit in 1 month with standing labs obtained prior to the visit.

Patient Impact:

Improved patient experience

- Baptist Memorial (DAX): 82% of providers reported increased provider/patient face time, increased quality of patient interaction; 3.1 min average reduction in time spent interacting with computer while in exam room.
- Novant Health (DAX): 72% reported improvements in patient experience.
- Wellspan Health (DAX): patients thought providers spent less time on computer and that the visit felt more like a personable conversation.

- Rush (DAX): 73% of providers reported increased provider/patient face time, increased quality of patient interaction; 2.7 mins average reduction in time spent interacting with computer while in exam room.
- University of Michigan Health West (DAX): 82% of providers reported increased provider/patient face time, increased quality of patient interaction; 4.8 mins average reduction in time spent interacting with computer while in the exam room.

Improved access to care

- Stanford Healthcare: Providers seeing more patients with DAX.

Improved accuracy of visit documentation
More personalized interaction with the patient during the visit

<u>Provider Impact</u>:

Improved physician satisfaction

- Baptist Memorial: 73% would be disappointed if they no longer had access to DAX
- Novant Health: 67% of providers would be very disappointed if services no longer available.
- Rush: 85% of physicians would be disappointed if they no longer had access to DAX
- University of Michigan Health West: 77% would be disappointed if they no longer had access to DAX

Reduced feelings of burnout/cognitive burden

- Baptist Memorial: 82% reported reduction in feelings of burnout and fatigue or reduced cognitive load
- Novant Health: 85% reported reduced burnout or fatigue; 67% found cognitive burden relief
- Rush: 54% reported reduction in feelings of burnout and fatigue or reduced cognitive load
- University of Michigan Health West; 64% reported reduction in feelings of burnout and fatigue or reduced cognitive load

Decrease in pajama time

- Baptist Memorial (DAX): 193 mins to 46.7 mins
- Novant Health (DAX): 89.48 mins to 51.39 mins
- University of Michigan Health West (DAX): 50.8 mins to 37.7 mins

Decrease in time outside scheduled hours

- Baptist Memorial (DAX): 166.6 mins to 58.4 mins
- Novant Health (DAX): 82.67 mins to 41.48 mins

Less time charting in EHR

- Novant Health (DAX): Average time saved in notes per day decreased from 82.5 mins. to 32.8 mins.; Average time saved in notes per appointment decreased from 7.86 to 2.88 mins. 83% of providers reported saving time on documentation.
- Stanford Health (DAX): minimal changes in time spent
- Texas Health (DAX): 4.6 mins. average time saved per patient encounter; 5.49 hrs. average documentation time saved per week

Improved work-life balance

- Novant Health (DAX): 61% reported better work/life balance

Improved Job satisfaction

- Novant Health (DAX): 84% reported higher job satisfaction

Improved career longevity
More time to catch up on in-basket, refills, letters, etc.

<u>Administrative and Professional Staff Impact</u>:

Improved operational efficiency

- Baptist Memorial: 8.3 minutes saved per encounter --> led to 3 additional patients per day that providers were comfortable seeing while using DAX
- Texas health: 4.38 min average reduction in time spent interacting with computer while in exam room per encounter --> 1.3 average additional patients per day added to provider schedule
- Wellspan: non-Dax users: added 3 appointments per month on average per physician. <60% of DAX utilization: added 5 appointments per month on average per physician. >60% of DAX utilization: added 12 appointments per month on average per physician. Clinicians who relied on DAX for >60% of patient encounters were able to see an average of 9 more patients each month without needing to extend clinic hours compared to non-DAX users.
- Rush: 6.9 minutes saved per encounter; 3 additional patients per day that providers were comfortable seeing using DAX
- University of Michigan Health West: 5.1 mins saved per encounter; 1.9 average additional patients per day that providers were comfortable seeing using DAX

Improved documentation quality

- Baptist Memorial: 73% of physicians state DAX improves documentation quality.
- Novant Health: 61% reported increased documentation quality
- Rush: 58% of physicians stated DAX improved documentation quality
- University of Michigan Health West: 79.5% of physicians stated DAX improved documentation quality

Provider retention
Novant Health: 86% stay at current organization

May result in decrease in number of days to close visits?

- University of Michigan Health West: for providers with >80% utilization, visits closed within 2 days increased from 56.9% to 88.5% while visits closed in 7+ days decreased from 1.9% to 0.5%

<u>Research Impact</u>: Notes might be longer

<u>Financial Impact</u>:

Increase in individual provider ROI

- Texas Health: Providers using Ambient listening technology had a change in wRVU of 176 per provider compared to 64 per provider for providers NOT using Ambient listening technology. This led to a collection difference of $20,173 per provider for those using Ambient listening technology vs $10,289 per provider for those not using Ambient listening technology. This translated into additional compensation of $1,282 per provider per month for providers using Ambient listening vs $417 per provider per month for those not using Ambient listening. Saving of 20-24k per provider in scribe costs. More recently they found that providers using ambient listening technology had 98 more wRVU per provider and this translated into collection difference of $12,733 per provider in a six-month period. On average providers on DAX had additional compensation of $5,469 per provider.

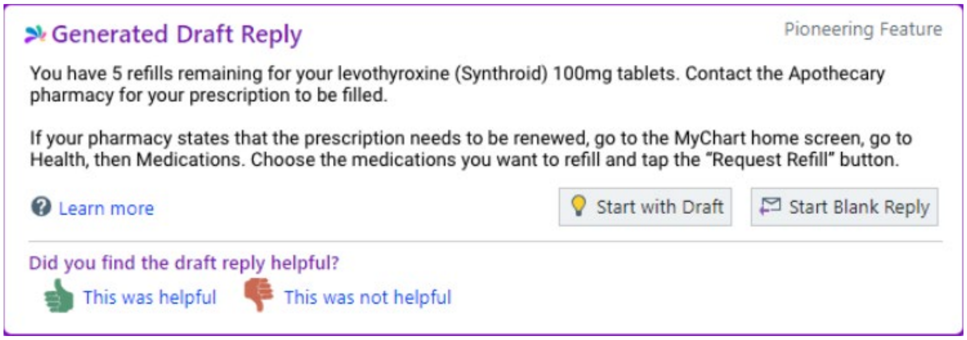
Feasibility (Technical/Operational Impact): Currently available to pilot. Other healthcare systems have implemented. This would be vendor supported (ex. Nuance/Abridge). Examples of other healthcare systems listed above. Need to pick the right physician phenotype for the pilot. Need to determine who would supply iPhones – would this be institution or physician?

Drafted Patient Message Replies

Clinician-Facing

Primary Author(s): Terri Kim

Description: Clinicians and other care team members spend considerable time responding to patient messages in the Inbasket, and the volume of messages received has increased considerably since the pandemic. Increased time spent in the Inbasket is associated with lower provider satisfaction and increased burnout. Epic has an AI tool in production that drafts a response to Inbasket messages in a conversational tone, which the provider can use as a starting message or can elect to compose a message without using the LLM-generated draft.



You have 5 refills remaining for your levothyroxine (Synthroid) 100mg tablets. Contact the Apothecary pharmacy for your prescription to be filled.

If your pharmacy states that the prescription needs to be renewed, go to the MyChart home screen, go to Health, then Medications. Choose the medications you want to refill and tap the "Request Refill" button.

Did you find the draft reply helpful?

Patient Impact: Preprint reports indicate improved documentation efficiency, thus reduced response times from providers would likely be of benefit for patient satisfaction. At this time there are no LLMs that can independently translate messages into other languages, but in the future if translation becomes available this would be of great benefit for DE&I.

Provider Impact: Results from early adopters indicate significant benefit for primary care providers (less voluntary adoption by specialists), with increased provider satisfaction. In simulated models, use of LLM-generated draft replies required editing about 40% of the time, but overall time spent in documentation was improved 77% of the time with low risk of harm from the generated draft messages.

Administrative and Professional Staff Impact: Benefit would be greatest for clinics who already rely on administrative and professional staff to assist with In Basket message review and triage, something that AI could automate, responding to and routing messages to appropriate clinical staff.

Research Impact: Inbasket messages are primarily used for communication between patients and their care teams, with some utility for research into patient-provider communication and measures of provider burnout.

Financial Impact: Similar effort and resources as other Epic-provided generative AI tools.

Feasibility (Technical/Operational Impact): Currently available through Epic to pilot, other institutions have implemented and reported early results.

References:

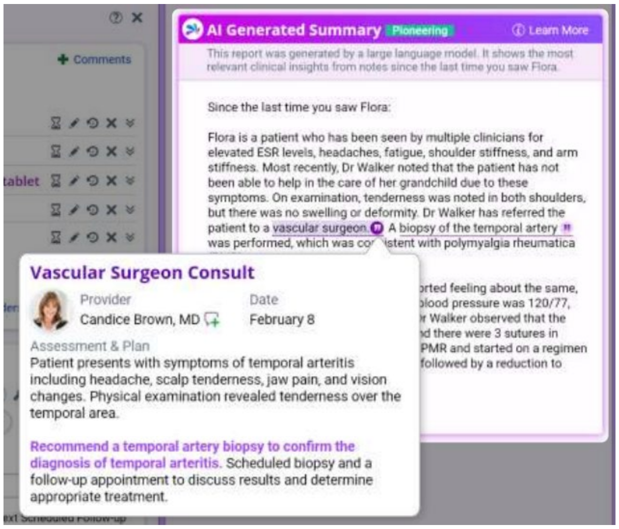
Chen S et al, The impact of responding to patient messages with large language model assistance (preprint) https://arxiv.org/abs/2310.17703
Liu S et al, Leveraging Large Language Models for Generating Responses to Patient Messages (preprint) https://www.medrxiv.org/content/10.1101/2023.07.14.23292669v1
Martinez K et al, Patient Portal Message Volume and Time Spent on the EHR https://link.springer.com/article/10.1007/s11606-023-08577-7
Nguyen O et al, A systematic review of contributing factors of and solutions to electronic health record–related impacts on physician well-being https://academic.oup.com/jamia/article-abstract/28/5/974/6124800

Drafting discharge or interim summaries

Clinician-Facing

Primary Author(s): Terri Kim

Description: Writing summary clinical documentation and narratives is time-consuming but is also critical to patient care, particularly for communication at transitions of care. Epic has an AI tool in pre-production stage that synthesizes information in the patient chart and produces a clinical summary document for the provider to review and then sign as an interim or discharge summary.



This report was generated by a large language model. It shows the most relevant clinical insights from notes since the last time you saw Flora.

Since the last time you saw Flora:

Flora is a patient who has been seen by multiple clinicians for elevated ESR levels, headaches, fatigue, shoulder stiffness, and arm stiffness. Most recently, Dr Walker noted that the patient has not been able to help in the care of her grandchild due to these symptoms. On examination, tenderness was noted in both shoulders, but there was no swelling or deformity. Dr Walker has referred the patient to a vascular surgeon. A biopsy of the temporal artery was performed, which was co· istent with polymyalgia rheumatica

orted feeling about the same,
blood pressure was 120/77,
r Walker observed that the
d there were 3 sutures in
PMR and started on a regimen
followed by a reduction to

Vascular Surgeon Consult

Provider: Candice Brown, MD

Date: February 8

Assessment & Plan

Patient presents with symptoms of temporal arteritis including headache, scalp tenderness, jaw pain, and vision changes. Physical examination revealed tenderness over the temporal area.

Recommend a temporal artery biopsy to confirm the diagnosis of temporal arteritis. Scheduled biopsy and a follow-up appointment to discuss results and determine appropriate treatment.

<u>Patient Impact</u>: Discharge and interim summaries are generally provider-facing more than patient-facing, so the impact to patients may be limited. However, since patients have access to read these documents, satisfaction with transparency around their care may be increased.

<u>Provider Impact</u>: Automated generation of discharge and interim summaries would significantly reduce workload for the primary services who currently write the summaries manually. The output would be of significant benefit for consulting services (e.g., specialists involved in the care of a patient) and support services (e.g., physical therapy, occupational therapy, nutrition, speech therapy, respiratory therapy, and others). Outpatient providers, such as primary care and skilled nursing facility providers, rely on timely discharge summaries for concise summarization of the hospitalization, therapy recommendations, pending results, and follow-up needs. Initial scholarly work indicates that discharge summaries produced by LLMs are of acceptable quality compared to those written by trainees.

<u>Administrative and Professional Staff Impact</u>: Automated summaries offer the potential to be more comprehensive, thus allowing billing capture of the full spectrum of diagnoses and services provided to a patient during their hospitalization.

<u>Research Impact</u>: Automated summaries would be able to store more data in discrete forms, making later queries for research purposes more efficient and effective.

<u>Financial Impact</u>: Epic's roadmap includes automated summaries, but the additional cost and requirement for more analysts is not yet clear.

<u>Feasibility (Technical/Operational Impact)</u>: This feature is currently in pre-production stage with Epic.

<u>References</u>:

Clough R et al, Transforming healthcare documentation: Harnessing the potential of AI to generate discharge summaries https://doi.org/10.3399/BJGPO.2023.0116

Order composer for inpatient and ambulatory settings

Clinician-Facing

Primary Author(s): Jesse Levin

Description: One way to increase provider efficiency would be to use AI/LLM to queue up orders via ambient listening dictation software, which would act like a medical scribe. Orders would be generated and Pended in the Order Composer to be reviewed and signed by the provider. This technology would be most effective if coupled with ambient listening dictation/documentation software.

Current technology includes “Hey Epic,” which must be triggered by the “wake phrase” each time and is limited to specific commands/tasks. Abridge’s roadmap includes order generation to function within the Epic API to draft orders during the patient visit, with release expected in early 2024. Epic is also developing an AI assistant feature to draft orders in response to patient messages based on clinical criteria.

AI technology could potentially be employed for dynamic, automated "Best Practice Alert” like rules to improve patient safety. Examples include propose/pend orders based on the clinical situation: e.g. goal-directed medical therapy based on patient data, venous thromboembolism prophylaxis or infection prophylaxis if on a threshold dose of immunosuppressive treatment, pre- and post-procedure orders and patient instructions, as well as clinical decision support for the most appropriate imaging/diagnostics, healthcare maintenance, and vaccinations.

Patient Impact: There are potential patient safety benefits, including ensuring core measures and prophylaxis are ordered/proposed, keeping up with age-appropriate screening, automatic ordering of follow-up appointments and repeat imaging for incidental radiographic and certain lab findings.

Patients would benefit from the increased efficiency of visits, since prescriptions, diagnostic orders, referrals, and follow-up scheduled could be generated during the visit / faster than if the provider manually entered orders. This could lead to more on-time appointments and less time waiting after visits wrap-up. The After-Visit Summary would contain the relevant orders, prescriptions, referrals, and follow-up plan. Patients would benefit from more face-to-face time with providers, since the provider would not need to spend as much time facing the computer.

Provider Impact: Clinicians would benefit from more efficient workflows in inpatient, outpatient, and Emergency Department settings. This would free up time for patient care, faster documentation, increased job satisfaction, decreased burnout, and improved work-life balance.

Administrative and Professional Staff Impact: Less impact on administrative staff, other than more efficient patient visits, and potential for increased patient visits/throughput.

Research Impact: For patients on research protocols, appropriate labs and diagnostics could be proposed automatically, reducing burden on clinical research staff.

Financial Impact: Financial benefits could include more patient visits related to increased visit and rounding efficiency. If used effectively, costs of the AI technology would eventually be offset.

Feasibility (Technical/Operational Impact): Less feasible until this type of technology is more widely available and able to integrate fully into the Epic EHR – likely would need to be coupled with an ambient listening documentation software package (and In Basket assistants for patient replies) for greatest impact.

References: UGM080 Cool Stuff Breakout - Artificial Intelligence

Drafting of prior authorization requests

Clinician-Facing Administrative

Primary Author(s): Trevor Cohen

Description: Writing prior authorization requests takes considerable clinician time. AI is capable of chart extraction to summarize the medications/treatments previously tried and generate rationale for specific requests. Supporting literature could also be pulled and used to strengthen the requests.

It is worth noting the payers are already using AI software to screen authorization requests and may deny care or treatments without sufficient medical director / clinician oversight. Groups like the American Medical Association are lobbying for regulatory oversight of the use of augmented intelligence for review of patient claims and prior-authorization requests, including whether insurers are using a thorough and fair process.

Patient Impact: Patients would benefit from access to recommended treatments faster due to more efficient prior authorization request submissions.

Provider Impact: This would be a major benefit to clinicians, as prior authorizations are some of the most frustrating and time-consuming tasks. Additional time could be spent on patient care and less after hours work, leading to increased job satisfaction, decreased burnout, and improved work-life balance.

Administrative and Professional Staff Impact: For clinical settings who get assistance from administrative and professional staff for certain aspects of prior authorization, there would be considerable benefits, freeing up time for other work.

Research Impact: Limited impact to research identified.

Financial Impact: Significant financial benefit by reducing low-value time spent by clinicians, which could be spent on higher-value patient care. Implementation costs would be lower if part of an AI application suite capable of chart extraction and summarization.

Feasibility (Technical/Operational Impact): Vendor applications such as Waystar are emerging which use LLMs to automate prior authorization processing, with the potential to integrate with Epic and payor systems. The regulatory landscape for use of AI in prior authorizations is uncertain, with pending litigation against payors alleging unfair denials issued by AI-powered systems.

References: Lee P, Goldberg C, Kohane I. The AI revolution in medicine: GPT-4 and beyond. Pearson; 2023 Apr 14, pg. 173
https://www.fiercehealthcare.com/health-tech/doximity-rolls-out-beta-version-chatgpt-tool-docs-aiming-streamline-administrative

Search, Synthesis & Analytics

Use Case	Patient	Provider	Administrative	Research
1. Augmented/automated determination of patients eligible for clinical trials or studies using clinical data				X
2. Compliance Surveillance and Coding of CPT coding for E&M and Procedural Services; ICD10-CM coding	X	X	X	X
3. Clinical Decision Support including interpretation of patient symptoms & signs; Physical examination findings, Imaging, lab and other diagnostic studies		X		
4. Data integration from multiple sources		X	X	X
5. Problem List Cleanup		X	X	
6. Analytics query development			X	X
7. Semantic search and knowledge extraction for policies, procedures, and job aids.	X	X	X	
8. General purpose HIPAA-compatible generative AI sandbox		X	X	

Summary

This topic area includes use cases that can be broadly placed into two categories: In the **first category**, proposals 1 - 5 represent aspirational opportunities for integrating LLMs with downstream systems such as rules engines, expert systems, or "conventional" predictive machine learning algorithms to overcome historical limitations in processing unstructured data and natural language. These use cases are generally unproven but promising given our current understanding of LLM capabilities; they should be considered high risk but high reward. At a very high level, the risks inherent in this group of projects include: 1) uncertainty of success; 2) high development and implementation costs; 3) stringent (and uncertain) requirements for validation; and 4) likelihood of harm given incorrect output. However, each of the proposals, if successful, have the potential to provide material cost savings, employee satisfaction, and quality of care.

