A Framework for Situating Innovations, Opportunities, and Challenges in Advancing Vertical Systems with Large AI Models

Gaurav Verma, Jiawei Zhou, Mohit Chandra, Srijan Kumar, Munmun De Choudhury
Georgia Institute of Technology
{gverma, j.zhou, mchandra9, srijan, munmund}@gatech.edu

Abstract

Large artificial intelligence (AI) models have garnered significant attention for their remarkable, often “superhuman”, performance on standardized benchmarks. However, when these models are deployed in high-stakes verticals such as healthcare, education, and law, they often reveal notable limitations. For instance, they exhibit brittleness to minor variations in input data, present contextually uninformed decisions in critical settings, and undermine user trust by confidently producing or reproducing inaccuracies. These challenges in applying large models necessitate cross-disciplinary innovations to align the models’ capabilities with the needs of real-world applications. We introduce a framework that addresses this gap through a layer-wise abstraction of innovations aimed at meeting users’ requirements with large models. Through multiple case studies, we illustrate how researchers and practitioners across various fields can operationalize this framework. Beyond modularizing the pipeline of transforming large models into useful “vertical systems”, we also highlight the dynamism that exists within different layers of the framework. Finally, we discuss how our framework can guide researchers and practitioners to *(i)* optimally situate their innovations (e.g., *when vertical-specific insights can empower broadly impactful vertical-agnostic innovations*), *(ii)* uncover overlooked opportunities (e.g., *spotting recurring problems across verticals to develop practically useful foundation models instead of chasing benchmarks*), and *(iii)* facilitate cross-disciplinary communication of critical challenges (e.g., *enabling a shared vocabulary for AI developers, domain experts, and human-computer interaction scholars*).

Introduction

Large artificial intelligence (AI) models have long served as powerful tools for advancing domain-specific research and development. Early examples include the adaptation of language embeddings (like word2vec (Mikolov et al. 2013)) for generating disease taxonomies (Ghosh et al. 2016), as well as the use of YOLO-based image models (Redmon 2016) for conducting animal population censuses (Parham et al. 2017). More recently, large models such as GPT-4o (Hurst et al. 2024) and SAM (Kirillov et al. 2023) have demonstrated more advanced capabilities, catalyzing further innovations in domains as varied as AI tutoring (Lin, Huang, and Lu 2023) and digital pathology (Deng et al. 2023), all the way to software development (Barenkamp, Rebstadt, and Thomas 2020). The success of these models has been propelled by the increasing scale of their architectures—rising from a few million parameters in 2013, to hundreds of millions in 2018, and now surpassing a trillion parameters in the most recent deployments (Elmeleegy et al. 2024)—as well as by the sheer volume of data used for their training (Bahri et al. 2024). Many widely used benchmarks, both general-purpose and domain-specific, have reached saturation or are rapidly approaching it (Chollet 2024; Phan et al. 2025). At the same time, frameworks built around these large models (such as Application Programming Interfaces; APIs) have lowered barriers to entry (Schillaci 2024), spurring a surge in applications and attracting researchers and practitioners from a broad range of disciplines.

Despite these successes, a decade of adapting large models in diverse verticals has highlighted persistent challenges. For example, research teams have discovered that BERT and CLIP-based models can be brittle to small input variations (Verma et al. 2022b; Ramshetty, Verma, and Kumar 2023), large language models (LLMs) often exhibit sensitivity to prompt formatting (Sclar et al. 2024), and performance of large models can diminish in highly specialized settings (Deng et al. 2023; Chandra et al. 2024). Moreover, AI systems sometimes struggle to effectively support diverse user groups, such as users with a lower Need For Cognition (Buçinca, Malaya, and Gajos 2021). These challenges, taken together, pose bottlenecks in the vertical-adoption of the large models. The process of adapting large models to verticals demands *coordinated efforts from stakeholders across varied disciplines* and continues to be an active area of research. Existing frameworks provide good starting points to address some of these problems at a more granular level—either focusing on one component of developing AI systems or a specific vertical. For instance, Ehsan et al. (2023) present a framework to bridge the gap between social and technical aspects of developing explainable AI systems. On the other hand, focusing on a specific vertical, Trotsyuk et al. (2024) present a framework to address potential misuse of AI in biomedical research. There is still an unmet need for a more *comprehensive* framework comprising various components required for vertical adoption of large models and has broad *applicability* towards generalizing to many verticals.

For instance, if a team were to develop an AI system to help tutor high-school children or if a different team intended to build AI that aids in the provision of psychiatric healthcare, what are the components that exist in adopting large AI models for these verticals? Which of these components would pose challenges and would need innovations from the team? On the other hand, which of these components could be addressed using existing solutions? We posit that the vertical adoption of large models can be made *manageable* and *modular* with a structured framework that systematically addresses the cross-disciplinary complexities. To this end, we propose a framework designed to guide both researchers and developers in optimizing large-model development and deployment by *situating* their innovations, opportunities, and challenges within a clearly defined structure.

Our proposed framework consists of 4 layers (see Figure 1), *(i)* starting with **large AI models** at the bottom, *(ii)* **vertical-agnostic properties**, *(iii)* **vertical adaptation**, and (iv) finally, **vertical-user intermediaries**. The 4 layers represent step-wise modular abstractions involved in developing systems with large models that deliver practical value to their intended users. Drawing on case studies from multiple verticals, we discuss how to *situate* innovations, challenges, and opportunities within each layer of this framework. We consider *innovations* as advances in algorithms, metrics, or interface designs that resolve identified pain-points; *challenges* as specific obstacles that hinder vertical adoption of large AI models; and *opportunities* as recurring unmet needs that signal room for high-leverage solutions. We use the term *situating* to indicate anchoring the contributions in one of the four layers of the framework. The overall aim of the situating contributions within the framework is to ensure that the holistic adoption of large AI models across various verticals remains effective.

Beyond describing the framework and its layers, we also highlight the inherent *dynamism* among these layers (i.e., how they interact with and influence one another over time). Finally, drawing on observed trends, we present actionable recommendations that benefit researchers and developers across various verticals and disciplines. For instance, we discuss whether aspects like robustness and privacy are better addressed as vertical-agnostic properties or as a vertical-specific concerns, how scoping feedback from many verticals and intermediary problems can lead to more effective development of newer large models, when domain-specific experts can borrow modeling and interfacing techniques from others while innovating on the domain-specific data curation and evaluation methods, and how interfacing AI systems with users remains ripe with opportunities. Collectively, we believe that adopting our framework will *(a)* guide optimally placed innovations, *(b)* highlight potential opportunities, and *(c)* enable cross-disciplinary dialogue.

Who can benefit from the framework? The framework is useful for interdisciplinary teams who want to adopt large AI models in their respective verticals, and for researchers who hope to position their work within a broad ecosystem for identification of cross-layer connections and translation of knowledge across contexts. For teams focused on a single vertical, the framework decomposes the adoption pipeline into modular components. Across multiple verticals, it provides a holistic view of the broader ecosystem that *(a)* fosters cross-vertical exchange of innovations, opportunities, and challenges (i.e., what can teams in healthcare learn from teams in education) as well as *(b)* funnels feedback from many verticals to improve the next iteration of large artificial intelligence models.

