PART II

Navigating the AI Frontier: The Team8 Cheat Sheet to Evaluating AI in Healthcare

By:
Assaf Mischari, Managing Partner, Team8
Eyal Eliakim, Head of AI, Team8

TEAM8™

Introduction

Artificial Intelligence is reshaping the landscape of major industries, including healthcare. It offers the potential to enhance patient outcomes, improve administrative efficiency, and reduce costs. However, realizing these benefits requires a rigorous and structured approach to evaluating AI solutions. Given the unique regulatory, ethical, and operational challenges inherent to healthcare, such evaluations must be nuanced, domain-specific, and comprehensive.

While Part 1 of this white paper compared Large Language Models to Traditional Machine Learning, Part II outlines fundamental aspects of evaluating healthcare AI technologies, from setting the right objectives and criteria to ensuring continuous monitoring and compliance.

As the healthcare industry races to adopt AI, many organizations face a flood of competing technologies and promises, often with little clarity about what truly works. To bring structure and insight to this challenge, Team8 developed **the Health Compass for LLMs.** This framework provides a rigorous, actionable approach to evaluating AI's role in healthcare as a particularly high-stakes domain. It's built around three key pillars:

1. A structured methodology for assessing AI solutions
2. A curated set of critical questions that decision-makers should ask to uncover hidden risks and align expectations
3. A practical guide to avoiding the most common pitfalls when selecting and deploying LLM-based solutions.

Throughout this white paper, we integrate the Compass by highlighting must-ask questions, empowering healthcare leaders to make smarter and safer strategic AI decisions.

1 AI Performance Evaluation

Defining Objectives and Evaluation Criteria

Evaluating AI solutions in healthcare requires clearly defined objectives and tailored evaluation criteria. It is essential to distinguish between administrative and clinical use cases, as each involves different validation processes and priorities.
For administrative applications, such as patient conversations, documentation, scheduling, billing, and resource management, the evaluation should focus on improvements in operational efficiency, accuracy in data processing, and the extent to which administrative burdens are reduced. Relevant performance metrics for these types of use cases include time saved, reductions in error rates, derisking, improvements in patient throughput, shorter wait times, and decreased patient frustration.

Clinical use cases, on the other hand, pertain to areas such as diagnosis, treatment planning, and patient monitoring. These require a more stringent focus on patient safety, diagnostic accuracy, and improvements in clinical outcomes. Evaluation metrics for clinical AI tools should include patient safety, diagnostic accuracy, clinical outcome improvements, alongside classic mathematical metrics like sensitivity, specificity, and precision when appropriate. Furthermore, the ability of the solution to integrate into existing clinical workflows and meet regulatory and ethical standards is essential. Given the complexity and critical nature of clinical environments, comparing AI outputs with the performance of human experts can provide an additional layer of validation, provided that the comparisons are made under controlled and fair conditions.

Benchmarks

Robust benchmarking is a vital component of AI solution evaluation in healthcare. Selecting the appropriate benchmark datasets is key to assessing the reliability, safety, and effectiveness of a proposed AI system. These datasets must be relevant to the intended use, of high quality, diverse and representative of real-world populations, as well as compliant with data privacy regulations.

For clinical applications, several benchmark datasets are widely recognized. The MIMIC Critical Care Database includes de-identified health data from intensive care patients and supports the development of models for critical care environments. MedQA and MedMCQA assess AI reasoning and general medical knowledge through exam-style questions. CheXpert is a large-scale dataset of chest X-rays used for image interpretation tasks, while the UK Biobank offers comprehensive health and genetic data. MedPerf is a federated platform that allows benchmarking across multiple datasets without compromising data privacy. OpenAI just recently released HealthBench, an open-source benchmark comprising 5,000 real-world medical conversations, co-created with 262 physicians across 60 countries.
Benchmarking for administrative tasks in healthcare remains relatively underdeveloped. As such, organizations may need to curate custom datasets to effectively evaluate these use cases. Nevertheless, several resources can provide valuable insights into administrative performance. The Healthcare Cost and Utilization Project (HCUP) includes data on hospital operations and billing, while the MGMA DataDive offers benchmarking information on productivity and compensation in medical group management. Axiom™ Comparative Analytics includes detailed financial and operational data from over 1,000 hospitals. Additionally, benchmarks for tasks such as PII/PHI de-identification, EHR querying, and medical note generation are publicly available as well.

Key Evaluation Questions for Objectives and Use Case Alignment:

Use-Case Clarity

- What specific problem does your AI solution address (administrative vs. clinical)?
- How does your solution improve existing processes (e.g., reduce time, reduce costs, enhance patient outcomes)?