In contrast, proposals in the **second category** of this section (6 - 8) represent more focused and well-understood uses of relatively mature LLM capabilities. In general, most of the risks inherent in projects in the first category are avoided by proposing the use of relatively proven technology in non-patient-care settings in contexts that allow “human in the loop” verification of outputs.

Key Insights

Unlike use cases presented in the section above (“Drafting Text & Ambient Listening Note Generation”), there are no existing products addressing use cases in the first category (1-5), and further consideration would require feasibility studies and an in-depth design process. These proposals may be challenging to evaluate and prioritize given the novelty of LLMs and the uncertain outlook for emerging solutions in Epic or from 3rd parties. Successful implementation of locally developed solutions for these proposals would require development of institutional skills, staff, and infrastructure and would involve substantial effort for development, implementation, and validation. That said, these proposals provide an important context for considering our long-term institutional strategy for development and adoption of large language models.

Proposals in the second category (6-8) are relatively low risk and complexity and can be implemented with minimal cost given existing institutional or departmental resources: prototype applications have been developed in at least one department addressing use cases 7 and 8, and LLM support for software development and analytics (#6) is already in use across the organization at an individual level. These projects provide opportunities to develop skills and infrastructure and gain institutional experience with both technology and governance while providing material benefits to the intended user groups and should be strongly considered for short-term prioritization.

Use Cases

Augmented/automated determination of patients eligible for clinical trials or studies using clinical data

Research

Primary Author(s): Nic Dobbins

Description: Determining patients who meet certain criteria (e.g., “Females over 65 diagnosed with osteoporosis”) is an important step for clinical trials, retrospective studies, hypothesis generation, and so on. The determination of which UW patients meet a given set of criteria is often done manually (i.e., chart review) or electronically using tools such as SlicerDicer or Leaf. Chart abstraction is time-consuming, while existing applications often have steep learning curve, do not have access to certain kinds of data, or are unable to represent complex real-world study needs quickly.

LLM-driven applications have the potential to greatly simplify the task of finding eligible patients by simplifying *inputs* - a user could enter eligibility criteria as free-text, and natural language – as well as *eligibility determination and outputs* – an LLM may generate database queries and analyze clinical notes, then describe in simple language the number of patients found (and their identification) and how this was determined.

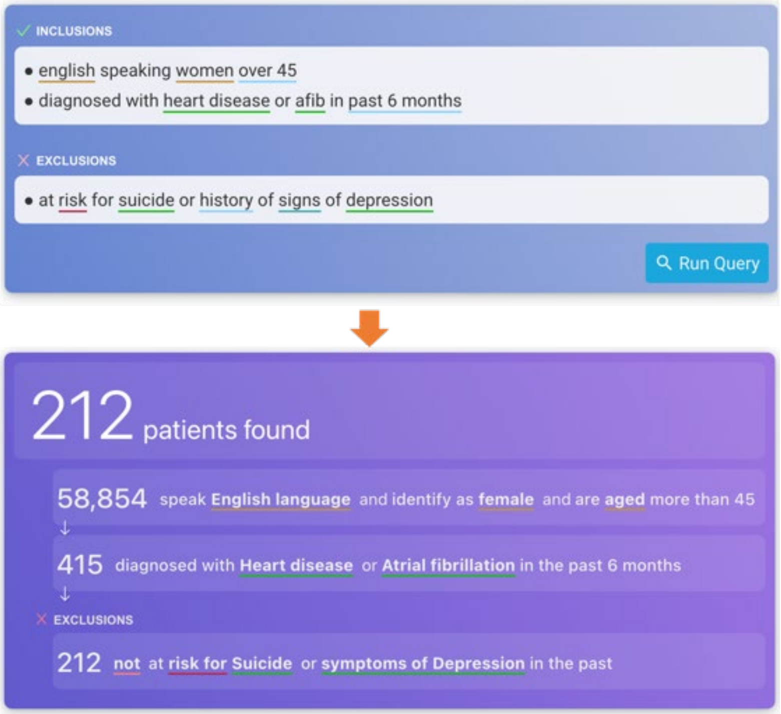
Such cases may thus dramatically quicken and simplify the process of clinical trials recruitment, retrospective studies, etc. and over current non-LLM based methods.

Importantly, LLM-based eligibility determination systems may utilize structured data (diagnosis codes, vitals, etc.), unstructured data (e.g., clinical notes, PDFs), or both. While unstructured data may ultimately contain the most rich and potentially accurate sources of information on patients' health (and thus important for determining eligibility), processing unstructured data quickly and at scale has traditionally proved challenging. Conversely, structured data, such as those within relational databases, may potentially be somewhat less rich but can generally be queried and analyzed much more quickly.

<u>Patient Impact</u>: N/A

<u>Provider Impact</u>: LLM-based cohort discovery systems would likely not be used in the course of patient care but could be used by providers who perform research (see Research Impact).
<u>Administrative and Professional Staff Impact</u>: Administrative staff, such as Research Coordinators, would likely be greatly impacted by such technologies. The time spent determining eligible patients could conceivably be greatly shortened, though some level of validation via chart review may still be necessary.

<u>Research Impact</u>: Widespread use of LLM-based eligibility determination tools may greatly lower barriers to performing research at UW, particularly in terms of time and costs for personnel such as Research Coordinators in recruiting for clinical trials.



✓ INCLUSIONS

- english speaking women over 45
- diagnosed with heart disease or afib in past 6 months

✕ EXCLUSIONS

- at risk for suicide or history of signs of depression

Run Query

212 patients found

58,854 speak English language and identify as female and are aged more than 45

↓

415 diagnosed with Heart disease or Atrial fibrillation in the past 6 months

↓

✕ EXCLUSIONS

212 not at risk for Suicide or symptoms of Depression in the past

Figure 1 Example natural language cohort definition and reporting screenshot from Dobbins et al, 2023

<u>Financial Impact</u>: Positive financial impacts would likely be observed in the research space, as explained above.

<u>Feasibility (Technical/Operational Impact)</u>: In technical terms, as discussed in the Usage Description section, applications utilizing structured data could likely be leveraged relatively quickly, given that UW Medicine has existing relational databases which could be utilized. If a given application required UW data to be transformed for use (i.e., into a specific data model such as OMOP), UW informatics teams have expertise in this as well.

In terms of research workflow, as such applications are typically accessible as secure web applications (or possibly directly within the EHR), they would likely be readily accessible to UW personnel.

UW can potentially leverage both vendor-provided solutions (e.g., it is likely that Epic's SlicerDicer will increasingly include LLM-driven capabilities) as well as internal solutions, such as LeafAI, which is UW intellectual property and the current state of the art.

References: Nicholas J Dobbins et al. LeafAI: query generator for clinical cohort discovery rivaling a human programmer, *Journal of the American Medical Informatics Association*, Volume 30, Issue 12, December 2023, Pages 1954–1964

Compliance Surveillance and Coding of CPT coding for E&M and Procedural Services; ICD10-CM coding

Clinician-Facing Administrative

Primary Author(s): Brad Henley

Description: This application has three primary users: Clinicians, UW Medicine Compliance (administrative), and Coders/Billers (administrative).

Inputting documentation by physicians and other healthcare professionals (HCPs) as free text or in a templated/structured format and allowing natural language processing/LLMs to determine Current Procedural Terminology (CPT) & International Classification of Diseases (ICD10) codes for the Evaluation and Management (E&M) level of services (e.g. office and hospital clinic visits), the procedural service(s) and the diagnosis(es) will simply and expediate coding of professional and hospital services and improve accuracy and compliance with CPT and ICD10 coding rules. As coding and billing are currently manual (chart abstraction by coders) and expensive processes, using LLMs should standardize coding, reduce expenses and capture additional services which may go unreported.

When used to monitor compliance, LLMs may identify more easily issues with billing (unbundling, over coding/under coding, etc.). The potential to automate the audit processes should result in an expense reduction as this process is manual and time consuming currently.

An application example: Jane, a patient with Type 2, diabetes mellitus had an outpatient visit for new onset cough with fever and chills which was ascribed to pneumonia. We need to generate CPT & ICD10 codes for the physician's professional office services and the associated diagnosis(es). An LLM would review the physician's documentation and generate the assignment of CPT and ICD-10 codes which would then be submitted to the patient's health insurer for reimbursement.

Patient Impact: There is little patient impact, though CPT and ICD10 codes used for billing are seen by patients and may be challenged if inaccurate.

Provider Impact: Physician coding for E&M services is a dissatisfier for many HCPs as coding rules change frequently and many do not feel competent or educated sufficiently in coding rules and their nuances. Instead, a majority would rather rely on UWP and hospital billing staff who have coding certification(s) to code their services. LLMs or natural language processing to determine accurately HCP and hospital services would increase HCP satisfaction while simultaneously improving compliance.

Currently, HCPs code their own outpatient E&M services whether performed in a hospital clinic setting or in a free-standing office environment (e.g., UW Primary Care Network). For hospital-based E&M and procedural services whether performed on the hospital wards or in the Operating Rooms/Procedural suites (e.g. GI endoscopy suite, Cardiac Catheterization lab, etc.) HCP documentation is read manually and coded by UWP's professional billers/coders for reimbursement of professional services and by hospital billers/coders for reimbursement of hospital services.

Administrative and Professional Staff Impact: Use of LLMs by coding staff and compliance administrators should simplify and expedite monitoring and auditing of coding/billing compliance. Given the manual and time-consuming auditing processes, automating and standardizing front-end coding/billing may permit a reduction in resources currently allocated to compliance monitoring with a concomitant reduction in expenditures.

Research Impact: CPT and ICD10 codes are used frequently to identify patient cohorts used in clinical and administrative research. Improving coding accuracy will improve the specificity of these independent variables used in research.

Financial Impact: Automating coding would likely result in a modest improvement in revenues and a concomitant reduction in expenses as described above.

Feasibility (Technical/Operational Impact): There is only a little published on the use of LLMs for coding of medical encounters using CPT codes though there are some publications referencing ICD-10 codes since this is an international system. Prior to LLMs, rule-based natural language processing has been implemented and monetized commercially. More than a decade ago, UW Medicine explored a product from LingoLogic. Their medical coding unit was acquired by Cerner and the name changed to Cerner Oracle. Their legacy product was used for a trial at the FHCRC in their Blood and Bone Marrow Transplant Unit with "remarkable results." Solutions for automated coding that do not make use of LLMs have been more extensively evaluated, and it remains unclear as to whether LLMs will offer an advantage in this context, especially given the risks associated with confabulation. However, "Intelligent Coding" along these lines is listed amongst Epic's Generative AI offerings in development, with a target date of May 2024.

I am not a member of any Epic workgroup, though it is likely that Epic (the Electronic Medical Record software used currently by UW Medicine) may be exploring the use of LLMs for these processes. Currently, if their rules are followed, documentation can be extracted from Epic, loaded into natural language processing software to achieve the desired coding automation results.

Using LLMs for these purposes is thus technically feasible and should have a large operational impact. Doing so would likely impact our relationship with coders and the union representing these employees.

References: Finneas Catling[1], Georgios P Spithourakis[2], Sebastian Riedel, **Towards Automated Clinical Coding**. Int J Med Inform. 2018 Dec:120:50-61.
Thomas H Payne[1], Aidan Garver-Hume, Sheila Kirkegaard, Jamie Sweeney, Michael Ash, K K Kailasam, Candace L Hall, Mika N Sinanan. **Natural language processing improves coding accuracy.** MGMA Connex. 2011 Oct;11(9):15-7.
Pei-Fu Chen[1,2], Ssu-Ming Wang[1], Wei-Chih Liao[1], Lu-Cheng Kuo[3], Kuan-Chih Chen[1,4], Yu-Cheng Lin[5,6], Chi-Yu Yang[7,8], Chi-Hao Chiu[9], Shu-Chih Chang[10], Feipei Lai. **Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning.** JMIR Med Inform. 2021 Aug 31;9(8):e23230.

Hang Dong[1 2], Matúš Falis[3], William Whiteley[4], Beatrice Alex[3 5], Joshua Matterson[6 7], Shaoxiong Ji[8], Jiaoyan Chen[9], Honghan Wu. **Automated clinical coding: what, why, and where we are?** NPJ Digit Med. 2022 Oct 22;5(1):159.

Clinical Decision Support including interpretation of patient symptoms & signs; Physical examination findings, Imaging, lab and other diagnostic studies

Clinician-Facing Research

Primary Author(s): Robert Doerning

Description: Since their introduction in the 1970s, Clinical Decision Support Systems (CDSS) have aimed to improve healthcare delivery by “enhancing medical decisions with targeted clinical knowledge, patient information, and other health information”(1). They are frequently broken into two major categories, knowledge-based and non-knowledge-based systems (Fig 1), with the latter category accommodating machine learning systems that do not require manual development of sets of rules for systems to follow deterministically. Large language models fall would typically fall under non-knowledge-based models and could use patient symptom, physical examination findings, and diagnostic testing as a semi-structured or unstructured data source.

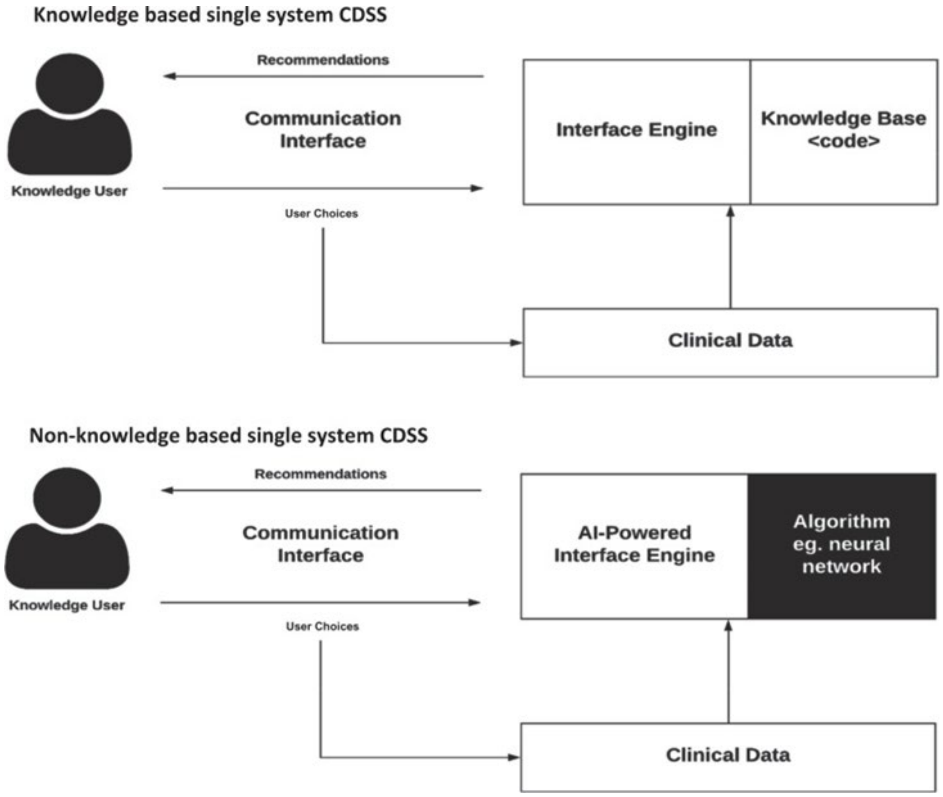


Fig 1: Knowledge vs non-knowledge based CDSS. Knowledge based systems use rules and retrieve data to evaluate the rule and produce an output. Non-knowledge-based systems use AI tools (such as LLMs) to interpret clinical data to produce an output.

LLMs offer specific advantages to CDSS through many potential applications including:

- Comprehensive Data Integration: LLMs have the ability to synthesize a vast array of medical literature, patient records, and clinical guidelines, providing a consolidated and up to date source for analysis and decision-making. Traditional knowledge based CDSS must be updated and maintained based on current medical literature and practice guidelines, a time intensive process.
- Enhanced Diagnostic Accuracy: By processing extensive datasets, they aid in accurate interpretation of patient symptoms, lab results, imaging, and other diagnostic findings, minimizing diagnostic errors.
- Real-time Decision Support: Immediate access to up-to-date information assists healthcare providers in making informed decisions swiftly, especially in critical scenarios.
- Personalized Medicine: These models can analyze patient-specific data to suggest tailored treatment plans based on individual health profiles and histories, optimizing care outcomes. LLMs can be used in an asynchronous manner both before and after the immediate care episode has occurred (Fig 2)(2).

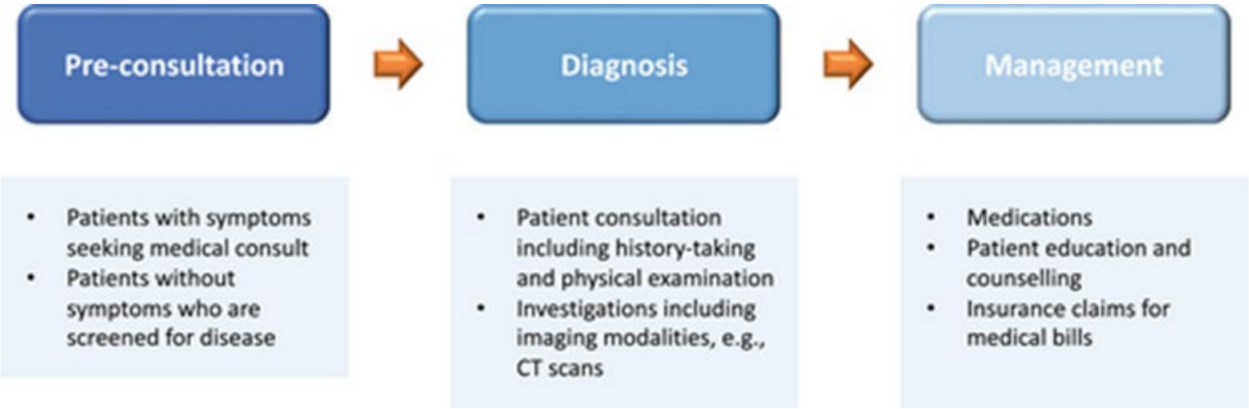

Pre-consultation
Diagnosis
Management
• Patients with symptoms seeking medical consult
• Patients without symptoms who are screened for disease
• Patient consultation including history-taking and physical examination
• Investigations including imaging modalities, e.g., CT scans
• Medications
• Patient education and counselling
• Insurance claims for medical bills

Fig 2: Phases of the patient care encounter where LLMs can be applied to CDSS

It will be important to frame any discussion of LLM applications to CDSS within the Five "Rights" of CDS (3):

1. The right information (e.g., evidence-based guidance)
2. To the right people (i.e., clinical care team—including the patient)
3. Through the right channels (e.g., electronic health records, patient portal, other means)
4. In the right intervention format (e.g., prompts, order sets, alerts)
5. At the right points in the workflow (i.e., for decision-making or action)

Patient Impact: Implementing large language models for Clinical Decision Support can significantly benefit patients through:

- Timely and Accurate Diagnosis: Access to large amounts of semi-structured and unstructured patient level information and diagnostic results can lead to higher quality CDSS and provider personalized care plans (Fig 3) (4)
- Reduced error rate: LLMs have the ability to refine currently implemented CDSS to uncover specific patient subgroups who may be harmed through application of knowledge-based CDSS (4)
- Empowered Patients: LLMs may allow for the creation of patient facing CDSS where a patient can communicate with the chatbot, explain their symptoms, and using non-knowledge based CDSS

(with some degree of provider oversight), receive a suggested next step. This may help reduce unnecessary visits or direct patients to more appropriate care areas.