A framework for advancing vertical systems with AI models

We developed our framework by building consensus among a group of experts with experience in creating and applying AI models across verticals such as well-being, web safety, and enterprise software. They brought expertise in artificial intelligence, human-computer interaction, social science, and healthcare, along with practical experience collaborating with leading industrial deployment teams, clinicians, non-profits, and non-governmental organizations. The group members engaged in reflective discussions, drawing on practical insights from their prior experiences deploying user-facing applications with large AI models. The identified recurring pain points were distilled into modular themes/layers that start with the underlying large models and end with users' needs. The group then discussed case studies and iteratively adapted the framework. The discussions also led to formulation of actionable recommendations for future work that aims to adopt large AI models in different verticals.

The framework is depicted in Figure 1 and described below, progressing from bottom to top, where each layer addresses specific functional aspects to achieve vertical-specific utility with large models.

1. Large AI models: These are large-scale (in terms of training dataset size and number of parameters) AI models, more recently dubbed foundation models (Bommasani et al. 2021) that power diverse applications across verticals. These include modality-specific models (e.g., language-only, vision-only) and multimodal models capable of handling multiple input modalities. Their functional utility lies in the general off-the-shelf capabilities they provide (e.g., natural language understanding, image-text alignment) and adaptability to new verticals through techniques such as fine-tuning, prompting, or in-context learning.

2. Vertical-agnostic properties: Large models might need general scaffolding around properties like robustness, interpretability, efficiency, and privacy, before they are useful as vertical-specific systems—such problems are considered vertical-agnostic properties. While improvements along these aspects generally benefit many verticals, nonetheless, some of these aspects may *also* require vertical-specific considerations.

3. Vertical adaptation: Designed for delivering specific value within verticals such as healthcare, web safety, and education, vertical-specific adaptations involve integrating large models' capabilities with vertical-specific data, modeling, evaluation, and interfacing the capabilities with end-users of the vertical.

4. Vertical-user intermediaries: These address vertical-agnostic challenges in interfacing systems built with large models with the end-users, focusing on aspects like trust calibration, feedback loops, and dynamic interfaces. While some interfacing challenges could be vertical-specific, others are broadly applicable challenges.

The following section presents case-studies that apply this framework to two verticals — healthcare and education.

Applying the framework: case studies

As we describe the innovations, opportunities, and challenges that exist in the vertical adoption of large models and situate them in our framework, two questions guide our efforts: "Who are we trying to help?" and "What do they care about?" These considerations gain prominence at higher levels—where user impact is tangible—but inform decisions throughout every layer of the framework. Table 1 depicts some of the example questions within different layers of the framework. These example problems show that each vertical has different data, modeling, and evaluation needs, yet they share cross-cutting challenges that exist between large AI models and vertical systems as well as between vertical systems and users.

Large AI models: The architectural scale and the pre-training dataset size have resulted in remarkable off-the-shelf capabilities of the large AI models, measured by their success on continually evolving benchmarks(Chollet 2024; Liang et al. 022). While there has been a push towards even larger models trained on huge datasets, equipping them with the ability to support inputs in multiple modalities — language, vision, audio, and in some cases even sensor data (Moon et al. 2024), has the potential to unlock new applications across many verticals. For instance, multimodal models could, in principle, process radiology scans along with diagnostic questions (Bhayana 2024), raw electrocardiogram (ECG) signals for health data analytics (Quer and Topol 2024), provide voice-based tutoring (Katsarou et al. 2023), and write and review lengthy codebases (Bairi et al. 2024). While the development of these AI models offers an affordance for multimodal input, questions remain around how *well the multimodal LLMs can reason over the non-textual forms of data*(Tong et al. 2024; Verma et al. 2024a), which requires further work within this layer of the framework. Advances that ensure multimodal LLMs indeed model all modalities reliably will ensure greater off-the-shelf capabilities in the future iterations of these models.

Vertical-agnostic properties: As we consider applying large AI models, the first set of problems are vertical-agnostic properties that apply to most verticals and pave the path for effective consideration of vertical-specific aspects. For instance, are multimodal LLMs robust to the plethora of plausible and realistic variations in user-provided inputs, given that it is unreasonable to assume users will constrain their inputs to the margins of the training distribution (Ramshetty, Verma, and Kumar 2023; Verma et al. 2022b; Nookala et al. 2023)? Will these models handle the personally identifiable information (PII) already encoded from its pre-training corpus (Carlini et al. 2021) and the sensitive data provided by users while using it for iterative refinement (e.g., from patients (Pan et al. 2024) or students (Yang and Beil 2024))? Relatedly, can large AI models provide interpretable predictions that foster transparency (Stiglic et al. 2020)? Addressing these problems as foundational steps will overcome bottlenecks across many verticals and enable a more effective use of the remedial techniques. Additionally, addressing these problems in the proximity of the large AI models layer could lead to integration of the remedial techniques in the development of upcoming large AI models; we elaborate on this in Section . We now move on to situating vertical-specific innovations, opportunities, and challenges within the framework, starting with healthcare and then exploring education. We chose to focus on healthcare and education verticals because they represent high-stakes domains with distinct yet complementary challenges: both verticals focus on building user-centered systems with healthcare demanding precision and rigorous safety measures while education prioritizing engagement and personalization to cater to diverse pedagogical needs of the learner.

Vertical adaptation: (a) MLLMs for healthcare: One of the well-explored clinical applications of multimodal deep learning models, including the recent multimodal LLMs, is as an assistant for radiologists. The models can help clinicians make diagnoses via conversational-assistance as well as writing medical reports (Johri et al. 2025; Zhang et al. 2024). Even though off-the-shelf LLMs encode medical knowledge (Singhal et al. 2023), they need to be vertically adapted to acquire diagnostically useful information from natural conversations with primary care providers. This vertical adaptation involves *curating the right data of diagnostic conversations*, which would require close involvement of domain-experts (Tu et al. 2024). Beyond curating the right data, off-the-shelf LLMs do not demonstrate properties that are required to accurately model the data–for instance, history-taking (Tanno et al. 2024; Tu et al. 2024). In the context of radiology, this would involve *equipping the underlying multimodal LLMs with temporal modeling capabilities* (Bannur et al. 2023). To evaluate whether the modeling approach is effective on the curated data, vertical-agnostic evaluations may not suffice. For instance, Yu et al. (2023) (Yu et al. 2023) demonstrate that generic 'natural language generation' metrics are not effective in capturing clinically pertinent differences between AI-generated radiology reports and those written by experts, and *propose meaningful metrics to guide future research* in this vertical. The evaluations uncovered that while both experts and AI systems can make mistakes while generating radiology reports alone, the instances of inaccuracies decrease when experts and AI work in collaboration to fix the errors (Tanno et al. 2024). Nonetheless, a crucial *challenge remains unaddressed when interfacing the AI systems with clinicians* – the collaboration loses effectiveness when the expert either overly relies on the AI predictions (Seah et al. 2021; Rajpurkar et al. 2020) or is excessively critical of them (Agarwal et al. 2023).

