Guide to Evaluating Vendors on AI Capabilities and Offerings

This guide offers questions and rating scales to structure discussions with vendors and IT partners, based on the AAMC's Principles for Responsible Use of AI in Medical School and Residency Selection.

Getting Started

Before you get started, we recommend:

1.) Reviewing the AAMC's Principles for Responsible Use of AI in Medical School and Residency Selection to understand their importance in AI use.
2.) Identifying your institution's priorities based on context and project goals.
3.) Using the Guide to Assessing Your Institution's Readiness for Implementing AI to identify your preparation needs and inform vendor evaluations.
4.) Consulting the Essential AI Terms and Definitions guide to align on key terminology essential for evaluating AI solutions.

How to Use the Guide

While using this guide, we recommend you:

- Use consistent questions across all vendors.
- Assign sections to team members based on expertise (e.g., IT lead questions).

Each section aligns with an AAMC principle. For each section:

1. **Select Relevant Questions**. Choose from the question bank. Follow up for clarification as needed to fully understand vendor capabilities.
2. **Notes**. Document key response details and follow-up conversation.
3. **Detailed Rating Scale**. Using the 3-point scale, check off behaviors/capabilities demonstrated.
4. **Overall Rating Scale**. Assess overall capability: (1) Limited to, (3) Comprehensive.

After vendor discussions, we recommend you:

- Collaborate as a team to create a comprehensive vendor assessment
- Use the Summary Ratings Table as a guide, customizing as needed.
- Remember that ratings are tools for discussion, not definitive measures — weigh principles based on your specific needs.

1. Balance Prediction and Understanding.

An effective AI system should target characteristics linked to student success, as defined by the institution. It must make accurate predictions based on these characteristics while providing clear explanations of its decision-making process to all users.

Questions

- How do you ensure the characteristics measured align with our institution's definition of an effective student or resident?
- How do you balance the complexity of your tool with the need for interpretable results?
- (Follow-up) How does your AI handle different data sources (e.g., academic, clinical, documents) in its decision-making?
- (Follow-up) Can you give an example of making your tool's output understandable to nontechnical users?
- (Follow-up) How do you incorporate our subject matter experts into the model-building and interpretation process?
- (Follow-up) What validation methods and metrics do you use to ensure large language model (LLM) outputs are accurate? Walk us through how you detect and prevent hallucinations or factual errors.

Notes

Use this space to record key points from the vendor's responses:

Detailed Rating Scale

Check off behaviors/capabilities demonstrated.

Limited	Moderate	Comprehensive
☐ Uses unclear methods to identify success characteristics.	☐ **Relies on general industry standards to identify success factors.**	☐ Collaborates with faculty to define program-specific success characteristics.
☐ Lacks research or analysis on success factors specific to the institution.	☐ Conducts basic analysis on success characteristics but lacks depth or relevance to specific programs.	☐ Conducts comprehensive research and analysis on success factors relevant to the institution.
☐ Provides no individual explanations for decisions.	☐ Only one method used to explain decisions and no explanations for specific groups.	☐ Two or more methods used to explain individual decisions and those for specific groups.
☐ Provides no evidence of reliability or validity for its methods.	☐ Briefly speaks to reliability or validity (e.g., AERA, APA, SIOP, NCME).	☐ Demonstrates thorough understanding of reliability and validity (e.g., AERA, APA, SIOP, NCME).

Overall Rating Scale

Combine notes and checkboxes to provide an overall rating.

(1) Limited	(2) Moderate	(3) Comprehensive
☐ Minimal or unclear prediction methods.	☐ Basic industry standards met.	☐ Advanced, institution-specific solutions with clear evidence.

2. Protect Against Algorithmic Bias.

An effective AI system should have a robust set of procedures to define, measure, monitor, and mitigate biases, especially for underrepresented groups in medicine (e.g., low income, rural) and individuals with disabilities.

Questions

- What are the historical biases found in your selection tool (e.g., how are they defined and measured)? How do you prevent biases from affecting your AI tool?
- How does the AI tool ensure fairness for all demographic groups, including underrepresented in medicine?
- (Follow-up) How do you ensure your training data are representative?
- (Follow-up) How do you communicate your bias mitigation efforts to users and applicants?
- (Follow-up) How does your AI system accommodate user needs, including accessibility features and assistive technology compatibility?

Notes

Use this space to record key points from the vendor's responses:

Detailed Rating Scale

Check off behaviors/capabilities demonstrated.