Provider Impact:

Patients and providers will have some overlapping areas where LLMs will impact CDSS, primarily when it comes to time and accurate diagnosis and reduced error rates. Providers may also benefit from:

- Decreased alert fatigue: it is estimated that 90% of alerts from CDSS are overridden or ignored which contributes to provider dissatisfaction with the EHR and burnout (4)
- Reduced CDSS administrative maintenance: current best practices require group of clinicians to periodically review CDSS in production to ensure that they are performing as expected and are based on current best practices. LLM derived CDSS can use non-knowledge-based approaches that require less maintenance or continually refer to trusted guidelines

Administrative and Professional Staff Impact: As CDSS are primarily patient care centered tools, administrative and professional staff will have little to rare interactions with LLM based CDSS.

Research Impact: While CDSS are primarily used for clinical care, given the improvements that LLMs may afford in development of these tools, there may be improvement in more accurate patient cohort creation. Additionally, if LLM based CDSS have the ability to decrease inappropriate alerts this may result in higher quality data collection for cohort prospective model creation.

Financial Impact: The global CDSS market is quite broad, estimated to be worth $1.7 billion in 2023 and is expected to grow at a compound annual growth rate of 7.5% from 2023-28. There are many vendors in the space offering a variety or rule-based and AI supported products. There are also a number of open source LLMs (Llama, Claude, MPT, Falcon, and Vicuna) that could be trained and tested on local data. Vendor-based solutions offer the advantage of seamless integration with the EHR and timely updates, decreasing the need for UW-IT resources at a large financial cost.

Feasibility (Technical/Operational Impact): AI based CDSS have been around from almost as long as CDSS were being used so the field is quite mature. Provider adoption of AI based CDSS has been less than widespread due to many factors including but not limited to, concerns around provider autonomy, bias in AI systems, and the black box nature of many AI algorithms. LLMs may be able to address some of these issues through their use of common language. There currently exists a robust governance structure at UW Medicine for the development and implementation of clinical CDSS. it is not unreasonable to use open source LLMs to improve current CDSS in production prior to looking at vendor based solutions. One problem with using a vendor based solution is, given that this is a relatively new market, many companies are bought out or change direction/scope which adds a layer of operational uncertainty. Amazon's recent acquisition of Health Navigator and the closure of Olive AI are good examples. Waiting for robust tools from EHR companies like Epic and Cerner may be prudent while developing homegrown solutions.

References:

1. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.

2. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: Development, applications, and challenges. Health Care Sci. 2023 Aug;2(4):255–63.

3. Osheroff JeromeA, Teich J, Levick D, Saldana L, Velasco F, Sittig D, et al. Improving Outcomes with Clinical Decision Support [Internet]. 0 ed. HIMSS Publishing; 2021 [cited 2023 Dec 3]. Available from: https://www.taylorfrancis.com/books/9781498757461

4. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc. 2023 Jun 20;30(7):1237–45.

Data integration from multiple sources

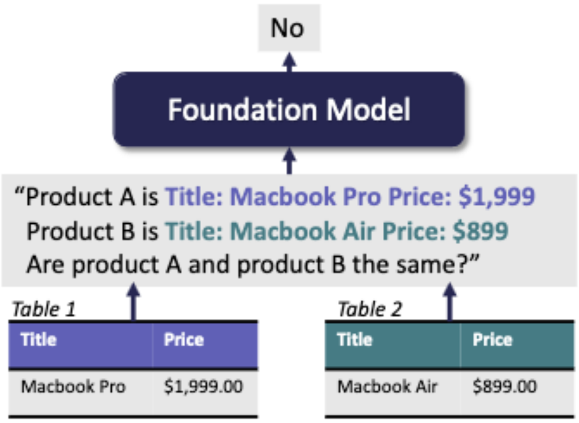
Clinician-Facing | Professional Staff | Administrative | Research

Primary Author(s): Nic Dobbins

Description: Relevant patient medical information can often be siloed across multiple systems, such as legacy EHR systems, other health systems, research databases, or unintegrated ancillary systems. Important relevant information could be utilized if these systems were better integrated, such as patient reported outcomes or historical health information.

LLMs have been demonstrated to be capable of semantically aligning and synthesizing data from various types of data, including tabular sources (e.g., relational database tables) and unstructured documents. These capabilities could be highly useful in cleaning and harmonizing heterogeneous data source across UW.

For example, when Northwest Hospital joined UW Medicine, significant human and financial resources were spent in identifying and aligning patients with medical records across both NWH and UW Medicine. If such an effort were to be done today, it could potentially be performed at significantly less manual human effort and cost by using an LLM to systematically compare the two sources, record by record.

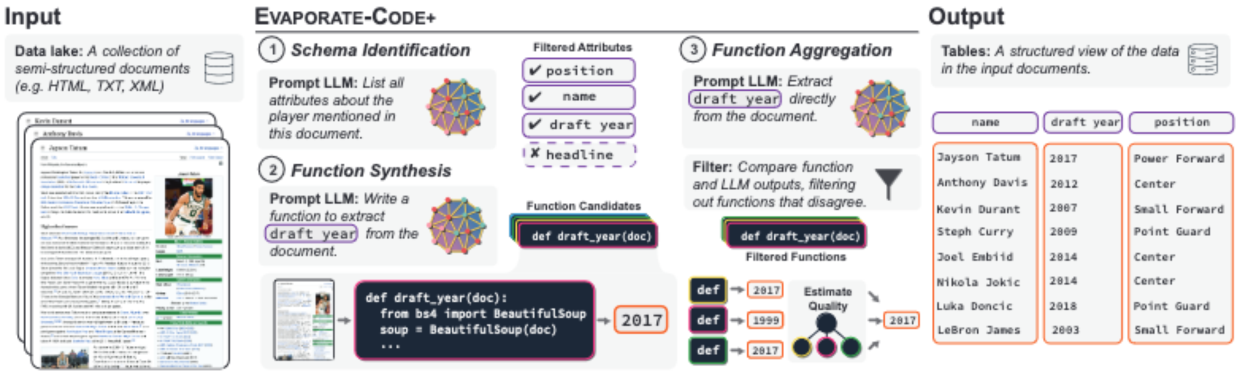


"Product A is Title: Macbook Pro Price: $1,999
Product B is Title: Macbook Air Price: $899
Are product A and product B the same?"

Title	Price
Macbook Pro	$1,999.00

Title	Price
Macbook Air	$899.00

Figure 2 An LLM can address entity matching tasks using prompting. Database table rows are serialized into text and passed to LLM with the question "Are A and B the same?" A similar strategy could be used to map patients, visits, or other data across multiple data sources. From Narayan et al, 2023.


Input
Data lake: A collection of semi-structured documents (e.g. HTML, TXT, XML)
Kevin Durant
Anthony Davis
Jayson Tatum
EVAPORATE-CODE+
1 Schema Identification
Prompt LLM: List all attributes about the player mentioned in this document.
Filtered Attributes
✔ position
✔ name
✔ draft year
✘ headline
2 Function Synthesis
Prompt LLM: Write a function to extract draft year from the document.
Function Candidates
def draft_year(doc)
2017
3 Function Aggregation
Prompt LLM: Extract draft year directly from the document.
Filter: Compare function and LLM outputs, filtering out functions that disagree.
def draft_year(doc)
Filtered Functions
def
2017
def
1999
def
2017
Estimate Quality
2017
Output
Tables: A structured view of the data in the input documents.

```
def draft_year(doc):
  from bs4 import BeautifulSoup
  soup = BeautifulSoup(doc)
  ...
```

name	draft year	position
Jayson Tatum	2017	Power Forward
Anthony Davis	2012	Center
Kevin Durant	2007	Small Forward
Steph Curry	2009	Point Guard
Joel Embiid	2014	Center
Nikola Jokic	2014	Center
Luka Doncic	2018	Point Guard
LeBron James	2003	Small Forward

Figure 3 Unstructured text, PDFs, or arbitrary relational data in heterogenous data lakes can be synthesized into structured information by using an LLM both as an information extraction platform as well as by executing LLM-generated code and parsers. Synthesized information can then be output into harmonized database tables or other structures. From Arora et al, 2023.

Patient Impact: This application would likely have minimal direct impact to patients, except in (likely rare) cases of critical medical information being absent from our EHR but present in some other accessible, integratable form.

Provider Impact: N/A

Administrative and Professional Staff Impact: For non-research work, this application would likely not significantly impact administrative and professional staff. An exception to this may be incorporation of patient information from Washington State resources, such as immunization records or vital statistics.

Research Impact: UW researchers have innumerable research datasets which exist outside of the EHR (e.g., in REDCap). Within UW Medicine, legacy and ancillary datasets are also stored within data lakes such as our enterprise data warehouse (DAWG) - many but not all of which can be readily linked to corresponding records within the Epic EHR. In both of these cases, LLM-based applications capable of automatically analyzing and aligning patient data across multiple sources could simplify and hasten many research efforts.

Financial Impact: Direct positive financial impact to UW Medicine may be minimal, except in possible future cases requiring large scale integration with outside data sources (such as the integration of NWH in the recent past, as discussed).

Feasibility (Technical/Operational Impact): While we are unaware of current vendor solutions specifically for this application, it is likely in the near future viable commercial products will be available. At the time of this writing, most LLM-based applications for this appear to be in the research realm (see References below).

Because open-source model such as Llama 2 may perform quite well at this task and given UW's available expertise and resources, it is possible (and in the near-term most likely) that such work would be conducted internally.

References:

- Narayan, Avanika, et al. "Can foundation models wrangle your data?." *arXiv preprint arXiv:2205.09911* (2022).
- Hegselmann, Stefan, et al. "Tabllm: Few-shot classification of tabular data with large language models." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023
- Arora, Simran, et al. "Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes." *arXiv preprint arXiv:2304.09433* (2023).
- Chen, Zui, et al. "Symphony: Towards natural language query answering over multi-modal data lakes." *Conference on Innovative Data Systems Research, CIDR*. 2023.
- Fernandez, Raul Castro, et al. "How Large Language Models Will Disrupt Data Management." *Proceedings of the VLDB Endowment* 16.11 (2023): 3302-3309.

Problem List cleanup and consolidation

Clinician-Facing Administrative

Primary Author(s): Robert Doerning

Description: A patient's problem list is an essential piece of the current and past medical history, "providing a comprehensive and accessible list of patient problems in one place" (1). They represent a list of illnesses, injuries, and other factors that affect a patient's health, usually identifying the time of occurrence and resolution. Currently, no single standard exists for the structure of problem lists. Meaningful use requirements state that "the provider must maintain an up to date problem list of current and active diagnoses based on ICD-9-CM or SNOMED-CT for 80% of patients, and 80% of all patients have to have at least one coded problem as opposed to their entire problem list coded" (2). The most common standard for interoperability when it comes to the problem list is the Continuity of Care Document, developed by ASTM International's Standard Specification for the Continuity of Care Record (E2369-05) and HL7. The CCD specifies SNOMED CT as the terminology standard for use in defining problems. While the CCD is generally accepted as the standard format for exchanging basic patient information, including the problem list, workflows for updating and maintaining the problem list are more challenging. Given that patients interact with many different care areas across many different health systems and EHRs, there is no 'patient master list' or overall owner of the problem list, resulting in many different disparate version of the problem list, often resulting in confusion both for the provider and patient. Using a LLM to automatically scan both local and outside records to add key problems to the Problem List, auto-resolve time-limited diagnoses, reduce duplicates, merge similar diagnoses, select most specific and appropriate diagnosis, optimize capture of most appropriate CPT codes for CMI, LOS, and reimbursement would not only reduce the administrative burden on healthcare providers but also help provide more accurate and timely patient care.

Patient Impact: Implementing large language models for Clinical Decision Support can significantly benefit patients through:

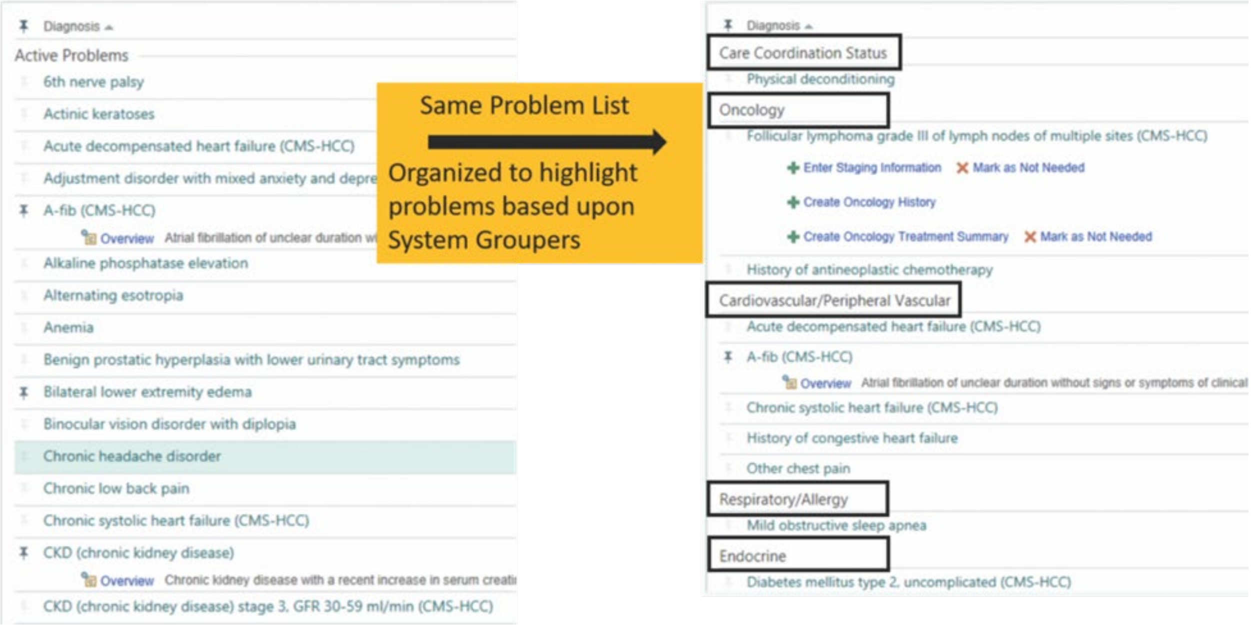
- Enhanced Quality of Care: Accurate problem lists improve care coordination, aiding in precise diagnoses and treatments. Complex patients may pursue care at multiple different organizations and using an LLM to keep an up to date 'master problem list' will improve care coordination across multiple specialties.

- Safety and Satisfaction: Reduced errors in problem lists contribute to increased patient safety and satisfaction by minimizing the risk of misdiagnoses and ensuring a more comprehensive understanding of patient history.

Provider Impact:

- Clinical Care Enhancement: improved workflows and more accurate data from multiple disparate problem lists, improving the efficiency of clinical decision-making and reducing the time spent on administrative tasks.
- Employee Satisfaction: Automation of problem list clean-up reduces the burden on healthcare professionals, improving satisfaction with EHR
- Level of Automation: High levels of automation in problem list clean-up reduce human intervention, minimizing errors and ensuring consistency.

An example of a problem list aggregator developed by Duke University uses 21 system/condition-based groupers using SNOMED-CT hierarchal concepts refined with Boolean logic (Fig 1) (3). This method cannot ingest external data or look for problems lists in unstructured fields.



Diagnosis
Active Problems
6th nerve palsy
Actinic keratoses
Acute decompensated heart failure (CMS-HCC)
Adjustment disorder with mixed anxiety and depre
A-fib (CMS-HCC)
Overview Atrial fibrillation of unclear duration wi
Alkaline phosphatase elevation
Alternating esotropia
Anemia
Benign prostatic hyperplasia with lower urinary tract symptoms
Bilateral lower extremity edema
Binocular vision disorder with diplopia
Chronic headache disorder
Chronic low back pain
Chronic systolic heart failure (CMS-HCC)
CKD (chronic kidney disease)
Overview Chronic kidney disease with a recent increase in serum creati
CKD (chronic kidney disease) stage 3, GFR 30-59 ml/min (CMS-HCC)

Same Problem List

Organized to highlight problems based upon System Groupers

Diagnosis
Care Coordination Status
Physical deconditioning
Oncology
Follicular lymphoma grade III of lymph nodes of multiple sites (CMS-HCC)

History of antineoplastic chemotherapy
Cardiovascular/Peripheral Vascular
Acute decompensated heart failure (CMS-HCC)
A-fib (CMS-HCC)
Overview Atrial fibrillation of unclear duration without signs or symptoms of clinical
Chronic systolic heart failure (CMS-HCC)
History of congestive heart failure
Other chest pain
Respiratory/Allergy
Mild obstructive sleep apnea
Endocrine
Diabetes mellitus type 2, uncomplicated (CMS-HCC)

Figure 4 Duke problem list aggregator and clean up tool

Administrative and Professional Staff Impact: LLMs for problem list cleanup would reduce but not eliminate the need for human intervention. Automating the task would free up administrative staff from the manual process of data entry and validation allowing them to focus on validating the output from the LLM with the patient. Many offices currently use medical assistants or support staff for this task who may not have the medical training necessary for manual problem list reconciliation but would benefit from LLM guidance over the final problem list, improving overall operational efficiency.

Research Impact: While problem list cleanup would primarily be a tool to help in the clinical setting, accurate problem lists could help with patient cohort identification allowing for more accurate data collection and predictive model development.

Financial Impact: Integration of an LLM tool for problem list clean up and aggregation would require a significant investment both financially to purchase and/or develop the tool as well as sizeable IT commitment for implementation and maintenance. While it is possible to develop an in-house tool, given the complexity of the problem, it seems more appropriate to pursue a vendor based solution. Implementation costs may be offset through reduced administrative costs, more accurate and appropriate patient care, and overall better coordination among medical teams.