Vertical adaptation: (b) MLLMs for Education: Large AI models, including multimodal LLMs, are already be-

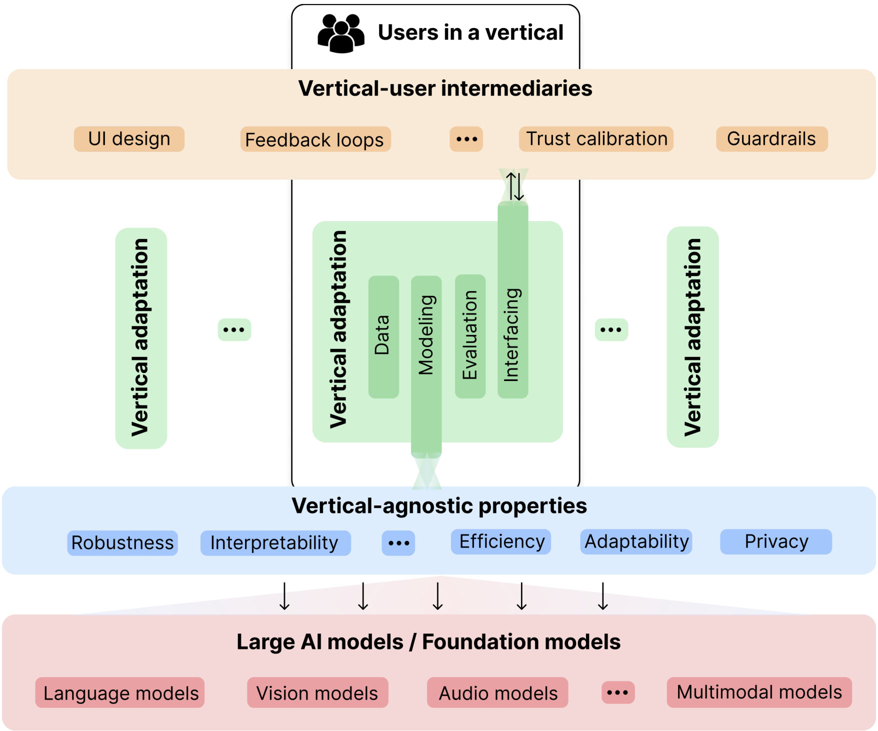

Figure 1: Overview of the proposed framework for situating innovations, opportunities, and challenges in advancing vertical systems with large AI models; read from bottom to top. Large models form the base for vertical systems. These models need scaffolding to demonstrate properties such as robustness, interpretability, efficiency, and privacy before being useful as vertical-specific systems. Vertical-specific adaptations are required to deliver value within specific verticals. This involves curating data, designing or adapting modeling approaches, vertical-centric evaluations, and interfacing the model's outputs with users. General problems in designing interfaces and interactions between the system and users are designated as vertical-user intermediaries. The dynamism between the framework layers is noteworthy (depicted by ↓ and ↑). Over time, vertical-agnostic properties, especially those applicable to many vertical systems, could become ingrained properties in future models as development strategies evolve. Similarly, modeling for vertical adaptation could become less prominent as large models become efficiently adaptable, exemplified by the success of in-context learning with large language models. Finally, vertical-specific insights for interfacing systems with users and general interfacing techniques influence each other over time.

ing used by students across the globe to assist with their learning (Zhu, Zhang, and Wang 2024), demonstrating their promise in tutoring. Even though their ad hoc capabilities are noteworthy, Macina et al. (2023) (Macina et al. 2023) note that their systematic impact on "tutoring has largely remained unaffected". A central challenge in this vertical concerns *curating data that captures diverse pedagogical strategies*, covering a broad range of topics, learner demographics, and instructional modes encountered in real classrooms (Jurenka et al. 2024). From a modeling perspective, besides fine-tuning (which may prove to be an inefficient strategy for adapting to different definitions of what constitutes effective pedagogy), it is *advantageous to allow learners to specify desired attributes across pedagogical dimensions and have the model reflect them* (Team et al. 2024). The *evaluation of such models should be grounded in learning science*, which often prioritizes motivating and promoting engagement from the learner, and not just "giving the right answer"(Foster et al. 2023). It is also crucial that the *interfaces support learner-tutor interactions such that the conversations are grounded in what the student "sees"*, which is one of the core principles of student-centric pedagogy.

Vertical-user intermediaries: Even though there are vertical-specific considerations involved in interfacing systems with users, some of these challenges are frequently encountered across many verticals. For instance, as we noted, clinician-AI collaboration tends to become less useful when the capabilities and limitations are not adequately understood by the experts. Similar observations around 'algorithm aversion or appreciation' exist across other verticals (Dietvorst, Simmons, and Massey 2015; Logg, Minson, and Moore 2019; Qian and Wexler 2024). Given the broad nature of the underlying challenge, it is effective to work on *trust calibration* as a vertical-agnostic vertical-user intermediary. E.g., introducing uncertainty expressions ("I'm not sure, but...") in AI-generated responses leads to decreased overreliance and calibrated trust among users (Kim et al. 2024) – an insight that can benefit many verticals, including generative information retrieval, healthcare assistance, and educational tutoring. Similarly, designing interfaces that cap-

Vertical-User Intermediaries

Trust calibration: How can AI systems effectively communicate their capabilities and limitations to users to prevent issues of overreliance or unwarranted skepticism?

Feedback loops:How can real-time user feedback be systematically captured and integrated into iterative refinement processes for large AI models?

Dynamic interfaces: How can interface designs dynamically accommodate users with varying cognitive engagement preferences, optimizing usability for both high and low Need For Cognition (NFC) individuals?

Vertical Adaptation in Healthcare	**Vertical Adaptation in Education**
Data: What methodologies can be used to curate specialized datasets of medical dialogues necessary for accurately tuning multimodal models for clinical use?	Data: How should educational data be curated to reflect diverse pedagogical strategies and learner demographics to improve AI-assisted tutoring?
Modeling: How can AI models effectively incorporate temporal patient history data to improve accuracy in clinical diagnostics?	Modeling: What modeling techniques can efficiently adapt AI systems to individual learner needs and preferred pedagogical strategies?
Evaluation: Which clinical standards and metrics most effectively measure the quality and reliability of AI-generated outputs in healthcare settings?	Evaluation: How can AI system evaluations accurately assess their impact on learner motivation, engagement, and educational outcomes beyond correctness of content?
Interfacing: How can AI systems best support collaboration with clinicians?	Interfacing: How can AI tutoring systems effectively ground responses within the immediate learning context and visual perspective of the student?

Vertical-Agnostic Properties

Robustness: How robust are AI models to realistic & plausible variations in user inputs spanning multiple modalities?

Privacy: How effectively can AI models ensure the privacy & security of personal and sensitive data?

Interpretability: In what ways can AI models be made to provide interpretable predictions to improve transparency?

Multimodal Large Language Models

How can modalities beyond language (visual, audio, sensor data) be reliably processed with large AI models?

Table 1: Operationalizing the framework for adoption of large AI models in the healthcare and education verticals; read from bottom to top. Each question is an example of existing/potential innovations, challenges, and opportunities in vertical adoption of large artificial intelligence models (in this case, multimodal large language models).

ture real-time, in-situ, and implicit user feedback, will unlock iterative refinement of underlying systems across many verticals(Shi et al. 2024). Furthermore, interfaces that dynamically adapt to deliver the optimal experience to a group of users with diverse Need For Cognition levels (NFC; a personality trait that considers users' tendency to engage with cognitive activities(Cacioppo and Petty 1982)) is an opportunity that will benefit many verticals. More specifically, for decades, researchers have observed that users with higher NFC levels benefit significantly from complex interfaces, while others struggle to adopt them(Carenini 2001) — a pattern that spans many verticals and has continued with recent generative artificial intelligence technologies (Toner-Rodgers 2024; Buçinca, Malaya, and Gajos 2021).