Value Proposition

- Can you clearly outline the expected ROI (efficiency gains, cost savings, or clinical outcome improvements)?

Performance Metrics

- Which statistical metrics do you track and why (e.g., sensitivity, specificity, etc.)?
- Which healthcare-oriented metrics are measured (e.g., diagnostic accuracy, length of stay, etc.)?

Limitations & Failure Modes

- What are the known limitations of your solution?
- When does performance degrade and how is it mitigated?

Key Evaluation Questions for Benchmarking and Comparison to Human Experts:

Benchmarking Datasets

- Which benchmarks have you used to validate performance (e.g., MIMIC, MedQA, etc.)?
- Have you used other custom datasets to evaluate your solutions?
- Do the datasets represent the target population and its diversity?

Comparison to Human Experts

- Has the model's performance been compared to human experts?
- How were these comparisons conducted to ensure fair evaluation?
- Which population of human experts was selected?
- Do you mitigate that?

Clinical Validation

To ensure that AI solutions designed for direct patient care are both safe and effective, clinical validation is a critical requirement. This process typically includes prospective studies that evaluate how the AI tool performs when deployed within actual clinical workflows. Such studies help to determine the tool's impact on patient outcomes, safety, and usability.

Regulatory approval is another essential step. Engagement with regulatory bodies such as the U.S. Food and Drug Administration (FDA) or the European Medicines Agency (EMA) ensures that the AI system meets the necessary legal and clinical standards for use in patient care. As mentioned in Part I, the FDA has approved few, if any, GenAI technologies for clinical use thus far. However, the FDA has made steps to incorporate LLMs within its approval process, the latest of which is the announcement of "Elsa".

Following deployment, continuous post-market surveillance must be conducted to monitor the AI's performance over time. This includes detecting any emerging issues, evaluating ongoing effectiveness, and ensuring that safety standards are upheld. These validation steps must meet the same rigorous standards applied to any other clinical process.

> **"We must hold healthcare AI to the same standards we expect from any critical system: measurable impact, clinical accountability, and continuous monitoring."**
>
> **Sachin Jain**
> President and CEO, SCAN Group & Health Plan

Key Evaluation Questions for Clinical Use:

Clinical Studies & Real-World Evidence

- Have real-world clinical studies been conducted to validate efficacy? If so, what were the outcomes of these studies?
- If you haven't yet conducted a clinical study, are there other case studies showing a quantifiable impact?
- Do you offer pilot programs or proofs-of-concept?

Post-Market Surveillance

- How is the solution's post-deployment performance monitored?
- What processes are in place to address any adverse events (e.g., hallucinations) or safety concerns?

2 Understanding Adaptation of AI Technology to Healthcare Use Cases

Domain-Specific Adaptations

AI solutions intended for healthcare often require adaptation of general-purpose AI. These adaptations enhance performance and relevance for domain-specific tasks. There are several methods of adaptation, each carrying implications for evaluation.

Organizations usually begin by "reshaping the conversation" rather than the model. Advanced prompt engineering tactics, domain-specific instruction templates, and carefully curated negative examples, can push out-of-the-box LLMs to use appropriate clinical terminology, cite authoritative guidelines, and reveal step-by-step reasoning without any code changes.

Another adaptation method involves providing LLMs with domain-specific knowledge bases. This could be done by adding context to a prompt or by developing a retrieval-augmented generation (RAG) layer that pipes in trusted data. These methods enable even a general-purpose LLM to ground its answers in verifiable facts and trusted, highly relevant sources, dramatically reducing hallucinations while keeping the infrastructure lightweight.

If deeper adaptation is needed, teams progress to post training techniques. One of which, supervised fine tuning - the process of teaching an existing AI model to perform a specific task effectively by showing it examples of correct behavior. It starts with a foundation model that has learned general language patterns from broad data. A dataset of input-output pairs, often curated from real-world examples, expert demonstrations, or synthetic data is then used to train the model further, but only on the specific task. The goal is to retain the model's general capabilities while making it much more reliable and relevant in the domain of interest.
An additional post training technique is reinforcement learning (RL) - a learning paradigm where an AI model interacts with an environment by taking actions, receiving reward signals in response, and updating its policy to maximize these rewards. Within RL, there's reinforcement learning from human feedback (RLHF) - A technique that augments standard RL

by replacing or guiding the reward signal with a model trained on human (expert) preferences, allowing the AI to align its behavior with human values or task-specific expectations even in complex domains.