Feasibility (Technical/Operational Impact): Problem list cleanup is technically a very challenging problem and a relatively novel application of LLMs in a healthcare environment. There are a few large vendors in this space that offer this service as part of their AI platforms including but not limited to: Health Catalyst's Healthcare.ai platform and Apixio. Like other use cases, a potential threat to using vendor-based solution is, given that this is a relatively new market, many companies are bought out or change direction/scope which adds a layer of operational uncertainty. For example, SyTrue was bought by ClaimLoqiq which was then merged with Apixio. Epic Systems, the current EHR vendor for UW Medicine, is currently developing their own LLM based problem list clean up tools. These are currently still in a beta testing phase and the timeline for integration into a production environment is likely many months away. The benefit of using Epic release tools is we have more control over what goes into production, and we can co-develop the tool with Epic. Some of the downsides are, we have less control over the pricing structure when partnering with Epic and many of the generative AI and LLM tools are available only on the most up-to-date version of Epic and UW Medicine has historically been on Epic environments 2-3 versions behind the current release.

References:

1. Blondeau C. Pocket glossary of health information management and technology. 2nd ed. Chicago, Ill.: American Health Information Management Association; 2010.
2. Centers for Medicare & Medicaid Services (CMS), HHS. Medicare and Medicaid programs; electronic health record incentive program. Final rule. Fed Regist. 2010 Jul 28;75(144):44313–588.
3. Problem List Aggregator and Clean Up Tool [Internet]. Available from: https://dihi.org/project/3771/

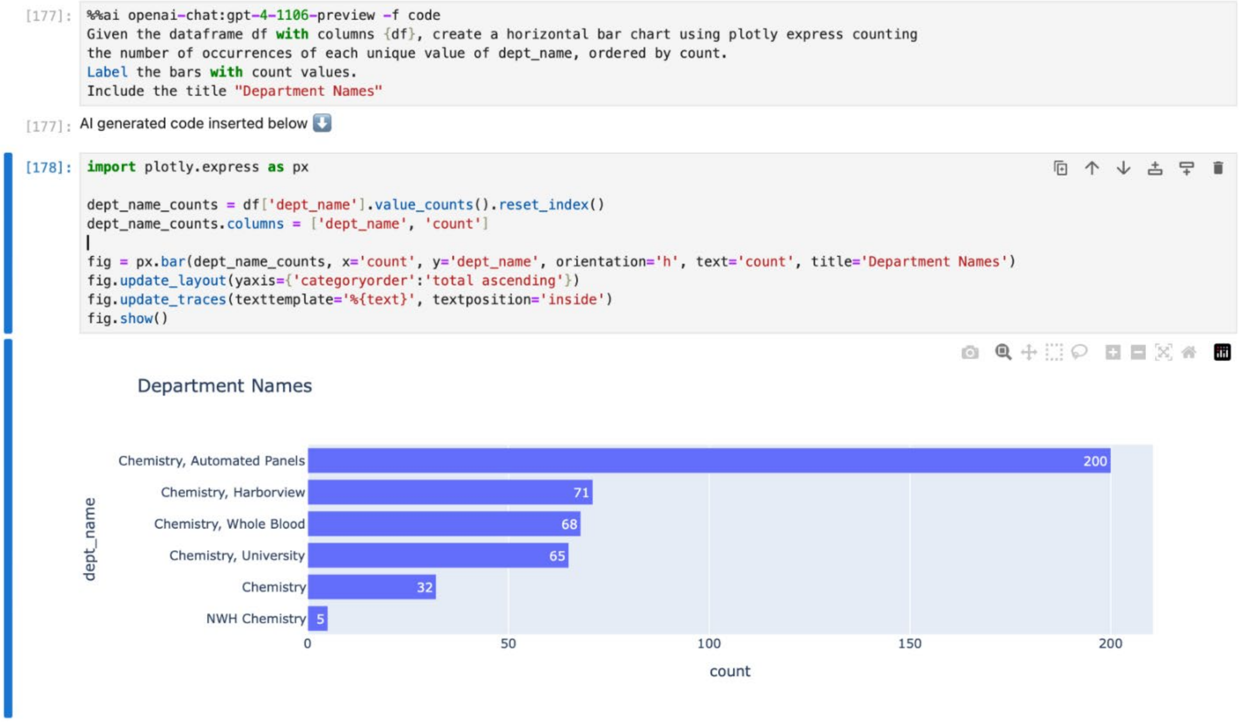
Analytics query development

Clinician-Facing Administrative

Primary Author(s): Noah Hoffman

Description: This proposal considers opportunities for enhancing analytics for both analysts and non-technical users (e.g., providers and administrative staff). Support for programming and software development (of which analytics can reasonably be considered a subset) using LLMs is increasingly mature and widespread. In its most generalized form, such support can take the form of standalone Chat-style user interfaces without direct integration with an editor or programming environment (e.g., see examples using ChatGPT; note that OpenAI Codex has been superseded by Chat Completions).

Generalized language models may also be integrated into programming environments widely used for analytics, e.g., via the jupyter-ai extension for JupyterLab (see figure below). Although graphical UI-driven environments such as Tableau and SlicerDicer provide powerful abstractions, the flexibility of queries and visualizations can be limited, and integration of external tools and packages (e.g., for statistical analysis or modeling) is typically not readily available. LLM integration can help offset the higher barrier to entry that developer-focused analytics environments present to beginners: in response to natural language queries, both general and special-purpose LLMs can write code, explain language features, and even execute queries.



```
[177]: %%ai openai-chat:gpt-4-1106-preview -f code
       Given the dataframe df with columns {df}, create a horizontal bar chart using plotly express counting
       the number of occurrences of each unique value of dept_name, ordered by count.
       Label the bars with count values.
       Include the title "Department Names"
```

[177]: AI generated code inserted below ⬇

```
[178]: import plotly.express as px

       dept_name_counts = df['dept_name'].value_counts().reset_index()
       dept_name_counts.columns = ['dept_name', 'count']
       |
       fig = px.bar(dept_name_counts, x='count', y='dept_name', orientation='h', text='count', title='Department Names')
       fig.update_layout(yaxis={'categoryorder':'total ascending'})
       fig.update_traces(texttemplate='%{text}', textposition='inside')
       fig.show()
```

Figure 1. Example of ChatGPT integration with JupyterLab and jupyter-ai. User input is in the first code block and includes a reference to the input data; the code and plot below were generated entirely in response to the prompt.

Beyond generalized language models, services such as GitHub Copilot are fine-tuned for software development and are intended to be used via integrations with code editors, such as Microsoft Visual Studio Code. LLMs can provide interactive support for a wide variety of programming languages, including SQL, Python, and R, dramatically improving productivity and lowering barriers for learning and applying language features. Considering possible support for analytics in DAWG/DEEP in particular, GitHub Copilot offers integrations with native MS SQL Server database clients. The richest possibilities for analytics support are likely to be provided by systems that allow the structure and content of data to be shared along with a prompt (examples of this can be found in commercial and open source products).

For end users of Epic, generative AI integration with SlicerDicer is available for Cosmos users via a new and developing feature called SideKick. Although many advanced features of SlicerDicer are not yet supported, filters, details, and measures can be defined using natural language queries. LeafAI (described in “Automated Determination of Patients Eligible for Clinical Trials or Studies using Clinical Data”) offers similar features using natural language queries and will likely be available for use at UW Medicine in the near future.

Patient Impact: No direct patient impact.

Provider Impact: Integration of generative AI with SlicerDicer, or similar functionality implemented in a locally developed application, such as Leaf, would lower barriers to providers performing self-service queries.

Administrative and Professional Staff Impact: Any productivity improvements for our limited analytics staff would be a great improvement. Access to assistive technologies for analytics and software development is widespread and is becoming normative in many environments, and it is likely that a lack of availability of these tools would result in a competitive disadvantage for attracting and retaining talent.

Research Impact: Similar to above: Research analytics would be a key area benefiting from integration with generative AI models.

Financial Impact: See Feasibility below.

Feasibility (Technical/Operational Impact): There are few barriers to introducing assistive technologies for analysts and administrative staff at an individual level. Open source programming environments with existing integrations such as VS Code and Jupyter Notebooks/JupyterLab are freely available; VS Code is already available in some environments (e.g., the TRT virtual environment that provides access to DAWG). The primary risk is the exposure of sensitive data for certain integrations, but this could be mitigated by using HIPAA-compatible endpoints hosted on Azure (for OpenAI models) or self-hosted alternatives (e.g., the Llama family of models). Generative AI capabilities in Epic would be available on Epic's timeline and constrained by the extent of native functionality, but presumably the effort for implementation would be extremely low. Alternatively, support for natural-language queries using LeafAI will likely be available to UW investigators in the near future. LeafAI is capable of incorporating data in queries outside of Epic, and the LLM models used can be fine-tuned and adapted within UW, which may be potentially faster and more accurate than those provided by Epic. Overall, there are many options to explore with low risk, at low cost, and offering significant potential benefits.

References: See inline hyperlinks.

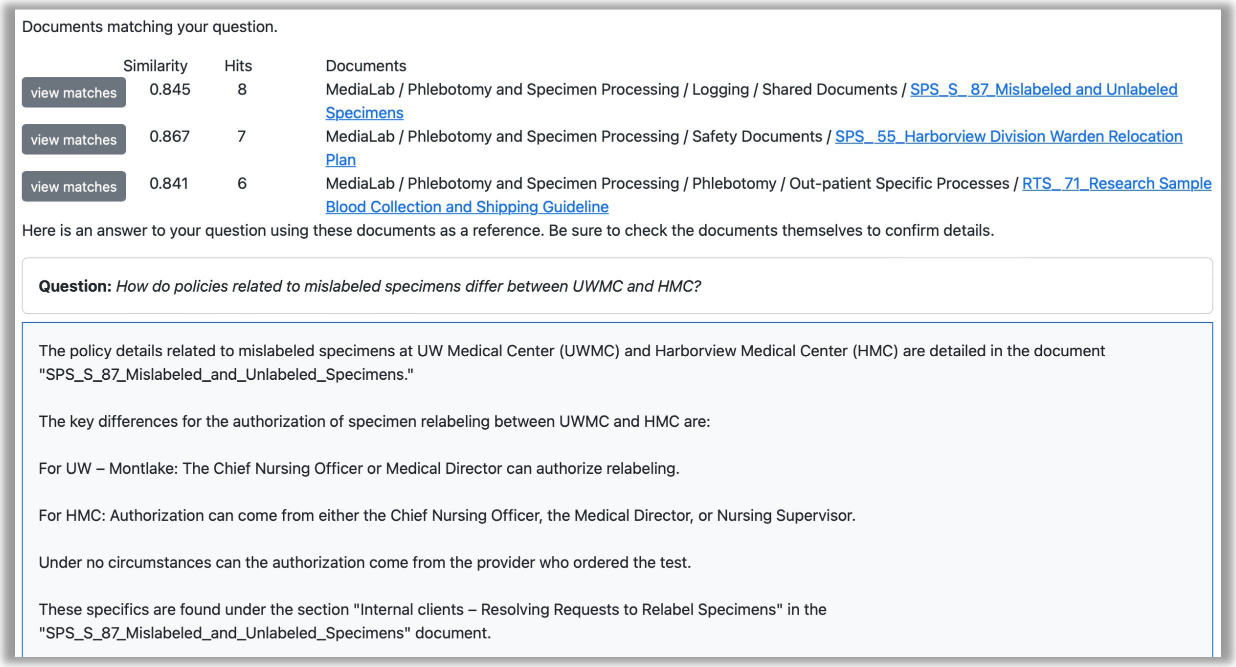
Semantic search and knowledge extraction for policies, procedures, and job aids.

Clinician-Facing | Professional Staff | Administrative

Primary Author(s): Noah Hoffman

Description: Generalized language models (like ChatGPT) can readily answer fact-based questions or provide a summary or synthesis when provided reference text from which to derive a response. But because the quantity of input text that can be included with a single prompt is limited, a strategy called "Retrieval Augmented Generation" (RAG) is commonly used to identify a smaller number of relevant documents from a much larger collection to include as context along with a prompt.

RAG consists of a two-step process in which a prompt or question is first used to search a library of documents; a second prompt is then submitted that includes the results of the first search as context. A particularly powerful mechanism for identifying relevant documents is known as "semantic search," which uses numerical representations of text ("embeddings") to retrieve content that is contextually related to the search query, even if the content does not contain the exact keywords initially submitted. One of the strengths of this strategy is that the model can identify the specific source documents used along with its response, allowing the user to verify the accuracy of the response. An LLM integration can also assist with maintenance and curation of a document library (e.g., "find all documents addressing policies related to specimen labeling and identify any points of disagreement" or "how do transfusion guidelines differ between HMC and UWMC?").



Documents matching your question.

	Similarity	Hits	Documents
view matches	0.845	8	MediaLab / Phlebotomy and Specimen Processing / Logging / Shared Documents / SPS_S_ 87_Mislabeled and Unlabeled Specimens
view matches	0.867	7	MediaLab / Phlebotomy and Specimen Processing / Safety Documents / SPS_ 55_Harborview Division Warden Relocation Plan
view matches	0.841	6	MediaLab / Phlebotomy and Specimen Processing / Phlebotomy / Out-patient Specific Processes / RTS_ 71_Research Sample Blood Collection and Shipping Guideline

Here is an answer to your question using these documents as a reference. Be sure to check the documents themselves to confirm details.

Question: *How do policies related to mislabeled specimens differ between UWMC and HMC?*

The policy details related to mislabeled specimens at UW Medical Center (UWMC) and Harborview Medical Center (HMC) are detailed in the document "SPS_S_87_Mislabeled_and_Unlabeled_Specimens."

The key differences for the authorization of specimen relabeling between UWMC and HMC are:

For UW – Montlake: The Chief Nursing Officer or Medical Director can authorize relabeling.

For HMC: Authorization can come from either the Chief Nursing Officer, the Medical Director, or Nursing Supervisor.

Under no circumstances can the authorization come from the provider who ordered the test.

These specifics are found under the section "Internal clients – Resolving Requests to Relabel Specimens" in the "SPS_S_87_Mislabeled_and_Unlabeled_Specimens" document.

Figure 5 Screenshot from an application in development implementing RAG for a knowledge base of > 300 policies, procedures, and job aids for DLMP Specimen Processing and Phlebotomy.

It may be particularly useful to combine currently disparate document libraries with an overlapping scope to ensure that requirements arising from different parts of the organization can all be accounted for.

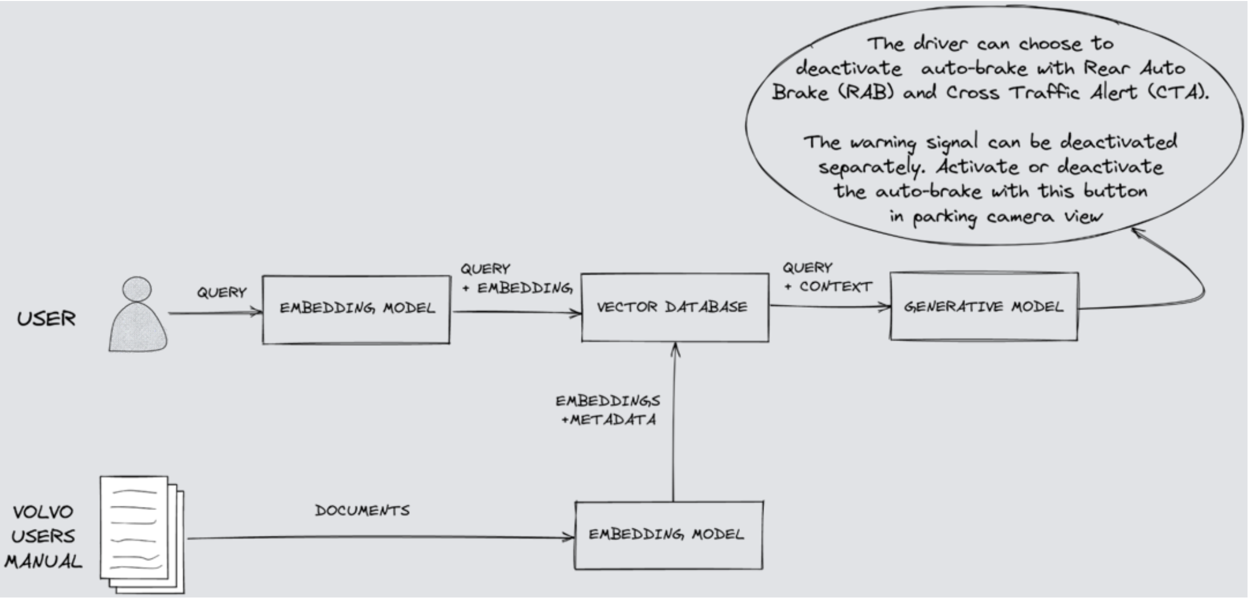

The driver can choose to deactivate auto-brake with Rear Auto Brake (RAB) and Cross Traffic Alert (CTA).
The warning signal can be deactivated separately. Activate or deactivate the auto-brake with this button in parking camera view
USER
QUERY
EMBEDDING MODEL
QUERY + EMBEDDING
VECTOR DATABASE
QUERY + CONTEXT
GENERATIVE MODEL
EMBEDDINGS +METADATA
VOLVO USERS MANUAL
DOCUMENTS
EMBEDDING MODEL

Figure 6 Illustration of knowledge discovery using RAG. From https://www.pinecone.io/learn/retrieval-augmented-generation/

This use case represents a tremendous opportunity for the many internal knowledge bases used within UW Medicine. In particular, our current repository of clinical policies, procedures, and job aids (tip sheets) in EHR Hub and our library of institutional and hospital-based policies have rudimentary search, and it is difficult to rapidly and confidently use these resources to answer questions. Another framing of this proposal might be to provide a generalizable platform for hosting knowledge bases for departments and administrative service areas throughout the organization; a specific example might include an employee searching for information regarding childcare resources or information related to retirement. In general the pipeline, infrastructure, and user interface should be agnostic of content.
Note that this use case is focused primarily on operational and administrative knowledge bases; the use of RAG and other techniques for delivery of medical knowledge is a major area of research (https://arxiv.org/abs/2303.01229) but is not considered here (see "Semantic Search and Synthesis" in this document). However, one important motivation for this proposal is to provide a relatively low-risk and low-cost context for the development of skills and infrastructure that could be readily extended into other domains.

Patient Impact: As proposed, none. It would of course be possible to develop patient-facing knowledge bases for operational and administrative questions (e.g., comprising content currently hosted on public UW Medicine web properties), but this would involve greater risk and effort. For example, we would expect higher thresholds for validation and more demanding requirements for the user interface.

Provider Impact: Improved information retrieval from operational and administrative knowledge bases might be expected to support activities related to training, troubleshooting, compliance, and quality improvement.