Dynamism between layers of the framework

The framework is structured in layers to support *modularity* in addressing innovations, opportunities, and challenges. Yet these layers do not serve the purpose of *rigidly demarcating* the underlying problems. Instead, as we discuss in this section, they exhibit extensive cross-layer interactions and mutual influences—referred to here as "dynamism". This section provides concrete examples that (a) illustrate this dynamism and (b) show how the dynamism can foster distributed-yet-collaborative synergy in AI's development.

Let us start with **vertical-specific considerations influencing vertical-agnostic properties**. As practitioners and researchers explore the applications of large models in different verticals, they highlight the successes as well as shortcomings. Certain aspects such as lack of robustness, poor handling of private data, and lack of efficient adaptability are frequently *encountered/identified* across verticals and make their way to being more effectively *addressed* in a vertical-agnostic manner. When addressing these challenges in a vertical-agnostic manner, **vertical-agnostic solutions can influence the development of the next iterations of large AI models**. Matryoshka representations (Kusupati et al. 2022) present a strong case study for this. In many real-world search systems (e.g., patient records, legal reviews, and educational video search), relevant items are often retrieved by computing similarity between neural representations. At large scales—hundreds of millions of items—this must be efficient. Kusupati et al. (2022) (Kusupati et al. 2022) introduced "Matryoshka doll" representations, where $14\times$ smaller embeddings match the performance of full-size ones for classification and retrieval. This innovation supports flexible representation sizes, enabling efficient retrieval and classification in a vertical-agnostic manner, even though the initial inadequacies were identified during many vertical-specific adaptations. Although originally the effectiveness of Matryoshka representations was demonstrated on ImageNet-scale datasets, it was later adopted by OpenAI and others to train large AI models (e.g., `text-embedding-3`) on web-scale data (OpenAI 2024).

Akin to the dynamism between vertical-specific considerations, vertical-agnostic properties, and large AI models, certain challenges may also **shift from interfacing-related aspects within verticals to broader vertical-user intermediaries**. Consider AI hallucinations in high-stakes fields like healthcare, where unreliable responses can lead to adverse outcomes and erode trust, prompting research on cognitive forcing interventions in medical AI-assisted decision-making (Buçinca, Malaya, and Gajos 2021). With LLMs now prevalent in many domains (e.g., web retrieval, law, education, and mental healthcare), hallucination has become a key challenge. This has spurred work on vertical-user intermediaries, including quantifiable or language-based uncertainty cues (Xiong et al. 2024; Kim et al. 2024) and referencing source documents for verifiable claims (Gao et al. 2023), with early evidence of reduced user over-reliance on AI outputs (Bo, Wan, and Anderson 2024).

The **dynamism between layers** is inherently desirable as it **facilitates the distributed-yet-collaborative development of AI systems**. The large AI models are often developed in highly resource-rich environments, while the practitioners and researchers who develop vertical-specific insights that trickle down to the bottom layers of the framework are often situated outside of these selective environments. Here, two points are worth noting: first, to deliver benefits within a vertical using large AI models, it is critical to actively engage with domain experts who have curated specialized data, engineered novel methods, or designed meaningful evaluations for those verticals. This enables specialized insights that arose from vertical-specific explorations, while simultaneously embedding them in more general-purpose model architectures. Second, it is *incorrect* to assume that vertical-specific explorations merely *piggyback* on advances in large AI models. Rather, vertical-specific advances play a critical role in the development of large AI models. When used carefully, these vertical-specific insights can make general-purpose models more capable and speed up the release of solutions for specific user needs.

Intended framework outcomes

Developing on the trends that illustrate the dynamism between layers of the framework and its collaborative-yet-distributed nature, we discuss how the framework can aid along the following axes: encouraging researchers to optimally situate or borrow innovations, discovering overlooked opportunities, and engaging in a structured cross-disciplinary dialogue. The actionable recommendations were derived through an iterative process, where the authors reflected on recurring pain points encountered across different applications of large AI models and deliberated on steps that could enable their effective integration.

Optimally situating and borrowing innovations

One of the key intended outcomes of the framework is to help situate the innovations such that they have increased potential for impact and also promote optimal use of resources. An important aspect of this is to **consider when vertical-specific innovations, particularly on the modeling and interfacing front, could have broader impact across many verticals**. For instance, studies across many verticals have found that adapting LLMs on benign datasets for healthcare, education, and law compromises their safety — making them more likely to respond to potentially harmful prompts (Qi et al. 2024). Individual research teams across verticals therefore go through additional steps to first *quantify the extent of compromise in safety upon fine-tuning* and then, if the extent is unacceptable, *improve the compromised safety of vertically-adapted LLMs* (Niknazar et al. 2024; De Freitas et al. 2024). However, as opposed to investigating and addressing the underlying issues independently in many verticals, Peng et al. (2024) study the loss of safety in adapted LLMs—across several LLMs and datasets—to propose a metric for safety in LLM adaptation by visualizing its safety landscape (Peng et al. 2024). The metric could be used across verticals to determine whether the LLM adaptation has compromised safety and the extent to which remedial strategies are required. This illustrates how addressing challenges that are encountered across several verticals as a vertical-agnostic properties could be more optimal in terms of impact and resource allocation.

Similar examples exist in how vertical-specific interfacing challenges could point to broadly applicable problems that are optimally addressed as vertical-user intermediaries. As a specific example, the tendency of LLM-based chatbots to neglect crucial aspects of user interactions as they primarily focus on outcomes has been noted in many verticals, including their applications in therapeutic conversations (Zhou et al. 2024), information seeking (Sharma, Liao, and Xiao 2024), writing code (Bajpai et al. 2024), or humanitarian frontline negotiations (Ma et al. 2024). While vertical-specific strategies have been proposed (like inducing Gricean maxims to structure code-related interactions (Bajpai et al. 2024)), it is promising to assess whether such strategies generalize across verticals as a vertical-user intermediary. If such a challenge can be addressed in a vertical-agnostic manner, it could enable more effective vertical-user interactions across many verticals. It is worth emphasizing that situating innovations as vertical-agnostic properties (whether in modeling or interfacing) does not aim to discourage vertical-specific innovation as it is the latter that provides insights into what could work (snowballing effect). The objective here is to encourage researchers and practitioners to consider the potential impact of their innovation and aid others in adopting it to establish vertical-agnostic generalization.

Conversely, optimality also requires that researchers and practitioners **adopt vertical-agnostic strategies when they are effective in their specific verticals**. For instance, while many vertical-specific works note that prompt engineering is often ad hoc, solutions like DSPy (Khattab et al. 2023) provide a structured, vertical-agnostic approach to optimizing prompts, thereby addressing common challenges across multiple domains and saving resources.