Limited	Moderate	Comprehensive
☐ **No bias testing or fairness metrics used in development or on real-world data.**	☐ **Ad hoc or reactive use of bias testing, unclear timing or cadence.**	☐ Conducts large-scale bias testing for each academic year.
☐ No bias mitigation or correction methods implemented.	☐ Implements some bias mitigation methods, but may rely heavily on newer, less-proven techniques.	☐ Implements robust bias mitigation methods, including both established and carefully vetted newer techniques.
☐ No consideration of demographic representation, especially for underrepresented groups.	☐ Some effort to ensure demographic representation, but gaps remain for underrepresented groups.	☐ Ensures training data and real-world applications fully represent multiple demographic groups, including intersectional.
☐ No accessibility considerations or WCAG compliance.	☐ Basic accessibility features, but not fully WCAG 2.1 compliant.	☐ Fully WCAG 2.1 compliant with robust accessibility features.

Overall Rating Scale

Combine notes and checkboxes to provide an overall rating.

(1) Limited	(2) Moderate	(3) Comprehensive
☐ Minimal or unclear methods for identifying, measuring, and mitigating bias.	☐ Basic industry standards met for bias testing and mitigation.	☐ Advanced, institution-specific solutions with clear evidence of robust bias protection.

3. Provide Notice and Explanation.

An effective AI system should provide clear and comprehensive information to applicants about how AI is used in the selection process.

Questions

- How do you help inform applicants about AI being used in the selection process?
- How do you advise institutions to address applicant concerns about AI being used in the admissions process while also maintaining the integrity of the process?
- How well would you be able to describe the process and explain the selection tools in a potential litigation?
- (Follow-up) What resources or templates do you provide for informing applicants about the use of AI in your selection system?
- (Follow-up) How would you address applicants that do not want to be screened using AI?

Notes

Use this space to record key points from the vendor's responses:

Detailed Rating Scale

Check off behaviors/capabilities demonstrated.

Limited	Moderate	Comprehensive
☐ No disclosure of AI use in the selection process.	☐ Basic disclosure of AI use but lacks detail.	☐ Clear, comprehensive disclosure of how AI is used in selection.
☐ No resources provided to applicants about AI use.	☐ Some resources available but not easily accessible or detailed.	☐ Comprehensive, easily accessible resources explaining AI use to applicants.
☐ No explanation of how AI impacts applicant evaluation.	☐ Basic explanation of AI impact but is not clear who is responsible for it.	☐ Detailed explanation of how AI specifically impacts applicant evaluation.
☐ No information on AI governance provided.	☐ Some information on AI governance but not comprehensive.	☐ Full transparency on AI governance policies and practices.

Overall Rating Scale

Combine notes and checkboxes to provide an overall rating.

(1) Limited	(2) Moderate	(3) Comprehensive
☐ Minimal or unclear AI disclosure practices.	☐ Basic industry standards met for AI transparency.	☐ Advanced, transparent AI communication.

4. Protect Data Privacy.

An effective AI system should have robust data protection measures in place and comply with relevant regulations.

Questions

- How do you ensure applicant data privacy and comply with U.S. guidelines (e.g., the NIST Risk Management Framework), in addition to European regulations (e.g., General Data Protection Regulation)?
- What processes do you recommend for allowing applicants to opt out of AI-assisted evaluation or limit the sharing of their data with external services?
- Can you support an in-house AI tool and database to avoid sharing sensitive data and minimizing the risk of a data breach?
- (Follow-up) How do you manage data sharing with external services or application programming interfaces (APIs), including AI tools like large language models (LLMs)?
- (Follow-up) How do you exceed compliance requirements and incorporate the latest best practices and technologies to protect our data?

Notes

Use this space to record key points from the vendor's responses:

Detailed Rating Scale

Check off behaviors/capabilities demonstrated.

Limited	Moderate	Comprehensive
☐ Minimal or no data protection measures in place for applicant data.	☐ Basic data protection measures in place focusing on compliance, but not detailed.	☐ Robust data protection measures going beyond compliance and incorporating latest best practices.
☐ No process for applicants to exercise their data rights.	☐ Basic process for data rights, but struggles with balancing applicant rights and institutional resources.	☐ Efficient, comprehensive process for applicants to exercise all data rights, with measures to manage excessive requests.
☐ No policies for third-party data sharing or API security measures.	☐ Basic policies exist for third-party data sharing and API security, but unclear how they would be enforced.	☐ Clear assurance and contractual agreements to ensure data protection for all third-party data sharing and robust API security measures.
☐ No specific protections for data used with LLMs or other AI tools.	☐ Some protections for LLM and AI tool data use, but not comprehensive.	☐ Comprehensive safeguards for all data interactions with LLMs and other AI tools.

Overall Rating Scale

Combine notes and checkboxes to provide an overall rating.

(1) Limited	(2) Moderate	(3) Comprehensive
☐ Minimal or unclear data protection methods.	☐ Basic industry standards met for data privacy.	☐ Advanced, secure data protection.