Administrative and Professional Staff Impact: Similar to provider impact. Leveraging semantic search to identify relevant information from knowledge bases might reduce the administrative overhead necessary to organize and maintain those libraries.

Research Impact: This proposal does not directly target research applications but could be expected to provide some benefits, depending on the knowledge base; for example, a corpus of documents related to research compliance might be expected to facilitate IRB applications or other administrative activities.

Financial Impact: See Feasibility: costs would vary depending on the solution, but in general would be expected to be low.

Feasibility (Technical/Operational Impact): A wide variety of commercial and open-source solutions exist for creating an LLM-enabled knowledge base. Cloud providers with an existing footprint in UW Medicine offer platform-as-a-service (PAAS) products (Azure AI Search, Amazon Bedrock) that provide the necessary components for building the back end of a RAG-enabled query engine; these PAAS offerings typically have the benefit of extensive examples and documentation. A large number of open-source tools also exist, as do third-party vendors. Given the relative maturity of the technology, abundance of options, and proliferation of instructions and examples, the overall cost and difficulty of developing even self-hosted solutions with existing infrastructure and personnel should be low.

References:

- https://learn.microsoft.com/en-us/semantic-kernel/memories/vector-db
- https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview

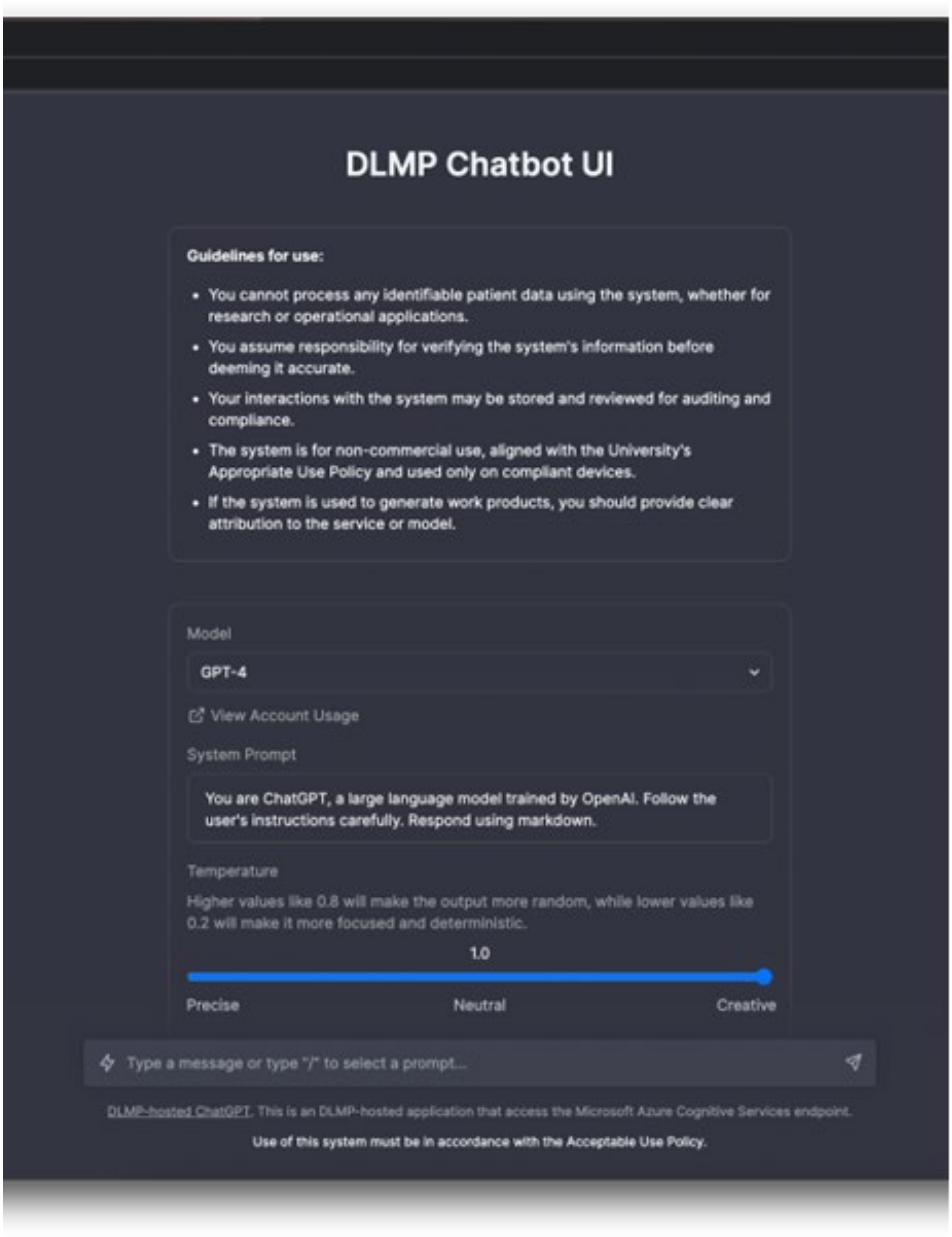
- https://docs.aws.amazon.com/sagemaker/latest/dg/jumpstart-foundation-models-customize-rag.html
- LangChain (open source): https://www.langchain.com/use-case/retrieval
- https://www.pinecone.io/learn/retrieval-augmented-generation/

General purpose HIPAA-compatible generative AI sandbox

Clinician-Facing Administrative

Primary Author(s): Noah Hoffman

Description: The objective of this proposal is to provide HIPAA-compliant access to generative text models in a self-hosted sandbox environment for general use by approved users. The initial scope would be the Department of Laboratory Medicine and Pathology along with invited users from other UW Medicine departments. Potential use cases for an interactive Chat interface are almost too numerous to count. Exploration of these use cases in our healthcare environment will be an ongoing project. We believe that it is important to provide hands-on experience with Large Language Models to our trainees, faculty, and staff so that we can all develop our intuition about their capabilities and limitations and more effectively participate in decision-making about their evaluation and appropriate uses. Even for use cases not explicitly involving sensitive data, a self-hosted service minimizes the risk of inadvertent exposure of PHI or other protected information when using consumer-facing services. As we gain experience with this application, we could consider the infrastructural and administrative requirements for a more broadly available service.



DLMP Chatbot UI

Guidelines for use:

- You cannot process any identifiable patient data using the system, whether for research or operational applications.
- You assume responsibility for verifying the system's information before deeming it accurate.
- Your interactions with the system may be stored and reviewed for auditing and compliance.
- The system is for non-commercial use, aligned with the University's Appropriate Use Policy and used only on compliant devices.
- If the system is used to generate work products, you should provide clear attribution to the service or model.

Model

GPT-4

System Prompt

You are ChatGPT, a large language model trained by OpenAI. Follow the user's instructions carefully. Respond using markdown.

Temperature

Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic.

1.0

DLMP-hosted ChatGPT. This is an DLMP-hosted application that access the Microsoft Azure Cognitive Services endpoint.

Use of this system must be in accordance with the Acceptable Use Policy.

Figure 3. Screenshot of the prototype application user interface (currently in use, but not yet approved for sensitive data).

A prototype application has been implemented by DLMP (and is currently made available to departmental users for non-sensitive data); this is awaiting the completion of a risk assessment. The existing user interface closely resembles the ChatGPT web application, which supports an open-ended dialogue between the user and a choice of language models.

Patient Impact: N/A

Provider Impact: In addition to alignment with the institution's goals in education and innovation, this project offers numerous immediate benefits to providers and staff.
Example use cases from our proposed appropriate use policy:

1. Medical Education
   - A resident uses an LLM to investigate the underlying molecular mechanisms of a rare disease she encountered during rounds. She cross-verified the information obtained from the AI with textbooks and scholarly articles before incorporating the knowledge into her clinical practice and research.
   - A clinical fellow uses an LLM to extract a list of genetic variants from a patient note and reformat them as a table to present at a clinical conference. The fellow is careful to verify the accuracy of the extracted variants.
2. Clinical/operational
   - A faculty member prepares for a tumor board conference by generating a case history from patient notes and emailed communications. All details included in the summary are verified from primary sources before the presentation.
   - A resident uses an LLM to summarize a long email chain discussing a clinical case; after verifying the accuracy of the summary, it is used to provide background information to a faculty member for input.
3. Administrative
   - An administrative staff member uses an LLM to automate responses to frequently asked questions about laboratory procedures, policies, and schedules.
   - An administrative assistant uses the service to prepare meeting minutes from notes and transcripts, providing a concise summary and identifying action items and assignments.

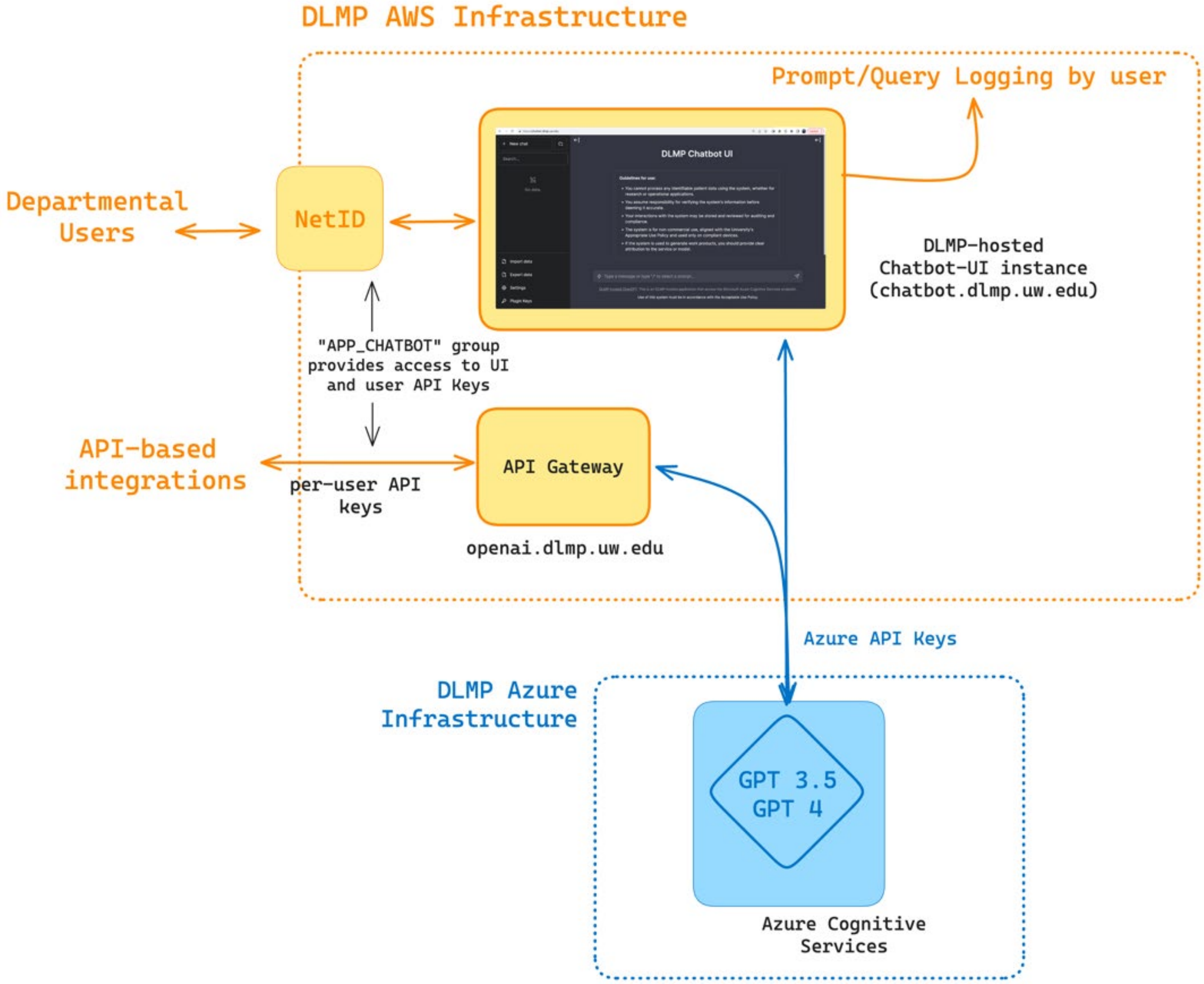
Administrative and Professional Staff Impact: Similar to provider impact.

Research Impact: The generative AI sandbox is not primarily intended to support research applications that involve prompts including sensitive data, but the administrative requirements can be evaluated if such consideration does not prove an impediment to implementation.

Financial Impact: A prototype application has been fully implemented using departmental resources. Ongoing usage fees from AWS (for application hosting) and Azure (for the LLM services) are expected to be on the order of several hundred dollars per year.

Feasibility (Technical/Operational Impact): The DLMP Chatbot web application uses Azure-hosted ChatGPT models to provide access to LLMs for DLMP staff, faculty, and trainees while ensuring infrastructure compatibility for the processing of PHI. Providing access to a centrally-managed LLM in this way makes it easy for users to “do the right thing” and mitigates privacy risks associated with users misconfiguring environments, or using public-facing interfaces to LLMs. This interface is integrated into our departmental process for identity and access management, ensuring secure user access through

NetID + 2FA, facilitated by an integration with AWS Cognito. Access is granted based on membership in a specific UW group, with user profiles managed by existing departmental on-/off-boarding processes to ensure appropriate access and security. Furthermore, signature of and adherence to an appropriate use policy is mandated, with a mechanism to prompt for a new signature on major updates.



Guidelines for use:

- You cannot process any identifiable patient data using the system, whether for research or operational applications.
- You assume responsibility for verifying the system's information before deeming it accurate.
- Your interactions with the system may be stored and reviewed for auditing and compliance.
- The system is for non-commercial use, aligned with the University's Appropriate Use Policy and used only on compliant devices.
- If the system is used to generate work products, you should provide clear attribution to the service or model.

Use of this system must be in accordance with the Acceptable Use Policy.

Figure 4. System diagram for the prototype application.

References:

- Data, privacy, and security for Azure OpenAI Service: https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy

Translation

Use Case	Patient	Provider	Administrative	Research
Chart summarization		X		
Patient-facing care or plan summarization	X			
Point-of-Care language translation	X	X		
Chart Abstraction		X		X
Interactive patient intake	X	X	X	
Patient-centered pathology reports	X	X		
Revenue Cycle – Customer Service inquires via MyChart			X	
Revenue Cycle - Denial Appeals			X	
Revenue Cycle - Coding			X	

Summary

Translation refers to use cases which involve the transformation, summarization, or abstraction of language from one format or audience to another. Such use cases may include translation from English to a non-English language, for example, but also summarization of information (such as that from a patient's chart) to a non-medical audience.

Key Insights

Our use case investigations lead us to a number of tentative suggestions;

1. Opportunities that have potential for improving patient-centered care and revenue cycle improvements should be prioritized.
2. LLMs can help address current limitations in resources (e.g., human translators) and serve as a force multiplier for otherwise laborious tasks.
3. There are likely good opportunities to use existing vendors to integration translation-related LLM tools.
4. LLMS can fundamentally change the way we offer care for patients by meeting patients "where they are" in a language and vocabulary which matches their needs and preferences.

Chart summarization

Clinician-Facing

Primary Author(s): Keith Eaton

Description: Helps identify key information from a patient's chart (e.g. what's changed since I last saw this patient) to bring someone quickly up to speed. This could include chart notes, procedures, summary of imaging and lab studies)
I am unaware of an existing application for chart summarization for medical professionals. There are several companies which offer this for other purposes (legal, insurance claims, billing, etc.) such as: https://www.digitalowl.com/ and others in references.
The goal would be to have a general tool that would work in a variety of situations and is adaptable to the clinician's specific focus. Such tools would likely be piloted with specific use cases then expanded to more general cases.

Patient Impact: A thorough review of the chart enables the provider to understand patients' health status and history. This can also include psychosocial information which could be used as a potential starting point for the clinician regarding understanding.
Patients greatly appreciate when the provider thoroughly understands their history.
Medical errors could be avoided by using this tool by surfacing important information that is "lost in the noise".

Provider Impact: Clinicians should review all relevant events prior to the patient encounter. This may involve hundreds of pages of medical records and significant time commitment. If this activity could reliably be performed by AI, it could improve the quality of the pre-charting review and result in significant reduction in clinician effort.
Decision making could also be improved with chart summarization tools which obtain needed data for risk calculators or clinical checklists (e.g. 10-year cardiovascular risk, lung cancer screening eligibility).

Administrative and Professional Staff Impact: AI chart summarization could greatly enhance the efficiency of intake staff and nurse navigators, potentially reducing FTE required. Ideally this summarization could also be used to improve coding for billing and risk adjustment for claims and Vizient.

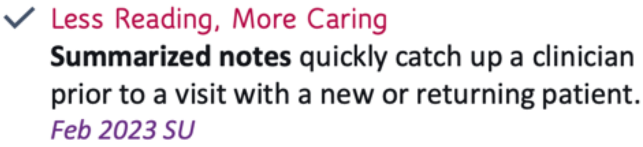
Research Impact: Such tools could inform research (i.e. better understanding of underlying health records for inclusion/exclusion criteria in trials or for a standard of care therapy). These tools could also synergize with abstracting tools or dedicated research tools.

Financial Impact: The effort involved in scoping the project, investigating implementation strategies (i.e., in house vs vendor), deploying in limited use cases and evaluating the technology is considerable.

Feasibility (Technical/Operational Impact): The technology would involve NLP, document scanning with OCR, and integration with Epic and other medical information systems. Ideally the technology would be trainable and flexible for multiple use cases.
The current products (I could find):
https://www.uptech.team/blog/ai-medical-records-summarization

https://www.wisedocs.ai
https://www.width.ai/post/gpt-4-medical-record-summarization-pipeline

Also, there is interest in this application from Epic (according to the one-pager, the checkmark indicates this is a playground feature already available for exploration):



✓ Less Reading, More Caring
Summarized notes quickly catch up a clinician prior to a visit with a new or returning patient.
Feb 2023 SU

(source: Epic Generative AI one-pager)

References:

Searle T et al. Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models. J Biomed Inform. 2023 May;141:104358. Epub 2023 Apr 5. PMID: 37023846.

Moen H et al. Comparison of automatic summarisation methods for clinical free text notes. Artif Intell Med. 2016 Feb;67:25-37. Epub 2016 Jan 21. PMID: 26900011.

Chi EA et al. Development and Validation of an Artificial Intelligence System to Optimize Clinician Review of Patient Records. JAMA Netw Open. 2021 Jul 1;4(7):e2117391.. PMID: 34297075.

Baron RJ. Using Artificial Intelligence to Make Use of Electronic Health Records Less Painful-Fighting Fire With Fire. JAMA Netw Open. 2021 Jul 1;4(7):e2118298. PMID: 34297077.