Uncovering overlooked opportunities

Guided by the philosophy of "turning frequent failures into signals", the framework can also help uncover overlooked opportunities. A compelling example exists in the development and adoption of large language models for languages in South Asia. Individual teams have noted the limitations of LLMs trained on predominantly English data when they are used in non-English languages for applications in different verticals (Jin et al. 2024; Kumar et al. 2024; Verma et al. 2022a). Such LLMs also lack the socio-cultural awareness to adapt to the needs of users in South Asian regions (Pawar et al. 2024). These failures have highlighted an opportunity for collaborative and participatory research to develop new LLMs for South Asian languages, such as BharatGPT (BharatGPT 2025) and SeaLLMs (Nguyen et al. 2023). Since their deployment, these models have been adopted across many verticals by businesses, governments, and non-profits (CoRover AI). By identifying the pattern among the failures in vertical adoption and addressing the major issues at the large AI model-layer of the framework, the vertical-agnostic and vertical-specific layers can build on top of advanced off-the-shelf capabilities of the newer models. Such **coordinated and targeted efforts also ensure that the burden of developing large AI models that are effective in specific verticals does not fall on individual teams** as the process is restrictively resource intensive and carries environmental costs (Strubell, Ganesh, and McCallum 2020).

Another bottleneck that exists in the vertical adoption of large models is careful handling of sensitive data – including health data (Pan et al. 2024), student data (Yang and Beil 2024), and proprietary workflows (Tang et al. 2024). Current large models readily allow adaptation via in-context learning or custom system instructions, which pose the risk of exposing sensitive data via jailbreak attacks (Liu et al. 2023). However, Rajendran et al. (2024) (Rajendran et al. 2024) argue that cross-cohort cross-category integration — "*the process of combining information from diverse datasets distributed across distinct, secure sites*" — is important for adopting large models in verticals like healthcare. This provides an **opportunity to develop large AI models or vertical-agnostic properties that facilitate adaptability while securely handling data without compromising on key performance metrics**. Training large models for adaptability via approaches like meta-learning (Verma et al. 2024b) or providing *secure* adaptability as a service (Tang et al. 2024) are some opportunities to accelerate vertical adoption of large AI models.

Beyond modeling-related opportunities, there are significant **opportunities on the human-AI interaction and interface design fronts, both in vertical-specific as well vertical-agnostic settings**. The large-scale user-adoption of AI is still relatively nascent (Houter 2024; Abril 2024) and while modeling-related problems (both vertical-specific and vertical-agnostic) are being actively addressed by newer iterations of the large models, the interfacing of these capabilities with the intended users is still a major challenge. For instance, beyond chatbot-like interfaces, recent studies show structured media such as notebooks provide a flexible interface for incrementally creating and consuming information, which is also effective for clinically mandated documentation standards (Cheng et al. 2024; Wang, Dai, and Edwards 2022; Adler-Milstein et al. 2022). Additionally, micro-prompting and using interactive graphical objects (an interaction paradigm that can be applied to many verticals) has shown to enhance user satisfaction in interactions between human and AI systems (Suh et al. 2023; Jiang et al. 2023; Butler et al. 2024).

Communicating cross-disciplinary challenges

Achieving real-world impact with large AI models demands coordinated expertise spanning AI architectures, high-performance hardware systems, data curation, domain knowledge (including regulatory considerations), and human-computer interaction. **Our framework offers a shared language for communicating these varied technical and societal needs**. It helps each group pinpoint bottlenecks and future directions across the layers. For instance, researchers who develop a modeling or interface solution in one vertical can highlight its potential generalizability, bringing it to the attention of others working on similar challenges in different verticals. Relatedly, researchers who develop vertical-agnostic solutions that address challenges across many verticals should persuade their adoption across verticals. If these solutions prove effective repeatedly across many verticals, large-model developers should incorporate them into next-generation models and architectures so they become standard off-the-shelf capabilities.

By helping researchers situate their efforts within a common structure, the framework promotes streamlined collaboration among AI developers, vertical experts, and human-computer interaction researchers, ensuring that large models evolve into useful and trustworthy tools across a wide range of real-world applications. The framework also helps avoid the pitfalls of viewing large AI models as self-sufficient solutions (Blodgett and Madaio 2021; Bommasani et al. 2021). At the same time, it **encourages vertical-specific teams to articulate their domain's unique requirements and allows large-model developers to spot recurring problems that merit general-purpose fixes**.

Discussion

Positioning with respect to other frameworks: As mentioned earlier, most existing frameworks focus on specific facets of adopting large AI models into verticals. For instance, Ehsan et al. (2023) chart the sociotechnical gap in explainable artificial intelligence, presenting a framework that conceptualizes the gap between model outputs and human interpretability needs. Similarly, Goldstein and Sastry (2024) present a framework for estimating the malicious use of systems built with advanced AI models. In the domain of conversational search, Azzopardi et al. (2024) develop a conceptual model of agent–user interaction, mapping out the dialogue acts and decision points that drive information-seeking conversations. Moving beyond facets like explainability and interfacing, at a systems level, Yan et al. (2024) provide a list of engineering-focused best-practices for LLM-based applications, emphasizing components like rigorous evaluation, retrieval-augmented generation, fine-tuning, and guardrails to integrate large models into products. In the same vein, institutions have also put forth adoption guides: a U.S. Department of Energy report (2022) underscores the need for robust data pipelines, high-performance computing, and reproducible model infrastructures to accelerate AI uptake in critical domains. Likewise, the NIST AI Risk Management Framework (2023) focuses on processes for identifying and mitigating AI system risks, defining functions such as Govern, Map, Measure, and Manage to ensure trustworthy development and deployment. Industry frameworks like IBM's "AI Ladder" (2025) similarly prescribe sequential steps (e.g., modernize, collect, organize, analyze, infuse) to guide organizations in scaling AI from data preparation to integration into workflows.

Each of these works offers valuable conceptual frameworks, but notably each isolates a specific sub-problem or component of AI adoption into verticals — be it explainability, user interfacing, engineering best practices, infrastructure needs, or product management cycles — rather than providing a unified framework. In contrast, our layered framework offers a holistic and modular structure that ties these aspects together. It spans from the core large-model layer, through vertical-agnostic properties and vertical-specific adaptation, up to user-facing intermediaries, explicitly situating innovations and challenges in context. Within this integrated view, we present several perspectives that could enable an ecosystem of efficient and effective vertical adoption of AI models — e.g., how improvements in one layer (for example, a robustness technique at the model layer or a novel interfacing paradigm at the user-intermediary layer) can propagate benefits across other layers and across domains. Cross-cutting concerns, such as data privacy, fairness, or trust, are better addressed at right layers such that they are not repeatedly "solved" in silos.

By providing a shared vocabulary and clear abstraction boundaries, the framework enables researchers and practitioners to communicate and build upon each other's advances. Our framework not only advocates to reduce redundant effort (teams in different verticals can reuse solutions or insights from analogous layers) but also establishes a basis for diving deeper into aspect-specific frameworks as needed. For instance, an explainability framework (as in the framework by Ehsan et al.) slots naturally into our vertical-user intermediary layer, and a conversational interaction model (as in Azzopardi et al.'s framework) can be viewed as a vertical-specific interface component. In essence, the layered approach "situates" specialized innovations within a bigger picture. All of these qualities address the unmet need for an integrated conceptual models for AI adoption: one that modularly encompasses the full pipeline of transforming a large AI model into a domain-specific, user-facing system, and thereby complements prior frameworks.