When examining AI solutions that underwent one or more of these adaptation techniques, evaluators should understand which additional datasets were used, how these datasets were collected and curated, what post training processes techniques were utilized and if modifications were made to the model architecture, as well as request visibility into any advanced prompting methods.

Compliance, Ethics & Privacy

Compliance with healthcare regulations and ethical standards is a non-negotiable requirement for AI solutions in this domain. However, modern AI technologies, particularly those using large language models (LLMs), present new challenges.

Transparency is a significant concern. Foundation models are often developed by external labs, with limited visibility into their training data or processes. This lack of insight complicates efforts to validate these models for clinical use.

Evaluation of modern AI systems is further complicated by their non-deterministic outputs and the absence of universally accepted standards. Organizations must ensure consistency in their evaluation approaches and compare solutions against recognized benchmarks whenever possible.

Explainability remains an ongoing issue. LLMs can generate outputs that are difficult to interpret, and existing explainability tools often provide probabilistic rather than deterministic insights.

Privacy and security concerns are elevated in modern AI workflows. These solutions often require data transfers at multiple stages, increasing exposure to risks such as unauthorized access to personally identifiable information (PII) or protected health information (PHI). Organizations should ensure that robust data protection measures are in place, such as encryption techniques and strict access controls.

Ethical and legal considerations also extend to the risk of hallucinations, biases, and irreproducible outputs. Evaluators must inquire how these challenges are addressed, including the application of guardrails and the implementation of safeguards for demographic fairness.

Additionally, relying on third-party models introduces intellectual property and data ownership ambiguity. Organizations should clarify who owns the training data, models, and generated outputs.

Finally, there is ongoing uncertainty around regulatory requirements. Definitions of what qualifies as a medical device and acceptable standards can vary significantly between jurisdictions. AI systems should be built with adaptability in mind to accommodate evolving policies and standards.

Key Evaluation Questions for Healthcare Adaptation:

Method of Adaptation

- How does your solution differ from other solutions for similar use cases? Does your adapted solution outperform general-purpose AI systems (e.g., ChatGPT, Claude)?
- What adaptation method did you use? Was the solution created using post-training techniques and/or by utilizing external knowledge bases?
- Were domain-specific datasets used for adaptation, and what were their sources? How were data quality and annotation standards ensured?

Post training techniques

- Which post-training methods were used to adapt the model to healthcare? Was the foundation model's architecture modified?

Advanced Prompt Engineering

- Which prompting techniques were used and what information was included to enhance performance on the healthcare-related task?

External Knowledge Bases

- How does the model access external knowledge bases (e.g., RAG)?

Key Evaluation Questions on Compliance, Ethics & Privacy:

Regulatory Status

- Is the AI solution considered a medical device?
- Is your solution approved by regulatory bodies (FDA/CE)?
- If not yet approved, what is the current regulatory approval status?
- How do you ensure HIPAA or GDPR compliance?
- Have you conducted any ethical reviews or impact assessments?

Data Security & Patient Privacy

- How is patient data protected? (e.g., encryption, de-identification)? Is securing data (at rest) addressed alongside AI usage (i.e., data leakage via prompts)?
- Does your solution require data transfer outside the organization? If so, what measures are in place to secure data in transit? Is the receiving organization certified to process PII/PHI?

Risk of Bias & Fairness

- How do you identify and mitigate biases in the training data or model outputs?
- Is there a monitoring mechanism for demographic biases or discrimination?

Transparency & Explainability

- How do you provide transparency into the AI decision-making process? What methods are used to justify AI outputs to end-users?

Intellectual Property & Data Ownership

- Who owns the solution's data, model, applications(or other components)?

3 Workflow Integration and Human-AI Collaboration

Workflow Compatibility

An AI solution's compatibility with existing healthcare workflows is critical to its success. Stakeholders must determine whether a solution improves an existing process or proposes a new workflow altogether. In either case, operational and strategic alignment is crucial.

Evaluation should begin with a needs assessment to identify inefficiencies or gaps the AI tool could address. It is also essential to consider the skillset of end users and evaluate the expected learning curve. Scalability and usability analyses should be conducted to determine whether the tool can be rolled out across departments with minimal disruption.

Certain use cases require modality-specific considerations. For example, clinical applications involving medical imaging may necessitate solutions that handle visual data, while patient-facing tools may require voice interfaces.

"Scalable AI in healthcare requires more than an algorithm—it demands infrastructure that is interoperable, secure, and deeply aligned with frontline workflows."