Patient-facing care or plan summarization

Patient-Facing

Primary Author(s): Trevor Cohen

Description: A LLM would be used to generate a summary of a patient's care or care plan, in language that is easily understood by patients. This is closely related to one of the use cases described in Epic's Generative AI white paper, as "Education they Remember" - personalized patient instructions in plain language. In these use cases, the content to be translated in plain language would be determined by a provider (*provider-initiated*), with the potential for review and approval of the generated summary.

Alternatively, this approach could be used to "translate" parts of the medical record (e.g. accessed via a patient portal) into language that is more easily understood by patients. In this case the patient themselves would select the components of the note they are having difficulty understanding (*patient-initiated*), and there is no designated intermediary available to review the translation concerned.

Plain language summaries of patient records and care plans are important because it is estimated that many patients in the United States have low or limited health literacy, even more so in vulnerable population groups. This may negate the beneficial outcomes (see e.g. Neves et al. 2020) of providing patients with access to their records. Conversely, better understanding of records and plans may lead to better adherence, as is the case with better health literacy (see e.g. Zhang et al. 2014).

Patient Impact: *Provider-initiated* care plan summaries have the potential to improve patient-provider communication, and perhaps adherence to treatment plans, though research is needed for formal evaluation of these outcomes.
Patient-initiated translations of aspects of the clinical notes have the potential to improve patient experience, satisfaction and autonomy. However, as there is no intermediary between the LLM output and the patient, confabulated or misrepresented content may lead to misinformed decision making.

Provider Impact: There is a potential decrease in provider workload, as these summaries would need to be reviewed and edited rather than written up from scratch.

Administrative and Professional Staff Impact: N/A

Research Impact: N/A

Financial Impact: For provider-initiated summaries, if Epic follow through with the proposal below, the cost model may be similar to that proposed for the Outbox Assistant. For patient-initiated summaries, with no commercial product on the horizon this would likely involve considerable internal development efforts to integrate access to a secure LLM with a patient portal used to explore clinical notes.

Feasibility (Technical/Operational Impact): Epic have expressed interest in supporting plain language translational of provider-authored educational material as a near-term LLM use case, though neither a demo nor an anticipated delivery date for one have been specified. Also, it may be possible to accomplish translation of a summary of a component of the record into patient-centered language with a minor modification of the prompts developed for clinical summarization using the Epic Sandbox.

An alternative might involve a locally hosted or secure Azure LLM instance with access to extracted patient notes. This would require considerable development effort, as well as hardware / infrastructure developments. Even with this development it would lack the ready access to structured data and indexed note components that an Epic product could likely leverage.

References: https://qualitysafety.bmj.com/content/29/12/1019.abstract

https://journals.sagepub.com/doi/full/10.1177/1060028014526562?casa_token=wYu1y_LCez4AAAAA%3A3aYStBj5a4Iha45BmrfsaNk2l_f2Xlx3-Y-23FVtuUH-OxixYnjyvGfcjfRqorj7GC4A9GCtxUj9

Point-of-Care language translation

Clinician-Facing Patient-Facing

Primary Author(s): Angad Singh

Description: LLM-driven translation of documents / messages (e.g. translation), with human oversight.

This includes two types of translation: 1. Language translation from English to another language or another language to English 2. Language translation to adjust for word choice based on reading level or other potential parameters that are patient-specific and patient-centered. LLM's have been shown to perform well on both of these tasks, outperforming customized machine translation solutions in translation between languages, and reducing the readability requirements of translated scientific articles.

To date, applications of automated translation tools center around creating translation drafts for patient education that be secondarily reviewed by a human. This is particularly relevant in Epic, who has already described functionality that can automatically generate translated drafts of patient instructions to be reviewed for accuracy by language translators They have ideated about being able to adjust instructions based on patient reading level on the fly (English to English) but have not demo'ed this or articulated clear plans if they will ultimately implement this. This use case could also be potentially considered with other vendors at a future juncture (imagining patient education vendors as an example).

Patient Impact: This will substantially improve access to medical resources and improve patient education, helping patients understand and engage in their healthcare.

Although a substantial segment of our patient population navigates healthcare in a primary language other than English, language-concordant tools are one of our biggest gaps in digital tools. The vast majority of our available patient-facing tools and educational materials are in English only. This gap is even larger when thinking of patient instructions provided by a patient's care team, which are typed on the fly and almost never translated away from English as we do not have system capacity for real-time translation by either in-house interpretation or a 3rd party vendor.

Reading level-based translation is an under-considered use case that would offer tremendous benefit even to those patients who speak English but may not have full comprehension of complex medical jargon and would benefit for more plain-English explanations.

Provider Impact: Providers are currently left with minimal to no options for translation. Many may resort to using a non-approved tool like Google Translate that has no element human oversight element to it. This might allow providers to request a draft be generated for selected text and to send it to a qualified translator who can vet accuracy within a short turnaround time.

Administrative and Professional Staff Impact: Improving communication across the health care system, including everything from scheduling to pre-operative planning to applying for financial assistance would benefit from real-time language translation.

Research Impact: This could be used in research spaces outside of an Epic-centric model to facilitate improved recruitment and communication with patients who navigate healthcare in a language other than English.

Financial Impact: Would allow us to maximize limited in-house translation resources by pre-creating draft translations and allow us to use AI to serve as a force multiplier for translation needs for the segment of our patient population that requires it.

Feasibility (Technical/Operational Impact): English<>non-English translation drafts is very feasible and will be implemented as an available Epic soon. Translation based on reading level is still a future idea in Epic.

This could be applied to non-Epic tools as well and, especially language-based translation drafts, could be implemented with relative ease.

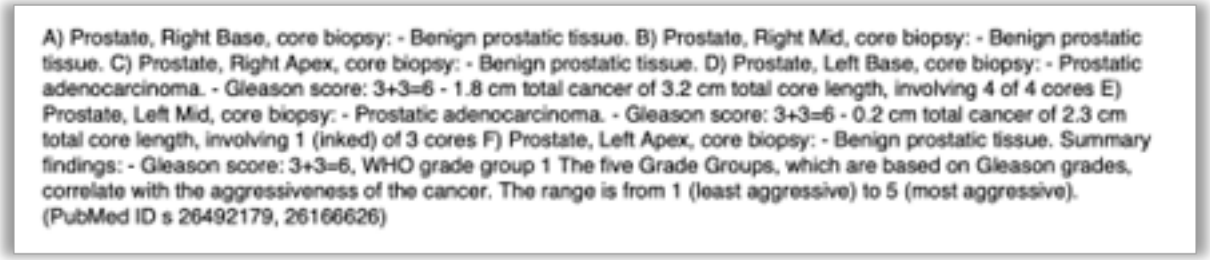
References:
https://aclanthology.org/2023.wmt-1.1/
https://www.sciencedirect.com/science/article/pii/S1532046423003015

Patient-centered pathology reports

Clinician-Facing Patient-Facing

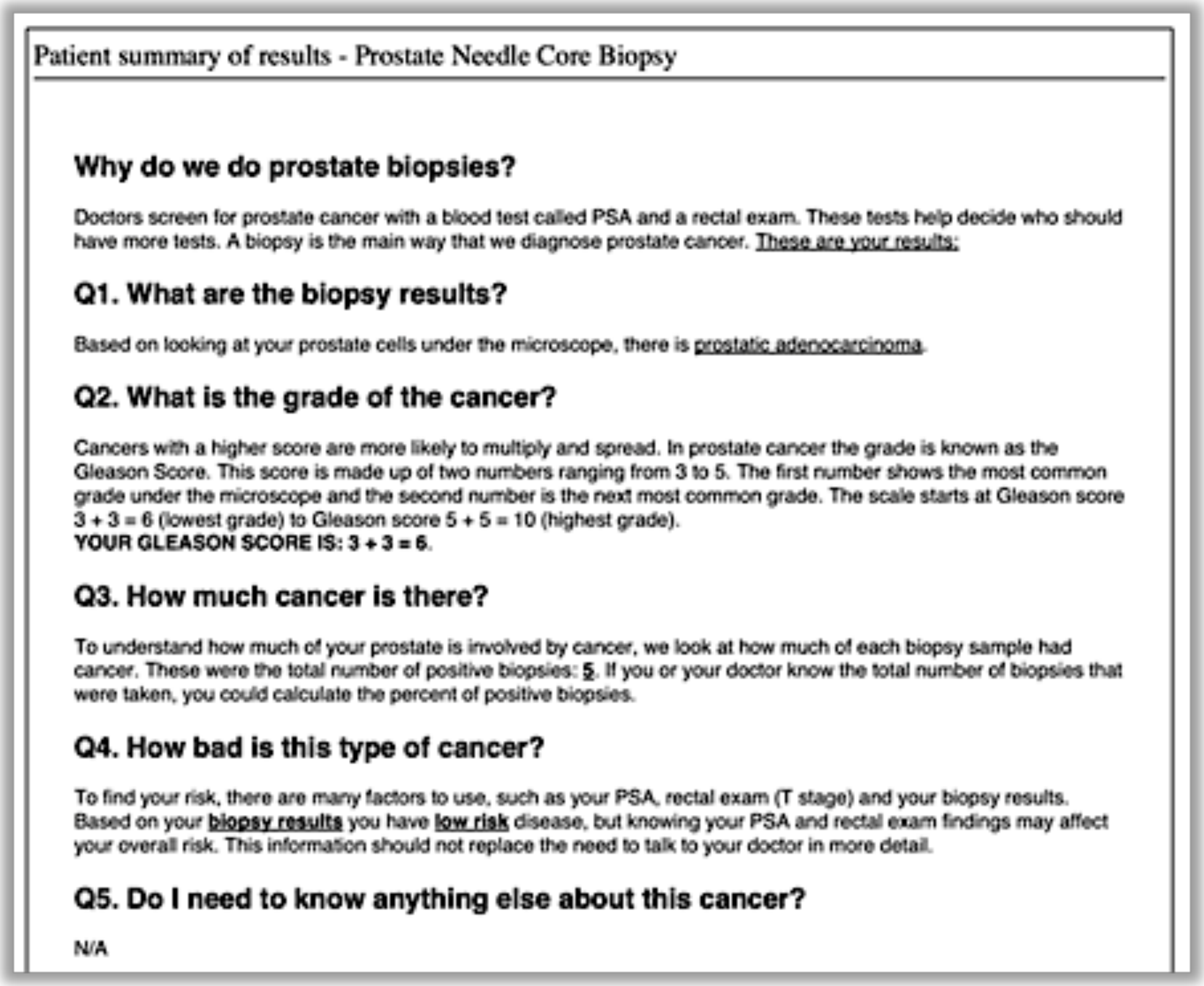
Primary Author(s): Noah Hoffman

Description: Pathology reports are difficult for patients to interpret without assistance, that LLMs could provide. For example, John Gore (Urologic Oncology) was awarded a grant from the Andy Hill CARE Fund for a project titled "**Development of Pathology Translator to Automate Creation of Patient-Centered Pathology Reports for Cancer Care.**" The overarching goal is to create systems and processes that translate diagnostic reports entered into the chart by pathologists into **patient-centered pathology reports (PCPRs)** that focus patients on the key elements foundational to prognosis and treatment decision-making presented in layouts and language that is understandable by patients. Ideally the PCPR becomes part of the medical record that can be shared within the patient portal.



A) Prostate, Right Base, core biopsy: - Benign prostatic tissue. B) Prostate, Right Mid, core biopsy: - Benign prostatic tissue. C) Prostate, Right Apex, core biopsy: - Benign prostatic tissue. D) Prostate, Left Base, core biopsy: - Prostatic adenocarcinoma. - Gleason score: 3+3=6 - 1.8 cm total cancer of 3.2 cm total core length, involving 4 of 4 cores E) Prostate, Left Mid, core biopsy: - Prostatic adenocarcinoma. - Gleason score: 3+3=6 - 0.2 cm total cancer of 2.3 cm total core length, involving 1 (inked) of 3 cores F) Prostate, Left Apex, core biopsy: - Benign prostatic tissue. Summary findings: - Gleason score: 3+3=6, WHO grade group 1 The five Grade Groups, which are based on Gleason grades, correlate with the aggressiveness of the cancer. The range is from 1 (least aggressive) to 5 (most aggressive). (PubMed ID s 26492179, 26166626)

Example input text from a pathology report for a Prostate biopsy.



Patient summary of results - Prostate Needle Core Biopsy

Why do we do prostate biopsies?

Doctors screen for prostate cancer with a blood test called PSA and a rectal exam. These tests help decide who should have more tests. A biopsy is the main way that we diagnose prostate cancer. These are your results:

Q1. What are the biopsy results?

Based on looking at your prostate cells under the microscope, there is prostatic adenocarcinoma.

Q2. What is the grade of the cancer?

Cancers with a higher score are more likely to multiply and spread. In prostate cancer the grade is known as the Gleason Score. This score is made up of two numbers ranging from 3 to 5. The first number shows the most common grade under the microscope and the second number is the next most common grade. The scale starts at Gleason score 3 + 3 = 6 (lowest grade) to Gleason score 5 + 5 = 10 (highest grade).
YOUR GLEASON SCORE IS: 3 + 3 = 6.

Q3. How much cancer is there?

To understand how much of your prostate is involved by cancer, we look at how much of each biopsy sample had cancer. These were the total number of positive biopsies: **5**. If you or your doctor know the total number of biopsies that were taken, you could calculate the percent of positive biopsies.

Q4. How bad is this type of cancer?

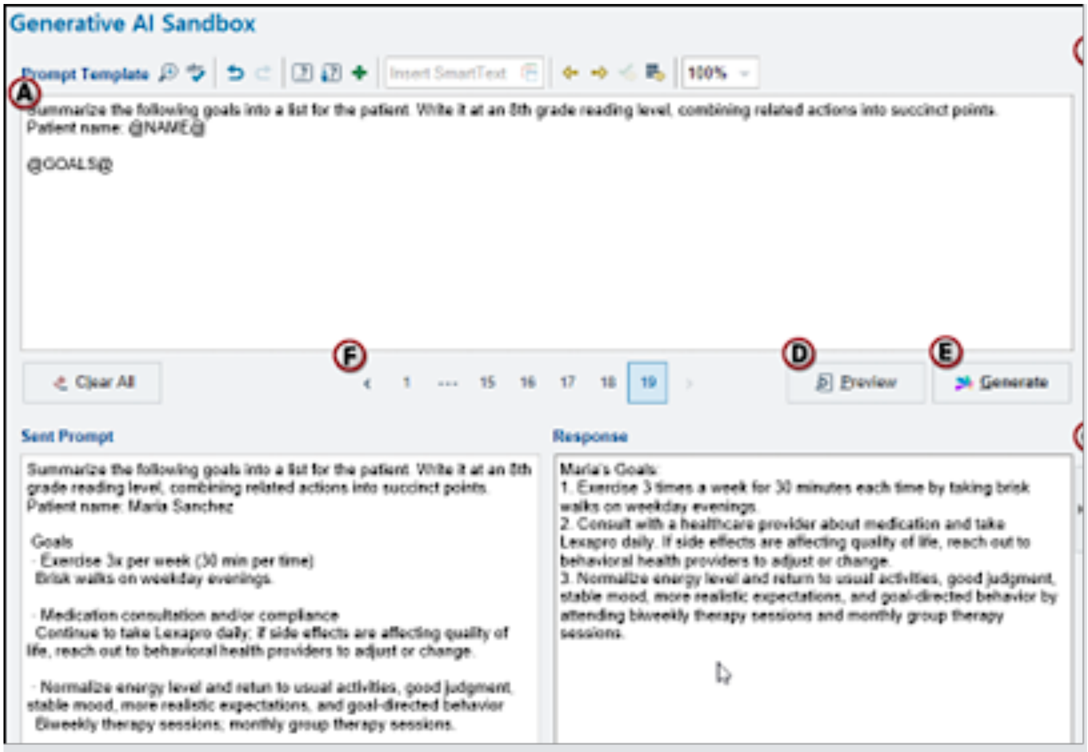
To find your risk, there are many factors to use, such as your PSA, rectal exam (T stage) and your biopsy results. Based on your **biopsy results** you have **low risk** disease, but knowing your PSA and rectal exam findings may affect your overall risk. This information should not replace the need to talk to your doctor in more detail.

Q5. Do I need to know anything else about this cancer?

N/A

Example PCPR derived from the above pathology report for a prostate biopsy.

A prototype application using technology predating the current generation of large language models has been implemented by a vendor partner; a plan for deployment to an infrastructure suitable for prospective validation is being considered. The existing product is being evaluated for translation of prostate cancer reports, but the ultimate goal would be to extend the service for other cancer types. It is likely, however, that translation of pathology reports to PCPRs would be well within the capabilities of generalized large language models already available in the Epic Generative AI Sandbox, particularly if fine-tuning becomes available. Crucially, planned features in Epic would most likely support integration for translation of pathology reports through the use of native features for note-writing.



Summarize the following goals into a list for the patient. Write it at an 8th grade reading level, combining related actions into succinct points.
Patient name: @NAME@

@GOALS@

Summarize the following goals into a list for the patient. Write it at an 8th grade reading level, combining related actions into succinct points.
Patient name: Maria Sanchez

Goals
- Exercise 3x per week (30 min per time)
Brisk walks on weekday evenings.

- Medication consultation and/or compliance
Continue to take Lexapro daily; If side effects are affecting quality of life, reach out to behavioral health providers to adjust or change.

- Normalize energy level and retun to usual activities, good judgment, stable mood, more realistic expectations, and goal-directed behavior
Biweekly therapy sessions; monthly group therapy sessions.

Maria's Goals:
1. Exercise 3 times a week for 30 minutes each time by taking brisk walks on weekday evenings.
2. Consult with a healthcare provider about medication and take Lexapro daily. If side effects are affecting quality of life, reach out to behavioral health providers to adjust or change.
3. Normalize energy level and return to usual activities, good judgment, stable mood, more realistic expectations, and goal-directed behavior by attending biweekly therapy sessions and monthly group therapy sessions.

Example of prompt generation from a note context in the Epic Generative AI Sandbox.

Patient Impact: Quoting material from the proposal receiving the award from the CARE fund:
"…the average American reads at below a basic literacy level, with half of all Americans unable to read a book written at an 8th grade reading level. Similarly, over 30 million US adults have low health literacy, which strongly correlates with overall health: 70% of those with low health literacy report poor overall health. This literacy gap disproportionately impacts individuals of lower socioeconomic status, the elderly, and racial and ethnic minorities. Thus, additional tools are needed to support patients' efforts to understand their new cancer diagnosis that are accessible to individuals of variable general and health literacy."
"Increased patient engagement is associated with improved health outcomes. The concept that a more engaged patient has better health outcomes is not new. However, recent work has shown that efforts to increase patient activation in their health care can improve health outcomes and lower costs of care."
"The content of the pathology reports for these biopsies that is exchanged between clinical disciplines (from pathologist to clinician) is not written with patients as an intended stakeholder. The complex language in these reports can be a barrier to patient understanding of the important elements of their new cancer diagnosis."