Adaptability with the changing interaction paradigm: While our framework primarily focuses on human-AI interactions, it is inherently adaptable to emerging ways of interactions with large models. For instance, Model Context Protocol (MCP) is an open standard recently designed to facilitate interactions between large models and external tools (Anthropic 2024). As agentic-AI workflows gain prominence, understanding signals and feedback from model-tool interactions becomes crucial for further improvements. However, such signals could differ from traditional measures commonly used in human-AI interaction paradigm (such as user trust) and may include newer measures such as agentic-coordination capabilities, and tool-usage efficiency. Our framework provides the agility to incorporate such changes. For example, in this case, an additional layer parallel to the vertical-user intermediary layer could be introduced which specifically adapted to analyze multi-agent system interactions.

Limitations and future directions: Our framework, while providing a valuable starting point for situating innovations, opportunities and challenges in vertical adoption of large AI models, has multiple limitations. The framework centers on end-users and their direct interactions with AI systems, leaving out broader social, institutional, and policy contexts. In the context of healthcare and well-being, De Choudhury et al. (2023) (De Choudhury, Pendse, and Kumar 2023) adopt a Social Ecological Model to explore how AI can influence not just individuals but also caregiving institutions and society at large. While these wider sociotechnical considerations are crucial for fully understanding AI's impact — especially in sensitive areas like mental health and telehealth — they fall beyond the scope of our current user-focused framework. Future work could integrate our framework with broader ecological models to capture the dynamism across multiple stakeholders and social layers, thereby offering a more holistic view of how AI systems both shape and are shaped by larger sociotechnical contexts. Furthermore, since our framework aims to cover the development pipeline that transforms large AI models into vertical systems, it does not dive deep into individual challenges. For instance, prior studies have developed frameworks around specific aspects like robustness, interpretability, explainability, and human-AI interaction and communication (Li et al. 2023; Fragiadakis et al. 2024; Bansal et al. 2024). Future work could involve a large-scale study that balances high-level coverage of the entire pipeline with deeper dives into each sub-issue. This would require a multi-disciplinary collaboration of experts in those sub-issues, ensuring that the final systems remain grounded in real user needs while thoroughly addressing the nuanced challenges in each aspect of deployment.

Conclusion

Our framework addresses the crucial gap between the remarkable capabilities of large AI models and the complex, context-specific needs that arise in real-world verticals. By modularizing the development pipeline into four layers — large AI models, vertical-agnostic properties, vertical adaptation, and vertical-user intermediaries, our framework offers a structured way to identify, situate, and address the challenges encountered when building practical AI systems. We also highlighted the dynamic interplay among these layers: insights from domain-specific challenges loop back into model-level improvements, and widespread interface challenges become generalizable solutions for human-AI interactions. To that end, the framework provides actionable guidance for effectively placing innovations, spotting opportunities that might otherwise be overlooked, and coordinating across disciplines to precipitate vertical-specific utility.

Large AI models hold tremendous promise but are not plug-and-play solutions that immediately translate into user-facing impact. Our work underscores the healthy dynamism needed to meaningfully apply these models: it calls on AI developers to incorporate feedback from vertical deployments (instead of merely chasing benchmarks), and it urges vertical-focused researchers to recognize where their innovations could broaden into vertical-agnostic improvements. Ultimately, this coordinated and interdisciplinary effort will ensure that large AI models truly deliver on users' needs.

References

Abril, D. 2024. Meet the 'super users' who tap AI to get ahead at work. https://www.washingtonpost.com/technology/2024/10/28/ai-work-superusers/. Accessed: 2025-01-29.

Adler-Milstein, J.; Aggarwal, N.; Ahmed, M.; Castner, J.; Evans, B. J.; Gonzalez, A. A.; James, C. A.; Lin, S.; Mandl, K. D.; Matheny, M. E.; et al. 2022. Meeting the moment: addressing barriers and facilitating clinical adoption of artificial intelligence in medical diagnosis. *NAM perspectives*, 2022.

Agarwal, N.; Moehring, A.; Rajpurkar, P.; and Salz, T. 2023. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research.

Anthropic. 2024. Introducing the Model Context Protocol — anthropic.com. https://www.anthropic.com/news/model-context-protocol. [Accessed 22-05-2025].

Azzopardi, L.; Dubiel, M.; Halvey, M.; and Dalton, J. 2024. A Conceptual Framework for Conversational Search and Recommendation: Conceptualizing Agent-Human Interactions During the Conversational Search Process. *arXiv preprint arXiv:2404.08630*.

Bahri, Y.; Dyer, E.; Kaplan, J.; Lee, J.; and Sharma, U. 2024. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27): e2311878121.

Bairi, R.; Sonwane, A.; Kanade, A.; Iyer, A.; Parthasarathy, S.; Rajamani, S.; Ashok, B.; and Shet, S. 2024. Codeplan: Repository-level coding using LLMs and planning. *Proceedings of the ACM on Software Engineering*, 1(FSE): 675–698.

Bajpai, Y.; Chopra, B.; Biyani, P.; Aslan, C.; Coleman, D.; Gulwani, S.; Parnin, C.; Radhakrishna, A.; and Soares, G. 2024. Let's Fix this Together: Conversational Debugging with GitHub Copilot. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 1–12. IEEE.

Bannur, S.; Hyland, S.; Liu, Q.; Perez-Garcia, F.; Ilse, M.; Castro, D. C.; Boecking, B.; Sharma, H.; Bouzid, K.; Thieme, A.; et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15016–15027.

Bansal, G.; Vaughan, J. W.; Amershi, S.; Horvitz, E.; Fourney, A.; Mozannar, H.; Dibia, V.; and Weld, D. S. 2024. Challenges in Human-Agent Communication. *arXiv preprint arXiv:2412.10380*.

Barenkamp, M.; Rebstadt, J.; and Thomas, O. 2020. Applications of AI in classical software engineering. *AI Perspectives*, 2(1): 1.

BharatGPT. 2025. BharatGPT - India's Very Own Large Language Model. https://bharatgpt.ai/. Accessed: 2025-01-29.

Bhayana, R. 2024. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*, 310(1): e232756.

Blodgett, S. L.; and Madaio, M. 2021. Risks of AI foundation models in education. *arXiv preprint arXiv:2110.10024*.

Bo, J. Y.; Wan, S.; and Anderson, A. 2024. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. *arXiv preprint arXiv:2412.15584*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1): 1–21.

Butler, J.; Vorvoreanu, M.; Janßen, R.; Sellen, A.; Immorlica, N.; Hecht, B.; and Teevan, J. 2024. Microsoft New Future of Work Report 2024. Technical Report MSR-TR-2024-56, Microsoft Research.

Cacioppo, J. T.; and Petty, R. E. 1982. The need for cognition. *Journal of personality and social psychology*, 42(1): 116.

Carenini, G. 2001. An analysis of the influence of need for cognition on dynamic queries usage. In *CHI'01 extended abstracts on Human factors in computing systems*, 383–384.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.

Chandra, M.; Sriraman, S.; Verma, G.; Khanuja, H. S.; Campayo, J. S.; Li, Z.; Birnbaum, M. L.; and De Choudhury, M. 2024. Lived Experience Not Found: LLMs Struggle to Align with Experts on Addressing Adverse Drug Reactions from Psychiatric Medication Use. *arXiv preprint arXiv:2410.19155*.