Nader Mherabi
Chief Digital and Information Officer, Executive Vice President, Vice Dean at NYU Langone

Human-AI Interaction

While AI technology will inevitably replace certain tasks and even professions entirely. However, at this point in time, given the current state of AI technology, the primary focus should be on augmentation and automation of small, well-defined tasks. Furthermore, replacing/automating or augmenting aren't binary states; there's an entire continuum of different implementations between them. Current AI technologies should be considered through this lens, and our Team8 compass certainly takes this approach. We recommend examining solutions and how they're placed on the augmentation-automation continuum, taking into account both sensitivity of use case and sensibility of the technical implementation (see the section on adaptation methods above). For clinical applications, which are likely considered highly sensitive, this means leaning towards augmentation and responsible AI practices, including delivering explainable outputs and intuitive interfaces that foster trust. Systems should allow for human override and include training protocols to ensure end users are equipped to use them effectively.

Feedback mechanisms should be integrated to enable iterative improvement of the AI system based on real-world use.

Key Evaluation Questions for Workflow Integration:

Integration with Existing Systems

- Can the solution integrate with our EHR or other existing IT systems?
- Which interoperability standards (e.g., HL7, FHIR) are supported?

Implementation & Scalability

- How does the intended solution meet the task's practical needs? Keep in mind the uniqueness and sensitivity tasks that are clinical/time-sensitive/patient-facing.
- How is minimal disruption to current workflows ensured?
- Can the solution easily scale across multiple departments or sites?

User Training & Change Management

- What is the user learning curve?
- What training and ongoing support is provided?

Key Evaluation Questions for Human AI Interaction:

Human-AI Interaction

- Are the user interface and outputs intuitive, clear, and explainable?
- Are there mechanisms for human override or second opinion?

4 Continuous Monitoring

Performance Monitoring

Once deployed, AI solutions must be continually monitored to avoid performance degradation and ensure ongoing relevance. This includes periodic assessment against initial benchmarks and test sets, constant adaptation of test sets based on an evolving reality, coverage of extreme cases, collection of user feedback, and comparison with emerging new models and methods where applicable.

Technical Oversight

Modern AI tools may suffer from technical limitations such as latency, high throughput demands, or prohibitive computational costs. These factors must be closely observed, especially in time-sensitive clinical environments.

Adaptability

Effective AI systems must be designed for adaptability. This includes maintaining the ability to update system components, applying version control, and instituting change management protocols. In addition, solutions must be responsive to shifts in regulatory landscapes, workflow needs, and clinical guidelines.

"The AI that will earn clinicians' trust isn't just the one that makes bold claims—it's the one that continuously proves its value through real-world outcomes, built on a foundation of transparent, evidence-based collaboration between researchers, technologists, and clinicians."

Dan Knecht
Chief Medical Officer, Emblem Health

Key Evaluation Questions for Continuous Monitoring:

Performance Monitoring

- How often is model performance re-evaluated?
- Which metrics/KPIs are monitored continuously? Do post-deployment metrics differ from pre-deployment ones, and why?

Model Drift & Updates

- How do you detect and address model drift (e.g., performance decline over time)? How are concept and data drift differentiated?
- Are version control or regular updates offered by the vendor for components under its control? How are third-party controlled system components addressed?

Support & Maintenance

- What's the SLA (Service Level Agreement) for technical support and downtime? How quickly are bugs or performance issues addressed?

Adaptability

- Will additional functionalities or new healthcare-related tasks be deployed?
- How is user feedback and evolving healthcare guidelines incorporated into future development?

Conclusion

Deploying AI in healthcare demands more than technical excellence—it requires trust, transparency, alignment with patient care goals, and workflows specific to this environment. This white paper provides a rigorous framework for evaluating healthcare AI solutions, ensuring that stakeholders ask the right questions and demand the right standards. Whether clinical or administrative, AI systems must demonstrate value, integrate seamlessly, and evolve safely within healthcare's complex ecosystem to ensure organizations make the most of this groundbreaking technology.

The Team8 Cheat Sheet to Evaluating AI in Healthcare

Key Evaluation Questions for Objectives and Use Case Alignment:

Use-Case Clarity

- What specific problem does your AI solution address (administrative vs. clinical)?
- How does your solution improve existing processes (e.g., reduce time, reduce costs, enhance patient outcomes)?

Value Proposition

- Can you clearly outline the expected ROI (efficiency gains, cost savings, or clinical outcome improvements)?

Performance Metrics

- Which statistical metrics do you track and why (e.g., sensitivity, specificity, etc.)?
- Which healthcare-oriented metrics are measured (e.g., diagnostic accuracy, length of stay, etc.)?