Provider Impact: This project would generate PCPRs **with minimal effort** within the context of the EHR.

Administrative and Professional Staff Impact: N/A

Research Impact: Although the primary objective of this proposal is to directly enhance patient care, this specific project is being designed and validated in a research context, so we would expect direct academic benefits to the team involved in the study.

Financial Impact: The expected cost would be low given the use of native Epic features.

Feasibility (Technical/Operational Impact): The proposal is to evaluate existing and planned native Epic features, so the expected effort would be low relative to projects requiring creation and integration of external services. An implementation team would need to be familiar with the build process for features using Generative AI, prompt engineering, and (if available) fine-tuning.

References:

- Mossanen M, Macleod LC, Chu A, Wright JL, Dalkin B, Lin DW, True L, Gore JL. Comparative Effectiveness of a Patient Centered Pathology Report for Bladder Cancer Care. J Urol. 2016 Nov;196(5):1383-1389. doi: 10.1016/j.juro.2016.05.083. Epub 2016 May 19. PMID: 27211289.
- Nayak JG, Scalzo N, Chu A, Shiff B, Kearns JT, Dy GW, Macleod LC, Mossanen M, Ellis WJ, Lin DW, Wright JL, True LD, Gore JL. The development and comparative effectiveness of a patient-centered prostate biopsy report: a prospective, randomized study. Prostate Cancer Prostatic Dis. 2020 Mar;23(1):144-150. doi: 10.1038/s41391-019-0169-7. Epub 2019 Aug 28. PMID: 31462701.
- Verosky A, Leonard LD, Quinn C, Vemuru S, Warncke E, Himelhoch B, Huynh V, Wolverton D, Jaiswal K, Ahrendt G, Sams S, Lin CT, Cumbler E, Schulick R, Tevis SE. Patient comprehension of breast pathology report terminology: The need for patient-centered resources. Surgery. 2022 Sep;172(3):831-837. doi: 10.1016/j.surg.2022.05.007. Epub 2022 Jun 15. PMID: 35715235.

Revenue Cycle – Customer Service inquires via MyChart

Professional Staff Administrative

Primary Author(s): Drew von Eschenbach

Description: Review messages coming to ERC customer service via MyChart to determine what is being requested and prepare a response for customer service team to review prior to submitting to patient.

Patient Impact: This could potentially improve the turnaround time to respond to patient inquiries, therefore improving customer service satisfaction

Provider Impact: N/A

Administrative and Professional Staff Impact: AI can potentially support this customer service function by scanning the incoming message from the patient to determine what is being asked or requested. AI then can use existing information native to Epic to prepare a response to answer the patient’s question or provide the information that has been requested. This will allow staff to work on other activities such as quality assurance or review of documentation.

Research Impact: N/A

Financial Impact: The financial impact is moderate to the organization in terms of the derived value of freeing up valuable staff time to perform other functions.

Feasibility (Technical/Operational Impact): Unknown.

References:

Revenue Cycle – Denial Appeals

Professional Staff Administrative

Primary Author(s): Drew von Eschenbach

Description: When an insurance company denies a claim, there are often times when we want to appeal the denial to pursue payment. This often requires a significant amount of time spent reviewing clinical documentation to support and justify the appeal. The appeal letter then needs to be written and submitted to the insurance company. Generative AI can be leveraged to assist in writing the appeal letter to be reviewed prior to payer submission.

Patient Impact: Patients are frustrated when insurance companies deny their claims, at times requiring the patient to be responsible or limits their ability to have continued similar/related services in future.

Provider Impact: Providers at times are asked to assist in appeal process either with providing supporting documentation or conducting a peer to peer.

Administrative and Professional Staff Impact: Appeals requires significant time and effort to collect, abstract, and prepare. Given the volume of accounts that require an appeal, we are often limited on the number that we can generate given limited staff to conduct the work to generate a satisfactory appeal letter.

Research Impact: N/A

Financial Impact: The financial impact is significant to the organization in terms of overturning a denial or allowing organization to take next steps in dispute resolution process with payers. This is a multi-million dollar opportunity.

Feasibility (Technical/Operational Impact): Although this is an unknown, it is estimated that generative AI could save a significant amount of time and resources in preparing appeal letters and also allow UW Medicine to increase the volume of appeal opportunities. A detailed review is needed to gauge feasibility.

References: Waters, Michael R., Sanjay Aneja, and Julian C. Hong. "Unlocking the power of CHATGPT, artificial intelligence, and large language models: practical suggestions for radiation oncologists." *Practical Radiation Oncology* 13.6 (2023): e484-e490.

Revenue Cycle – Coding

Professional Staff   Administrative

Primary Author(s): Drew von Eschenbach

Description: Review clinical documentation to abstract diagnosis and procedures codes or descriptions to cross walk to ICD-10 diagnosis code and ICD-10/CPT procedure codes.

Patient Impact: N/A

Provider Impact: Providers should just need to document the services that are provided to a patient and allow AI to abstract the documentation to identify all appropriate diagnosis and procedure codes.

Administrative and Professional Staff Impact: AI can potentially support the coding function by scanning documented content to determine if a diagnosis and/or procedure code is present on the information. This will allow staff to work on other activities such as quality assurance or review of documentation.

Research Impact: N/A

Financial Impact: The financial impact is moderate to the organization in terms of the derived value of freeing up valuable staff time to perform other functions.

Feasibility (Technical/Operational Impact): Epic is already working on this technology in the outpatient arena and is expected to have available to clients in the middle of 2024 year.

Augmentation/Automation & Scheduling

Use Case	Patient	Provider	Administrative	Research
Calling patients for appointment scheduling or research data collection	X	X	X	X
Telemedicine/nurse triage call in		X	X	
Improving utilization of clinic appointments and OR block time		X	X	
Improving processes to reduce the need to capture Medicare ABNs and commercial insurers' waivers (to reduce non-coverage determinations) and to improve ABN/waiver utilization when required		X	X	
Training and Education of patients and providers	X		X	

Summary

The augmentation/automation and scheduling use cases allow for large-scale augmentation/automation of repetitive tasks typically requiring specific training or knowledge. The list of use cases in this method group tends to be more aspirational with limited studies and/or commercial products available. The use cases have the potential to be impactful financially and otherwise but are still mostly theoretical. The one exception is the first use case on the list 'calling patients for appointment scheduling or research data collection'. This would automate the standard practice of scheduling an appointment once a referral is received. Humans calling patients is still the standard practice. LLM phone callers are available and can relatively easily schedule visits in an automated way by phone. These calls can describe the appointment, any prep required, directions to the clinic, and answer other questions automatically. Transcripts of the call can then be extracted. Bland.ai is a tool available today that performs these tasks quite effectively.

Key Insights

Experimenting with related but more feasible, less disruptive use cases such as automated phone calls for research use, may be a reasonable short-term option for gaining institutional knowledge of shortcomings and how to deploy. This would be low risk and potentially highly impactful.

Use Cases

Calling patients for appointment scheduling or research data collection, etc.

Administrative Research

Primary Author(s): Sean Mooney

Description: A standard practice after a referral is a call from a clinical scheduling a referral appointment. While MyChart based scheduling is available, a human placing a phone call for clinic to a patient is still the standard practice. LLM phone callers are available and can relatively easily schedule visits in an automated way by phone. These calls can describe the appointment, any prep required, where to go, and answer other questions automatically. Tools such as Bland.ai are close to being able to do this today.

Patient Impact: Quality of care would likely be reduced during the phone call, however, I would argue that care would improve as getting referral calls often takes time (months) and is often inconvenient.

Provider Impact: N/A

Administrative and Professional Staff Impact: This could have a significant ROI as many calls could be avoided.

Research Impact: N/A

Financial Impact: Significant cost savings in reduction in human placed calls.

Feasibility (Technical/Operational Impact): Technical difficulty is high and maturity is low. Feasibility is possible.

References: https://www.bland.ai

Training and Education of patients and providers

Clinician-Facing Patient-Facing

Primary Author(s): Nathan Cross

Description: LLMs are exceptionally well suited to training and education for both patients and providers.

On the provider side, LLMs could be used to modernize static training materials and make them interactive or provide further detail and explanations dynamically. They could also be used for scoring and assessment and targeting training materials to different subgroups. LLMs have also shown the ability to simplify materials to suit those with varying technical or educational backgrounds which would help training to be more concise and relevant. We provide significant training for employees and LLMs could potentially lower costs, develop training materials, augment/automate delivery of trainings, elaborate or consolidate content, and assess understanding.

On the patient side, LLMs could provide automated summarization of chart information incorporating background materials to explain issues to patients in need of guidance. LLMs could also generate clinical guidance materials in preparation for procedures, studies and surgeries. The capability to rewrite to different reading levels could also facilitate appropriate usable information to a broad audience.

Patient Impact: These tools can help our system stay more current and unified resulting in higher quality care delivery which benefits patients

Provider Impact: This can help provide a much more targeted and concise collection of training materials to keep our providers current with a broad collection of training which could increase the quality of care delivered.

Administrative and Professional Staff Impact: This use case would require training of those responsible for training and educational materials so that they are aware of prompt creation best practices and would still require substantial proof reading.

Research Impact: N/A

Financial Impact: Expense of inferencing.

Feasibility (Technical/Operational Impact): Particularly with models that have some clinical material training this use case is likely to be quite feasible in the near term. It would take significant effort for prompt engineering and proof reading to develop the content. The other challenge is for much of this use case there is likely need for multimedia including images and photos that cannot be easily generated. Other AI tools for image generation have not displayed the level of nuance and understanding that LLMs have with text so far.

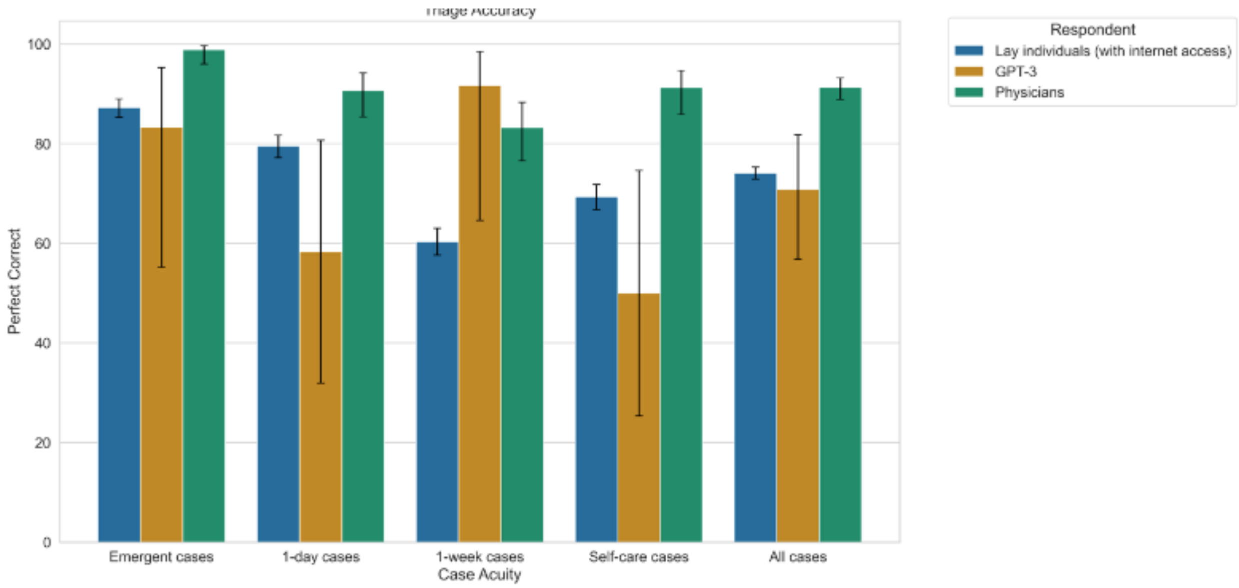
Telemedicine/nurse triage call in

Clinician-Facing | Professional Staff | Administrative

Primary Author(s): Robert Doerning, Nic Dobbins

Description: Clinics use nurse triage lines to help guide patients to appropriate care either as an outpatient or as a referral to the ED. LLMs may be able to do preprocessing on patient complaints to help guide better care referral.

Levine et al. (2023) used the GPT-3 LLM model (predating ChatGPT and more powerful, recent models (e.g., GPT-3.5, GPT-4)) using 47 vignettes of actual patient admission descriptions and compared the LLM-suggested triage category to actual triage categories determined by medical personnel. The LLM predictions performed less well than a physician (accuracy 70% vs. 91%), though given recent breakthroughs with better performing models these results may reasonably be expected to meet or exceed physician performance given GPT-4 performance on other benchmarks.



Moreover, LLM-driven voice and conversation services (e.g., BlandAI, ElevenLabs) may enable triage directly by audio conversation.

Patient Impact: Patients may be able to potentially be triaged faster, in an automated fashion.

Provider Impact: Nurses who currently manage triage lines may be freed up to work more directly with patients.

Administrative and Professional Staff Impact: N/A

Research Impact: N/A

Financial Impact: A certain proportion of nurses currently engaged in triage may be freed to work in other aspects of care.

Feasibility (Technical/Operational Impact): Recent studies suggest LLM-driven triage may perform close to or better than a human medical professional, though given the high stakes, intensive validation and testing is recommended. This use case is feasible in the intermediate term but likely not before.

References: https://learn.microsoft.com/en-us/semantic-kernel/memories/vector-db

Levine, David M., et al. "The diagnostic and triage accuracy of the GPT-3 artificial intelligence model." *medRxiv* (2023): 2023-01.

Exhibit C

Risk identification Subgroup Report

Scope

The Risk Identification Subgroup (Risk Subgroup) was tasked with creating an overview of the legal, regulatory, ethical and mission related risks associated with the use and training/fine-tuning of generative AI, including large language models (LLMs) in the clinical environment, including research in the clinical environment. As part of this scope, the subgroup was asked to consider, at a high level, what policies, processes, procedures, education, resources, infrastructure, oversight, investment, communication, equity considerations, etc. may be needed over time to successfully navigate the increased use of generative AI in the healthcare environment.

The Risk Subgroup included the following individuals: Beth DeLair (co-lead), Grace Lin (co-lead), Ana Anderson, Sally Beahan, Augie D'Agostino, Malia Fullerton, Marcia Gonzales, Lisa Hammel, Margaret Lane, Leo Morales, Shawntá Mosely-App, Adina Mueller, Aimée Olivier, Adam Parcher, Kelly Patrick and Gerianne Sands.

Approach

The Risk Subgroup was an interdisciplinary group made up of members with various backgrounds and expertise in different healthcare risks from UW and Fred Hutchinson Cancer Center. Using the collective knowledge of the group, we identified and described 14 risks, which we put into 8 categories: 1) legal, 2) privacy, 3) accuracy and integrity, 4) security, 5) bias and discrimination, 6) medical/patient care, 7) human resources, and 8) other. Next, each subgroup member was given the opportunity to rank the risks using a scoring tool and an overall risk ranking was determined (see "Risk Ranking Methodology" for more information). Finally, the subgroup discussed mitigation strategies, considering both existing resources and identified gaps. Based on the identified gaps, the subgroup developed eight recommendations to help mitigate risk with generative AI implementation in the clinical environment.

Risk Ranking

Methodology

The 14 risks were scored based on a modified UW Medicine Compliance scoring tool. The scoring tool included a likelihood score and an impact score. The **likelihood score** represents the potential that the risk could occur as generative AI is implemented. The **impact score** represents the potential for negative legal, reputational or financial impact to the organization if the risk occurs. Both scores were on a 1 to 10 scale, with the higher number equaling a great risk. The likelihood score and impact scores were multiplied together to determine an **overall score** for each risk. Subgroup members were asked to complete the scoring tool individually. Using the average overall scores, the 14 risks were ranked from highest to lowest risk. The subgroup reviewed the risk rankings and discussed whether any risks should move within the ranking. The group determined that medical malpractice should be moved up in the risk rankings right below clinical care given the correlation between the two risks.

Summary

The chart below shows the risks as ranked using the above methodology. We have included the average overall scores, as well as the scoring range. The subgroup acknowledges that each member approached the scoring with their own subject matter expertise and experiences, so some risks may have been top of mind for certain members but not others. This is reflected in the broad scoring of many of the risks.

Ranking	Risk Area	Risk	Average Overall Score	Range
1	Legal	Legal/Regulatory Landscape	82.73	81 - 100
2	Privacy	Data Breach	81.18	63 - 90
3	Accuracy & Integrity	Model outputs	74.36	63 - 90
4	Security	Data use and Storage	73.73	35 - 80
5	Other	Concerns re LLM as a new initiative	67.73	64 - 81
6	Legal	Contracts	66.55	30 - 90
7	Model Bias	Discrimination	56.82	48 - 100
8	Medical/Patient Care	Clinical Care	54.45	42 - 80
9	Medical/Patient Care	Malpractice Risk	36.18	24 - 81
10	Privacy	Sale of PHI	45.73	32 - 72
11	Human Resources	Human Resources	42.27	24 - 56
12	Other	Brand/Reputation	33.64	25 - 35
13	Medical/Patient Care	Patient Experience	32.27	20 - 64
14	Legal	Intellectual Property	28.18	25 - 40

Identified Risks

The following is a description of each risk.

Risk 1: Legal – Legal/Regulatory Landscape

- Currently there are minimal, and potentially inconsistent, laws regulating generative AI, including LLMs. However, the group also recognizes that laws and law enforcement around these areas are rapidly evolving and involve competing state, federal, and international regulatory bodies that will require monitoring (e.g., FDA -medical device, OIG – anti-kickback, CPB – consumer complaints). By way of example, in the 2023 legislative session, at least 25 states along with Puerto Rico and the

District of Columbia introduced artificial intelligence bills. Since October 2023, the White House issued an Executive Order on AI and the European Union reached a provisional deal on the world's first comprehensive laws to regulate AI.

- Potential concerns related to the practice of medicine without a license.
- Billing compliance risks related to the medical necessity of and accuracy of documentation, charge capture and coding for services provided.
- Potential legal liability for failure to comply with anti-discrimination laws.
- For state agencies, use of an AI system could create a public record under the WA State Public Records Act.

Risk 2: Privacy – Data Breach

- The potential for inappropriate use, access or disclosure of PHI or PII (including PHI or PII entered into a search engine function).
- Potential for re-identification when using de-identified data or a limited data set.
- Effectiveness of de-identification (could lead to breach if not properly de-identified).
- Inability to pull back data in the event of a breach.
- There may not be a contractual obligation for LLM to report or respond to a breach.
- Cost of data breach; UW Medicine does not have cyber liability coverage.