Cheng, R.; Barik, T.; Leung, A.; Hohman, F.; and Nichols, J. 2024. BISCUIT: Scaffolding LLM-Generated Code with Ephemeral UIs in Computational Notebooks. *arXiv preprint arXiv:2404.07387*.

Chollet, F. 2024. OpenAI o3 breakthrough high score on ARC-AGI-pub.

CoRover AI. ???? A Conversational AI Platform. https://corover.ai/. Accessed: 2025-01-29.

De Choudhury, M.; Pendse, S. R.; and Kumar, N. 2023. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*.

De Freitas, J.; Uğuralp, A. K.; Oğuz-Uğuralp, Z.; and Puntoni, S. 2024. Chatbots and mental health: Insights into the safety of generative AI. *Journal of Consumer Psychology*, 34(3): 481–491.

Deng, R.; Cui, C.; Liu, Q.; Yao, T.; Remedios, L. W.; Bao, S.; Landman, B. A.; Wheless, L. E.; Coburn, L. A.; Wilson, K. T.; et al. 2023. Segment anything model (SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1): 114.

Ehsan, U.; Saha, K.; De Choudhury, M.; and Riedl, M. O. 2023. Charting the sociotechnical gap in explainable AI: A framework to address the gap in XAI. *Proceedings of the ACM on human-computer interaction*, 7(CSCW1): 1–32.

Elmeleegy, A.; Raj, S.; Slechta, B.; and Mehta, V. 2024. Demystifying AI Inference Deployments for Trillion Parameter Large Language Models. https://developer.nvidia.com/blog/demystifying-ai-inference-deployments-for-trillion-parameter-large-language-models/. Accessed: 2025-01-29.

Foster, D.; McLemore, C.; Olszewski, B.; Chaudhry, A.; Cooper, E.; Forcier, L.; and Luckin, R. 2023. EdTech Quality Frameworks and Standards Review: DfE Quality Characteristics Project (ref: PQFFSR). *UK Department for Education*.

Fragiadakis, G.; Diou, C.; Kousiouris, G.; and Nikolaidou, M. 2024. Evaluating Human-AI Collaboration: A Review and Methodological Framework. *CoRR*.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488.

Ghosh, S.; Chakraborty, P.; Cohn, E.; Brownstein, J. S.; and Ramakrishnan, N. 2016. Characterizing diseases from unstructured text: A vocabulary driven word2vec approach. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, 1129–1138.

Goldstein, J. A.; and Sastry, G. 2024. The PPOu framework: A structured approach for assessing the likelihood of malicious use of advanced AI systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 503–518.

Houter, K. D. 2024. AI in the Workplace: Answering 3 Big Questions. https://www.gallup.com/workplace/651203/workplace-answering-big-questions.aspx. Accessed: 2025-01-29.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiang, P.; Rayan, J.; Dow, S. P.; and Xia, H. 2023. Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–20.

Jin, Y.; Chandra, M.; Verma, G.; Hu, Y.; De Choudhury, M.; and Kumar, S. 2024. Better to ask in English: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM on Web Conference 2024*, 2627–2638.

Johri, S.; Jeong, J.; Tran, B. A.; Schlessinger, D. I.; Wongvibulsin, S.; Barnes, L. A.; Zhou, H.-Y.; Cai, Z. R.; Van Allen, E. M.; Kim, D.; et al. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 1–10.

Jurenka, I.; Kunesch, M.; McKee, K. R.; Gillick, D.; Zhu, S.; Wiltberger, S.; Phal, S. M.; Hermann, K.; Kasenberg, D.; Bhoopchand, A.; et al. 2024. Towards responsible development of generative AI for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*.

Katsarou, E.; Wild, F.; Sougari, A.-M.; and Chatzipanagiotou, P. 2023. A systematic review of voice-based intelligent virtual agents in EFL education. *International Journal of Emerging Technologies in Learning (iJET)*, 18(10): 65–85.

Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Kim, S. S.; Liao, Q. V.; Vorvoreanu, M.; Ballard, S.; and Vaughan, J. W. 2024. "I'm Not Sure, But...": Examining the

Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 822–835.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kumar, S.; Kholkar, G.; Mendke, S.; Sadana, A.; Agrawal, P.; and Dandapat, S. 2024. Socio-Culturally Aware Evaluation Framework for LLM-Based Content Moderation. *arXiv preprint arXiv:2412.13578*.

Kusupati, A.; Bhatt, G.; Rege, A.; Wallingford, M.; Sinha, A.; Ramanujan, V.; Howard-Snyder, W.; Chen, K.; Kakade, S.; Jain, P.; et al. 2022. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35: 30233–30249.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9): 1–46.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 022. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*.

Lin, C.-C.; Huang, A. Y.; and Lu, O. H. 2023. Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments*, 10(1): 41.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; Wang, K.; and Liu, Y. 2023. Jailbreaking ChatGPT via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Logg, J. M.; Minson, J. A.; and Moore, D. A. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151: 90–103.

Ma, Z.; Mei, Y.; Bruderlein, C.; Gajos, K. Z.; and Pan, W. 2024. "ChatGPT, Don't Tell Me What to Do": Designing AI for Context Analysis in Humanitarian Frontline Negotiations. *arXiv preprint arXiv:2410.09139*.

Macina, J.; Daheim, N.; Wang, L.; Sinha, T.; Kapur, M.; Gurevych, I.; and Sachan, M. 2023. Opportunities and Challenges in Neural Dialog Tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2357–2372.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Moon, S.; Madotto, A.; Lin, Z.; Nagarajan, T.; Smith, M.; Jain, S.; Yeh, C.-F.; Murugesan, P.; Heidari, P.; Liu, Y.; et al. 2024. Anymal: An efficient and scalable any-modality augmented language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1314–1332.

National Institute of Standards and Technology (NIST). 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report, National Institute of Standards and Technology, Gaithersburg, MD. Version 1.0, released January 26, 2023.

Nguyen, X.-P.; Zhang, W.; Li, X.; Aljunied, M.; Tan, Q.; Cheng, L.; Chen, G.; Deng, Y.; Yang, S.; Liu, C.; et al. 2023. SeaLLMs–Large Language Models for Southeast Asia. *arXiv preprint arXiv:2312.00738*.

Niknazar, M.; Haley, P. V.; Ramanan, L.; Truong, S. T.; Shrinivasan, Y.; Bhowmick, A. K.; Dey, P.; Jagmohan, A.; Maheshwari, H.; Ponoth, S.; et al. 2024. Building a domain-specific guardrail model in production. *arXiv preprint arXiv:2408.01452*.

Nookala, V. P. S.; Verma, G.; Mukherjee, S.; and Kumar, S. 2023. Adversarial Robustness of Prompt-based Few-Shot Learning for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2196–2208.

OpenAI. 2024. New Embedding Models and API Updates. Accessed: 2025-02-07.

Pan, W.; Xu, Z.; Rajendran, S.; and Wang, F. 2024. An adaptive federated learning framework for clinical risk prediction with electronic health records from multiple hospitals. *Patterns*, 5(1).

Parham, J.; Crall, J.; Stewart, C.; Berger-Wolf, T.; and Rubenstein, D. I. 2017. Animal population censusing at scale with citizen science and photographic identification. In *AAAI spring symposium-technical report*.