5. Incorporate Human Judgment.

An effective AI system should complement human expertise rather than replace it and provide clear processes for human oversight and intervention.

Questions

- How does the AI system complement human expertise in the admissions and selection process?
- Does the system provide recommendations to help staff focus on certain candidates, or does it make autonomous selections? How would you resolve a disagreement?
- What kind of training and ongoing support is provided for using the AI system?
- (Follow-up) How do you incorporate subject matter experts (e.g., administrative professionals, faculty) into the model-building and interpretation process?
- (Follow-up) What safeguards are in place to prevent over-reliance on AI decisions by human evaluators?

Notes

Use this space to record key points from the vendor's responses:

Detailed Rating Scale

Check off behaviors/capabilities demonstrated.

Limited	Moderate	Comprehensive
☐ AI makes decisions without human involvement in the model-building or decision-making process.	☐ Humans can review AI decisions, but with limited understanding or ability to intervene in the decision-making process.	☐ Seamless integration of AI insights with human decision-making, with clear processes for human involvement, oversight, and intervention.
☐ No mechanism for overriding or appealing AI decisions.	☐ Basic override/appeal process exists, but is cumbersome or unclear.	☐ Clear, efficient processes for reviewing, overriding, and appealing AI decisions.
☐ No formal initial training provided for using the AI system.	☐ Basic initial training provided, but not comprehensive or tailored.	☐ Comprehensive, role-specific initial training provided for all users of the AI system.
☐ No ongoing support provided after initial implementation.	☐ Limited ongoing support available, but not proactive or comprehensive.	☐ Proactive, comprehensive ongoing support, including regular check-ins and updates.

Overall Rating Scale

Combine notes and checkboxes to provide an overall rating.

(1) Limited	(2) Moderate	(3) Comprehensive
☐ Minimal or unclear human oversight methods.	☐ Basic industry standards met for human involvement.	☐ Advanced human-AI integration with institutional support.

6. Monitor and Evaluate.

An effective AI system should have robust processes for continuous improvement and adaptation.

Questions

- How do you ensure adherence to established standards for fairness, performance, and responsible AI practices over time?
- How do you assess the effectiveness and user-friendliness of your training and support services, particularly when adapting to AI system updates?
- What steps do you take for ongoing improvement and alignment with institutional goals?
- (Follow-up) How do you balance standardized, academic years with real-time AI monitoring benefits?
- (Follow-up) What is your process for demonstrating that system improvements lead to better outcomes in the selection process?

Notes

Use this space to record key points from the vendor's responses:

Detailed Rating Scale

Check off behaviors/capabilities demonstrated.

Limited	Moderate	Comprehensive
☐ No regular reviews of AI system performance to catch shifts in fairness, accuracy, or data patterns.	☐ Occasional performance reviews, but not systematic or comprehensive.	☐ Regular, comprehensive performance reviews with clear protocols for addressing shifts in fairness, accuracy, and/or data patterns.
☐ No clear process for incorporating user or institutional feedback.	☐ Some feedback collected, but not systematically incorporated into improvements.	☐ Robust system for collecting and incorporating diverse feedback into ongoing improvements.
☐ No mechanism for ensuring ongoing alignment with institutional goals.	☐ Basic checks for alignment with institutional goals, but not comprehensive or regular.	☐ Regular, in-depth assessments of AI system alignment with evolving institutional goals.

Overall Rating Scale

Combine notes and checkboxes to provide an overall rating.

(1) Limited	(2) Moderate	(3) Comprehensive
☐ Minimal or unclear monitoring methods.	☐ Basic industry standards met for evaluation.	☐ Advanced monitoring with institutional alignment.

Summary Ratings Table

Use this space to compile ratings across all interviewers and questions and calculate a total for each vendor.

Remember, these are tools for comparison and discussion, not definitive measures of an AI tool's suitability. Your institution should decide how to weigh different principles based on your specific needs and goals.

Note that all numbers in the table below are for demonstration purposes only.

Principle	**Ratings**		
	Vendor 1	Vendor 2	Vendor 3
Balance Prediction and Understanding	2	3	2
Protect Against Algorithmic Bias	3	2	1
Provide Notice and Explanation	2	2	2
Protect Data Privacy	1	2	2
Incorporate Human Judgment	3	1	1
Monitor and Evaluate	2	2	2
Total Rating	13	12	10

Rating scale: 1 = Limited; 2 = Moderate; 3 = Comprehensive.

Next steps:

- Identify areas where more information or clarification is needed and set up a second round of questions, if necessary.
- Discuss how the AI tool's strengths and weaknesses align with your insights from the Guide to Assessing Your Institutions Readiness for Implementing AI and institutional priorities.
- Plan for potential implementation, including staff training and integration with existing processes.