Limitations & Failure Modes

- What are the known limitations of your solution?
- When does performance degrade and how is it mitigated?

Key Evaluation Questions for Benchmarking and Comparison to Human Experts:

Benchmarking Datasets

- Which benchmarks have you used to validate performance (e.g., MIMIC, MedQA, etc.)?
- Have you used other custom datasets to evaluate your solutions?
- Do the datasets represent the target population and its diversity?

Comparison to Human Experts

- Has the model's performance been compared to human experts?
- How were these comparisons conducted to ensure fair evaluation?
- Which population of human experts was selected?
- Do you mitigate that?

Key Evaluation Questions for Clinical Use:

Clinical Studies & Real-World Evidence

- Have real-world clinical studies been conducted to validate efficacy? If so, what were the outcomes of these studies?
- If you haven't yet conducted a clinical study, are there other case studies showing a quantifiable impact?
- Do you offer pilot programs or proofs-of-concept?

Post-Market Surveillance

- How is the solution's post-deployment performance monitored?
- What processes are in place to address any adverse events (e.g., hallucinations) or safety concerns?

Key Evaluation Questions for Healthcare Adaptation:

Method of Adaptation

- How does your solution differ from other solutions for similar use cases? Does your adapted solution outperform general-purpose AI systems (e.g., ChatGPT, Claude)?
- What adaptation method did you use? Was the solution created using post-training techniques and/or by utilizing external knowledge bases?
- Were domain-specific datasets used for adaptation, and what were their sources? How were data quality and annotation standards ensured?

Post training techniques

- Which post-training methods were used to adapt the model to healthcare? Was the foundation model's architecture modified?

Advanced Prompt Engineering

- Which prompting techniques were used and what information was included to enhance performance on the healthcare-related task?

External Knowledge Bases

- How does the model access external knowledge bases (e.g., RAG)?

Key Evaluation Questions on Compliance, Ethics & Privacy:

Regulatory Status

- Is the AI solution considered a medical device?
- Is your solution approved by regulatory bodies (FDA/CE)?
- If not yet approved, what is the current regulatory approval status?
- How do you ensure HIPAA or GDPR compliance?
- Have you conducted any ethical reviews or impact assessments?

Data Security & Patient Privacy

- How is patient data protected? (e.g., encryption, de-identification)? Is securing data (at rest) addressed alongside AI usage (i.e., data leakage via prompts)?
- Does your solution require data transfer outside the organization? If so, what measures are in place to secure data in transit? Is the receiving organization certified to process PII/PHI?

Risk of Bias & Fairness

- How do you identify and mitigate biases in the training data or model outputs?
- Is there a monitoring mechanism for demographic biases or discrimination?

Transparency & Explainability

- How do you provide transparency into the AI decision-making process? What methods are used to justify AI outputs to end-users?

Intellectual Property & Data Ownership

- Who owns the solution's data, model, applications(or other components)?

Key Evaluation Questions for Workflow Integration:

Integration with Existing Systems

- Can the solution integrate with our EHR or other existing IT systems?
- Which interoperability standards (e.g., HL7, FHIR) are supported?

User Training & Change Management

- What is the user learning curve?
- What training and ongoing support is provided?

Implementation & Scalability

- How does the intended solution meet the task's practical needs? Keep in mind the uniqueness and sensitivity tasks that are clinical/time-sensitive/patient-facing.
- How is minimal disruption to current workflows ensured?
- Can the solution easily scale across multiple departments or sites?

The Team8 Cheat Sheet to Evaluating AI in Healthcare

Key Evaluation Questions for Human AI Interaction:

Human-AI Interaction

- Are the user interface and outputs intuitive, clear, and explainable?
- Are there mechanisms for human override or second opinion?

Key Evaluation Questions for Continuous Monitoring:

Performance Monitoring

- How often is model performance re-evaluated?
- Which metrics/KPIs are monitored continuously? Do post-deployment metrics differ from pre-deployment ones, and why?

Model Drift & Updates

- How do you detect and address model drift (e.g., performance decline over time)? How are concept and data drift differentiated?
- Are version control or regular updates offered by the vendor for components under its control? How are third-party controlled system components addressed?

Support & Maintenance

- What's the SLA (Service Level Agreement) for technical support and downtime? How quickly are bugs or performance issues addressed?

Adaptability

- Will additional functionalities or new healthcare-related tasks be deployed?
- How is user feedback and evolving healthcare guidelines incorporated into future development?