Pawar, S.; Park, J.; Jin, J.; Arora, A.; Myung, J.; Yadav, S.; Haznitrama, F. G.; Song, I.; Oh, A.; and Augenstein, I. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Peng, S.; Chen, P.-Y.; Hull, M.; and Chau, D. H. 2024. Navigating the Safety Landscape: Measuring Risks in Finetuning Large Language Models. *arXiv preprint arXiv:2405.17374*.

Phan, L.; Gatti, A.; Han, Z.; Li, N.; Hu, J.; Zhang, H.; Shi, S.; Choi, M.; Agrawal, A.; Chopra, A.; et al. 2025. Humanity's Last Exam. *arXiv preprint arXiv:2501.14249*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2024. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *The Twelfth International Conference on Learning Representations*.

Qian, C.; and Wexler, J. 2024. Take It, Leave It, or Fix It: Measuring Productivity and Trust in Human-AI Collaboration. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 370–384.

Quer, G.; and Topol, E. J. 2024. The potential for large language models to transform cardiovascular medicine. *The Lancet Digital Health*, 6(10): e767–e771.

Rajendran, S.; Pan, W.; Sabuncu, M. R.; Chen, Y.; Zhou, J.; and Wang, F. 2024. Learning across diverse biomedical data modalities and cohorts: Challenges and opportunities for innovation. *Patterns*.

Rajpurkar, P.; O'Connell, C.; Schechter, A.; Asnani, N.; Li, J.; Kiani, A.; Ball, R. L.; Mendelson, M.; Maartens, G.; van Hoving, D. J.; et al. 2020. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ digital medicine*, 3(1): 115.

Ramshetty, S.; Verma, G.; and Kumar, S. 2023. Cross-Modal Attribute Insertions for Assessing the Robustness of Vision-and-Language Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15974–15990.

Redmon, J. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Schillaci, Z. 2024. LLM Adoption Trends and Associated Risks. In *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, 121–128. Springer Nature Switzerland Cham.

Sclar, M.; Choi, Y.; Tsvetkov, Y.; and Suhr, A. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Seah, J. C.; Tang, C. H.; Buchlak, Q. D.; Holt, X. G.; Wardman, J. B.; Aimoldin, A.; Esmaili, N.; Ahmad, H.; Pham, H.; Lambert, J. F.; et al. 2021. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health*, 3(8): e496–e506.

Sharma, N.; Liao, Q. V.; and Xiao, Z. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–17.

Shi, T.; Wang, Z.; Yang, L.; Lin, Y.-C.; He, Z.; Wan, M.; Zhou, P.; Jauhar, S. K.; Xu, X.; Song, X.; et al. 2024. WildFeedback: Aligning LLMs With In-situ User Interactions And Feedback. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; and Cilar, L. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5): e1379.

Strubell, E.; Ganesh, A.; and McCallum, A. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13693–13696.

Stryker, C. 2025. The 5 Biggest AI Adoption Challenges for 2025. Accessed: 2025-05-09.

Suh, S.; Min, B.; Palani, S.; and Xia, H. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–18.

Tang, X.; Shin, R.; Inan, H. A.; Manoel, A.; Mireshghallah, F.; Lin, Z.; Gopi, S.; Kulkarni, J.; and Sim, R. 2024. Privacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation. In *The Twelfth International Conference on Learning Representations*.

Tanno, R.; Barrett, D. G.; Sellergren, A.; Ghaisas, S.; Dathathri, S.; See, A.; Welbl, J.; Lau, C.; Tu, T.; Azizi, S.; et al. 2024. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 1–10.

Team, L.; Modi, A.; Veerubhotla, A. S.; Rysbek, A.; Huber, A.; Wiltshire, B.; Veprek, B.; Gillick, D.; Kasenberg, D.; Ahmed, D.; et al. 2024. LearnLM: Improving Gemini for Learning. *arXiv preprint arXiv:2412.16429*.

Toner-Rodgers, A. 2024. Artificial intelligence, scientific discovery, and product innovation. *arXiv preprint arXiv:2412.17866*.

Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9568–9578.

Trotsyuk, A. A.; Waeiss, Q.; Bhatia, R. T.; Aponte, B. J.; Heffernan, I. M.; Madgavkar, D.; Felder, R. M.; Lehmann, L. S.; Palmer, M. J.; Greely, H.; et al. 2024. Toward a framework for risk mitigation of potential misuse of artificial intelligence in biomedical research. *Nature Machine Intelligence*, 1–8.

Tu, T.; Palepu, A.; Schaekermann, M.; Saab, K.; Freyberg, J.; Tanno, R.; Wang, A.; Li, B.; Amin, M.; Tomasev, N.; et al. 2024. Towards Conversational Diagnostic AI. *arXiv preprint arXiv:2401.05654*.

U.S. Department of Energy. 2022. Artificial Intelligence & Technology Office FY22 Program Plan and FY23 Forecast. Technical report, U.S. Department of Energy, Washington, D.C. Digitally signed by Pamela K. Isom, Director, Artificial Intelligence and Technology Office.

Verma, G.; Choi, M.; Sharma, K.; Watson-Daniels, J.; Oh, S.; and Kumar, S. 2024a. Cross-modal projection in multimodal LLMs doesn't really project visual attributes to textual space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 657–664.

Verma, G.; Kaur, R.; Srishankar, N.; Zeng, Z.; Balch, T.; and Veloso, M. 2024b. AdaptAgent: Adapting multimodal web agents with few-shot learning from human demonstrations. *arXiv preprint arXiv:2411.13451*.

Verma, G.; Mujumdar, R.; Wang, Z. J.; De Choudhury, M.; and Kumar, S. 2022a. Overcoming language disparity in online content classification with multimodal learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1040–1051.

Verma, G.; Vinay, V.; Rossi, R.; and Kumar, S. 2022b. Robustness of Fusion-based Multimodal Classifiers to Cross-Modal Content Dilutions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 360–374.

Wang, Z. J.; Dai, K.; and Edwards, W. K. 2022. StickyLand: Breaking the Linear Presentation of Computational Notebooks. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM.

Xiong, M.; Hu, Z.; Lu, X.; LI, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Yan, E.; Bischof, B.; Frye, C.; Husain, H.; Liu, J.; and Shankar, S. 2024. What We've Learned From A Year of Building with LLMs. *Applied LLMs*.

Yang, E.; and Beil, C. 2024. Ensuring data privacy in AI/ML implementation. *New Directions for Higher Education*, 2024(207): 63–78.

Yu, F.; Endo, M.; Krishnan, R.; Pan, I.; Tsai, A.; Reis, E. P.; Fonseca, E. K. U. N.; Lee, H. M. H.; Abad, Z. S. H.; Ng, A. Y.; et al. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Zhang, K.; Zhou, R.; Adhikarla, E.; Yan, Z.; Liu, Y.; Yu, J.; Liu, Z.; Chen, X.; Davison, B. D.; Ren, H.; et al. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 1–13.

Zhou, J.; Chen, A. Z.; Shah, D.; Reese, L. S.; and De Choudhury, M. 2024. ”It's a conversation, not a quiz”: A Risk Taxonomy and Reflection Tool for LLM Adoption in Public Health. *arXiv preprint arXiv:2411.02594*.

Zhu, T.; Zhang, K.; and Wang, W. Y. 2024. Embracing AI in Education: Understanding the Surge in Large Language Model Use by Secondary Students. *arXiv preprint arXiv:2411.18708*.