Risk 3: Accuracy and Integrity – Model Outputs

- The models may generate outputs that are false or incorrect (hallucinations).
- LLMs do not verify facts or resources; bad actors may jailbreak the systems to obtain dangerous information or to engineer the models to act in harmful ways.
- Bad actors may insert malicious codes or false information into LLMs or orchestrate data extractions.
- Lack of visibility into data used to train the models.
- Lack of accounting for temporality in training models (e.g., outdated treated the same as new data/evidence).

Risk 4: Security – Data Use & Storage

- Degree of information security risk related to implementing generative AI is based on two factors:
    - Existence and degree of key internal controls
        - Inventory (where and how generative AI is being used)
        - Logging and monitoring (data input and output requests)
        - Software development lifecycle (ability to review/amend coding changes)
        - Access (who has access to what part(s) of data)
        - Data protection (transmission, storage, processing)
    - How and where data is stored and maintained
        - Closed environment (we control the data)
        - Contained environment (data in the cloud or we contract with a vendor to store/maintain/administer)
        - Open environment (data open to the internet, e.g., "google" or "siri")

Risk 5: Other – Concerns regarding LLM as a New Initiative

- Rapid speed at which the generative AI space is evolving, requiring a balance between an ability to move quickly to keep pace with innovation and new practices while still ensuring necessary due diligence is conducted and appropriate operational infrastructure is in place.
- Lack of investment of money and in structure and process to enable efficient review, assessment, approval and ongoing monitoring of tools for use.
- Creation and use of generative AI and tools using generative AI without guidelines, education and training.
- Allowing abstract risks to slow down benefiting from tools versus taking a surgical approach to assessing and mitigating risks.
- Lack of transparency and clear communication as work continues to evolve.
- Effectively coordinating within and across closely aligned organizations.
- People using generative AI without realizing it; also, how to define and identify generative AI.
- Inability to switch generative AI models/processes; need for due diligence process.
- Lack of competitiveness (e.g., as a healthcare provider and employer of choice) if we are slow to adopt.
- Risk to us as an institution as third parties optimize use of generative AI tools in a way that might be disadvantageous to us (e.g., payors).

Risk 6: Legal – Contracts

- Contractual provisions (e.g., IT rider, indemnification, staffing provisions, etc.) will need to be considered, added, and/or modified.
- Contractual language should be reviewed to ensure limited scope of data usage. For example, it may be permissible to use data to enhance current purchased product/services, but it should not extend to the development of future products—may be avenue for commercialization.
- Contractual risk-shifting with respect to regulatory compliance, data breach, etc.
- Data sharing agreements and how generative AI changes how we review.
- Decentralized contracting.

Risk 7: Model Bias – Discrimination

- Biases, including confirmation and perpetuation of bias discrimination, given lack of visibility into how data models are trained.
- Discriminatory or biased outcomes could potentially be caused or exacerbated by model recommendations or predictions, for example in personnel recruitment and the delivery of care.

Risk 8: Medical/Patient Care – Clinical Care

- How generative AI impacts a provider's approach to the care of a patient.
- Use of generative AI without independent clinical judgement, validation, and engagement.
- Generative AI in the development of medical devices.
- Risk of not engaging in generative AI as it becomes more commonly used and changes the standard of care. Using AI as a clinical co-pilot (i.e., in combination with a healthcare provider) may be expected and potentially required for better patient care and outcomes.

Risk 9: Medical/Patient Care – Malpractice

- Reliance on generative AI in decision-making process may change the standard of care (as noted above in Risk 8). Risks potentially exist in both over-reliance and under-reliance on generative AI tools.
- Informed consent concerns.
- How does the use of these tools impact the practice of medicine (consider the changes to the standard of care)?

Risk 10: Privacy – Sale of PHI

- The possibility that our PHI will be sold.
- The potential that products trained on our clinical data (which includes de-identified PHI) will be monetized or otherwise commercialized in a manner that is inconsistent with institutional comfort.

Risk 11: Human Resources

- Labor and personnel risks (e.g., technology affecting staffing levels, replacing work that is currently being done by represented employees, changes to employees' scope of work).
- Employee data breach risks.

Risk 12: Other – Brand/Reputation

- Brand and reputational risks associated with the many other risks identified (e.g., data breach, patients not understanding if/how we use this functionality in their care, etc.).
- Includes UW Medicine's reputation as a leading research and teaching institution and its ability to stay on the cutting edge of research and clinical teaching practices which may impact recruitment and retention of faculty and students.

Risk 13: Medical/Patient Care – Patient Experience

- Disconnect with clinician, and patient perception that a "real doctor" is not involved in their care.
- Patient trust in use of generative AI, including LLMs, for diagnosis and treatment.
- Patient request for use of generative AI, including LLMs, for diagnosis and treatment tying back to risks identified above about the potential changes to the standard of care.
- Patients' individual use of publicly available generative AI.

Risk 14: Legal – Intellectual Property (IP)

- IP risks related to both data from vendors as well as data we are sharing.
- Plagiarism concerns.
- Data fabrication and falsification.

Mitigation Strategies and Recommendations

Based on the risks, the Risk Identification Subgroup discussed potential mitigation strategies. This included identifying existing resources and offices that provide services that support mitigation of the above risks. The subgroup also identified several gaps where additional infrastructure and resources are recommended to mitigate the risks of generative AI use in the clinical environment.

Recommendation #1: Policy

A written (and evolving) policy that broadly addresses when and how generative AI, including LLMs, may be used in the healthcare setting, and, at minimum, specifically addresses:

- The use, access and disclosure of PHI and PII.
- The institutional approach/risk tolerance regarding:
    - Protections required to share very large de-identified data sets
    - Level of allowable uncertainty related to data accuracy given unknowns of external models
- Defining allowable use of generative AI, including LLMs, in clinical care, including a definitive statement that no care should be provided without clinical judgement and decision making.
- Use of generative AI in hiring practices.
- Plagiarism concerns (e.g., when to pause and vet both inbound and outbound data before use).
- Prohibition on the sale or other "commercialization" of PHI or PII.
- Potential for bias and discrimination.
- Requiring human interaction/verification with data output versus full automation.
- Ensuring equitable access to and use of generative AI tools.
- Clear decision-making pathway and authority.

Recommendation #2: Governance Structure

In addition to clinical, operational, and IT staff, inclusion of representatives from the following in the governance structure:

- Fiscal/financial to ensure initiatives are understood and funded appropriately.
- Human Resources to ensure potential hiring and labor risks are represented.
- Marketing/Communications
- Clinical Risk Management
- Compliance
- Legal
- Equity
- Patient safety and quality
- Subject matter experts who have diverse backgrounds and understand generative AI/LLM data models, tools, and features.

Recommendation #3: Committees/Workgroups

The Risk Subcommittee recommends the following workgroups or committees be convened to address potential risks with generative AI, including LLMs:

- A workgroup that reviews proposed uses of PHI and PII outside of research (as opposed to clinical data used for research).

- A clinical workgroup that identifies potential uses and includes clinicians, risk management, clinical operations leaders, and representatives from clinical schools and programs (e.g., residency program, medical school).
- A security/technical team that has subject matter expertise in all areas related to generative AI, and who can develop internal capacity to respond to a compromised security environment.
- A legal workgroup that monitors, analyzes, and communicates legal/regulatory changes. The workgroup could include contracting review, language revisions, and education.
- A workgroup that considers matters from the perspective of the "patient experience", and includes at a minimum, representatives from the following:
  - Clinicians
  - Operations
  - Marketing and communications
  - Patient representative or champion

Recommendation #4: Education

An education program and plan for those who use generative AI, including LLMs, which addresses/includes:

- When and how to use generative AI tools (may include a checklist/review process)
- Privacy and security risks
- Clinical care issues, including standard of care expectations and potential malpractice risk
- Bias and discrimination risks
- Ethical principles

In addition, the workgroup recommends a separate training module/checklist for contracting groups that addresses special provisions that must be reviewed and included when contracting with vendors that offer generative AI tools.

Recommendation #5: Communication

Easily accessible and transparent communication to end users, patients and other stakeholders as identified, which includes:

- A distinct, easily accessible, and navigable website that links to resources (e.g., policies, guidance documents etc.), any required/available training, best practices, etc.
- Messaging from Marketing and Communications teams related to use of generative AI, including LLMs, in our healthcare system.
  - Internal e-mail communications to staff
  - Epic MyChart and other communication options to patients
- Pro-active vs. reactive communications.
- A list of terms/guardrails on how we communicate externally regarding generative AI.
- Tools to help the patient community understand when a tool/functionality relies on generative AI as it is not always clear.
- A process to communicate emerging trends, issues, laws and best practices with relevant stakeholders.

Recommendation #6: Resources

Resourcing that allows for:

- Funded infrastructure to support generative AI initiatives, including but not limited to:
  - Intake, assessment and decision making when considering engaging in a specific tool or functionality (including any pilot projects or developing of any prototypes to test concepts prior to releasing any generative AI tools/functions).
  - Any necessary monitoring or auditing of processes (e.g., data accuracy, looking at clinical outcomes from use of tools, behavior analysis of tools on end users).
  - Issue and risk management.
- Opportunities for relevant stakeholders, including legal teams, to attend conferences or otherwise receive education on generative AI, including LLMs, emerging trends, issues, laws and best practices.

Recommendation #7: Contracting Process

Contracting is decentralized throughout UW Medicine and the emergence of these technologies will require a more thorough review of specific provisions that will need to be added to existing templates and considered during the negotiation process to decrease institutional risk. This may include, but is not limited to:

- Definitions related to data used/shared (e.g., de-identified, etc.).
- Require vendors to describe what they have done to minimize bias and discrimination.
- Consider the need for representations and warranties or indemnification by vendor for use of generative AI (e.g., third party indemnification re: IP infringement).
- A provision prohibiting the sale or other commercialization of PHI (including de-identified PHI) and PII.
- Labor impact and clauses.
- Monitoring/identifying additional services to existing contracts and their impact.
- Breach notification.
- Ability for us to audit.
- Contracting due diligence when the model our vendor is relying on is not theirs (e.g., Epic/Microsoft).

Recommendation #8: Miscellaneous

- Ensure partnership, collaboration and consistency with upper campus units, existing and emerging policies, and approach as well as organizations like Fred Hutchinson Cancer Center and Seattle Children's Hospital (e.g., exploring insurance needs). Review policies as they evolve at affiliate and collaborating organizations where UW Medicine faculty members may be at multiple worksites/campuses and subject to different and possibly conflicting/varying policies.
- Consider whether any changes will need to be made to our Notice of Privacy Practices.
- Integrate risk recommendations relating to generative AI into UW Medicine Strategic Plan.

Exhibit D

Governance Subgroup Report

Scope

The LLM Governance Subgroup was responsible for developing a proposed committee or governance structure for developing policies, addressing issues and overseeing our institutional approach to the use and training of generative AI, including large language models (LLMs) in the clinical environment, including research in the clinical environment. As part of this scope, the subgroup was asked to consider, at a high level, what policies, processes, procedures, education, resources, infrastructure, oversight, investment, communication, equity considerations, etc. may be needed over time to successfully navigate the increased use of generative AI in the healthcare environment.

The Governance Subgroup included the following individuals: Ana Anderson (lead), Trevor Cohen, Jeff Leek, Kristal Mauritz-Miller, Sean Mooney, Adam Parcher, Kelly Patrick, Anneliese Schleyer and Peter Tarczy-Hornoch.

Broad Concept of Governance

UW Medicine's ability to efficiently and effectively leverage generative AI, including LLMs, in the healthcare setting will require a strong approach to governance. The governance sub-group began its work by identifying several principles of effective governance in this space including, but not limited to:

- Making the right thing to do the easy thing to do
- Taking a proactive (versus reactive) approach
- Implementing formal/transparent operational processes
- Developing policies that are not overly restrictive but include clear consequences
- Ensuring clarity of scope (e.g., generative AI vs. other forms of AI, clinical vs. research)
- Identifying accountable leadership, trusted and empowered by the organization
- Designing a scalable approach
- Ensuring governance is efficient, integrated and aligned with other governance structures
- Framing our institutional approach positively (here's what you can do and how to do it versus here's what you cannot do)
- Ensuring broad stakeholder involvement
- Appropriately resourcing the work to support our approved institutional approach

These principles led the group to a broad concept of governance, inclusive of the following key areas: 1) strategy, 2) governance structure, 3) leadership and accountability, 4) policy, 5) operational workflows, 6) risk assessment, and 7) communication and education. It is the view of the sub-group that a successful governance approach should address each of these areas. *See* Figure 1.

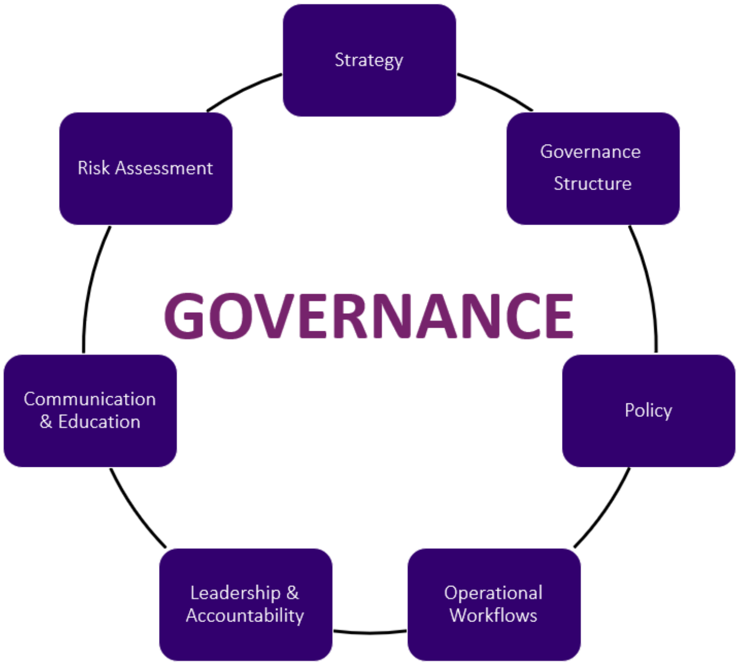


Figure 1: Governance

Recommended Approach – Phased Governance

In order to design and advance the governance approach described above, the sub-group recommends a phased approach:

- Phase I (Now): Develop the Foundation
    - ★ LLM Workgroup (~ 9 months; Aug. 23 – Jan. 24)
- Phase II (Near): Build the Infrastructure
    - ★ Generative AI Taskforce (~ 6 months)
- Phase III (Far): Launch and Operate under Steady State (permanent)

As the work of the initial LLM Workgroup concludes, the governance sub-group recommends establishing a Generative AI Taskforce ("Taskforce") to continue the Phase I work and develop a defined and comprehensive business approach to generative AI in the healthcare setting (Phase II). The Taskforce (and its supporting structure) should:

1. **Develop a strategy for generative AI in the healthcare setting.** This work should include, without limitation, an assessment of how generative AI can be leveraged to advance the UW Medicine clinical strategic plan, a prioritization framework to guide what tools/functionality we pilot,

exploration of partnership opportunities (e.g., vendor partnership, consortia/collaboratives), learnings from peer institutions, consideration of building internal capability to develop generative AI tools, and potential philanthropic or other funding to support implementation.

2. **Draft institutional policy to govern the use of generative AI in the healthcare setting.** This policy (or policies) should address use of these tools/functionality for patient care, business operations supporting the healthcare enterprise, and use of publicly available tools in the course of everyday work. The policies also should address any unique requirements associated with clinical research involving LLMs or the use of clinical data to support research involving generative AI.
3. **Design a long-term governance structure (including draft charters and proposed membership).** This long-term structure should be designed to oversee UW Medicine's institutional approach to generative AI in the healthcare setting. In developing a structure, the Taskforce should:
    a. Consider existing governance structures and determine how this new structure might fit within, replace and/or complement what exists today[1]
    b. Take care to avoid duplication, confusion and overloading those with key subject matter expertise
4. **Develop (or define) operational workflows.** The Taskforce must ensure that there are operational processes (whether net new or existing) to support, at a minimum, intake, assessment and approval for the following three categories:
    a. Tools/functionality to support clinical activity or activity in support of the clinical enterprise;
    b. Clinical research involving generative AI; and
    c. Sharing of clinical data for research involving generative AI

The Taskforce also should consider requirements for early pilots to ensure operational rigor post-approval.[2]

[1] • UW Medicine Information Technology Governance
- o Tier I Strategic Technology Committee
- o Tier II Clinical Research Informatics/Tier III RAPiD
- o Tier II Data and Analytics/Tier III Predictive Analytics
- o Tier II Clinical Operations

• UW/Fred Hutch Joint Clinical Data Oversight Committee
• Fred Huth data governance structure
• UW/Fred Hutch Joint Clinical Trials Governance
• Institutional Review Board (IRB)
• Mission Forward (automation)
• Institute for Medical Data Science
• Quality/clinical risk management structures
• Operational venues, such as Clinical Operations Roundtable (COR)
• Departmental governance

[2] This could include, for example, identifying the accountable operational leader, specifics for the pilot (e.g., timeline, scope of implementation, intended use), key performance indicators and plan to track defined metrics, monitoring plan, etc.

5. **Define processes to minimize risk.** These processes should include, at a minimum:
    a. A framework or rubric that can be used operationally to assess risk on a case-by-case basis and enable tailored risk assessment;
    b. Plans to monitor, audit and/or decommission tools/functionality;
    c. Identified pathway(s) to address unintended consequences, as appropriate.
6. **Creation of a robust education, communication and engagement strategy.** This strategy should target a variety of audiences (e.g., faculty, staff, trainees, patients, policymakers and labor unions), propose what types of materials will be needed to support the strategy and outline the risks of failure to implement a comprehensive approach to education, communication and engagement around the use of generative AI in the healthcare setting.

Based on the concrete deliverables outlined above, the Taskforce should be charged with developing financial, resource and other recommendations, as appropriate, to support the proposed business approach in the next one (1) to three (3) years.

Final Thoughts

As the sub-group developed the recommended governance approach outlined in this report, several themes emerged. First, generative AI is a rapidly evolving space. We must be nimble and continue to move forward with our work quickly, but also ensure the necessary due diligence and operational infrastructure to support the work. Second, resourcing will be critical. The level of resourcing dedicated to this work needs to be sufficient to execute on our institutional approach. Insufficient investment will create risk, operational bottlenecks and ultimately, could prevent us from taking advantage of these tools in a timely (and responsible) manner. Finally, our work must balance innovation and opportunity with our obligation to use this technology in a safe, ethical and responsible way. Only by striking the right balance will we be able to use these incredible new tools in the healthcare setting to advance our mission to improve the health of the public.