April 2025

Tool for Assessing AI Vendors

A resource for decision-makers
in international development,
humanitarian,
and social impact sectors

This tool was developed by Grace Lyn Higdon of Revolution Impact with contributions from Linda Raftree of The MERL Tech Initiative (MTI). It is part of a suite of public good tools developed for MTI's Natural Language Processing Community of Practice.

This is Version 1 of the "Tool for Assessing AI Vendors: A resource for decision-makers in international development, humanitarian, and social impact", published in April 2025.

The Natural Language Processing Community of Practice brings together monitoring, evaluation, research, and learning practitioners, artificial intelligence experts, and data responsibility advocates to learn and collaborate. We focus on responsible, appropriate, and effective applications of NLP (including Generative AI) to address demand-driven, real-world MERL challenges. Visit merltech.org/nlp-cop for more information about this and other resources.

The MERL Tech Initiative (MTI) is a social venture that sits at the intersection of digital technology and the social sector. We support thoughtful tech-enabled program design, implementation, and monitoring, evaluation, research and learning (MERL). We help organizations with responsible design, use, and governance of digital technologies and digital data to achieve better outcomes. MTI convenes and supports the NLP-CoP. Visit merltech.org for more information.

Revolution Impact is a boutique consulting firm working with a wide range of stakeholders who prioritize economic justice and human rights, including public and private foundations, impact investors, funds, INGOs, and civil society networks. Visit revolution-impact.org to learn more.

revolution impact

THE MERL Tech INITIATIVE

NLP Community of Practice

Contents

About this assessment tool

Who we are and why we developed this assessment tool

We are Steering Committee members of The Natural Language Processing Community of Practice (NLP-CoP), which has been exploring the use of generative AI (GenAI) and natural language processing (NLP) since January 2023.[1] Our community has voiced consistent needs for frameworks to evaluate AI tools and services. Our experience spans roles in monitoring, evaluation, research and learning (MERL), program design and implementation, and grant management across various social impact organizations and funding institutions. We are techno-pragmatists — aware of the purported benefits, while attuned to the risks technologies pose, and sensitive to the narratives shaping incentives for increased use.

AI is increasingly being woven into the day-to-day tools most of our organisations use. As a community, we are interested in maintaining a responsible and critical lens when adopting AI-powered tools. We believe a balanced view that neither exaggerates the utility of AI nor avoids it altogether best serves the sector. At the same time, the high-level and practical ethical challenges with AI are becoming more and more apparent. In our role as co-leads of the Ethics and Governance working group, we have been challenged to identify AI tools that meet both quality standards for implementation and ethical standards across the development and supply chain in the creation of AI. That is why we have created this assessment tool.

Who is this assessment tool for?

This assessment tool is designed for decision-makers who work in the international development, humanitarian, or social impact sectors and who need to assess AI vendors but may not have specialized knowledge in AI systems. These could be program managers, MERL professionals, and/or technical staff who are considering AI tool procurement. Organizations with varying levels of technical expertise, including smaller teams with limited technical capacity, may also find this assessment tool useful. Some questions will be more relevant to certain vendor types than others, and as the AI space evolves, the assessment tool will need to evolve as well!

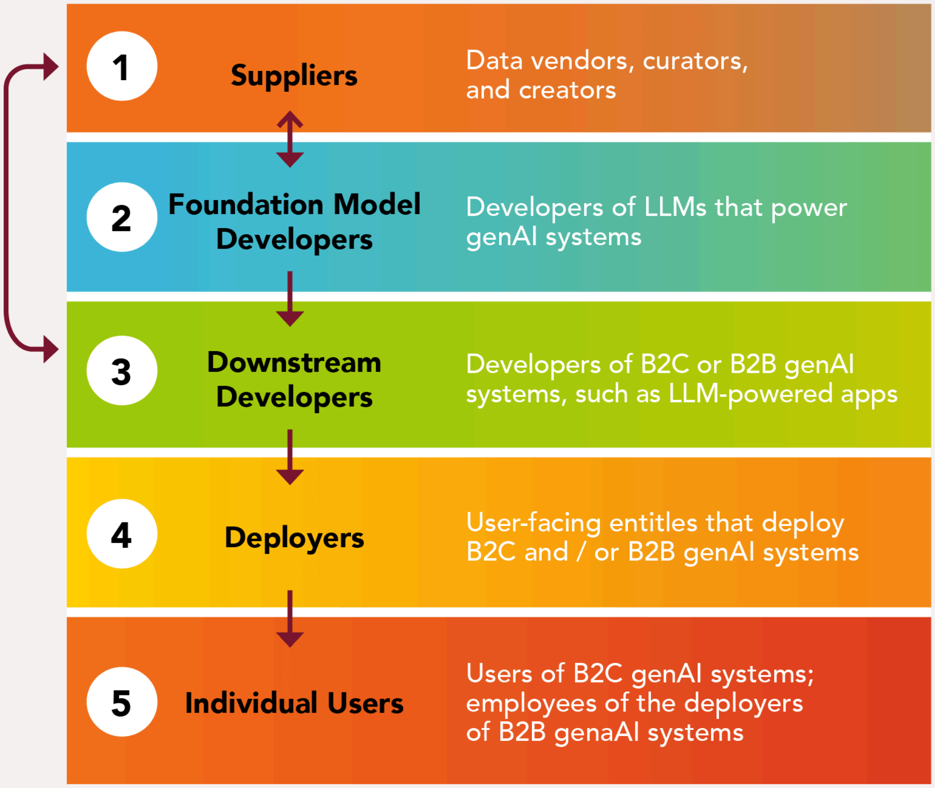
AI vendors offer diverse services that require different assessment questions. This assessment tool covers questions relevant to:

1) Off-the-shelf AI products: AI solutions with fixed capabilities
2) Custom AI development: Bespoke solutions built specifically for your requirements
3) AI integration services: Embedding new AI capabilities into existing systems

[1] More information about the NLP-CoP is available at: https://merltech.org/nlp-cop/

What does this assessment tool aim to do?

This assessment tool aims to provide a straightforward, criteria-based analysis of vendor credibility and implementation track record. In the simplified AI supply chain diagram below, this assessment tool could support conversations with **downstream developers** and **deployers**.[2,3]


1
Suppliers
Data vendors, curators, and creators
2
Foundation Model Developers
Developers of LLMs that power genAI systems
3
Downstream Developers
Developers of B2C or B2B genAI systems, such as LLM-powered apps
4
Deployers
User-facing entitles that deploy B2C and / or B2B genAI systems
5
Individual Users
Users of B2C genAI systems; employees of the deployers of B2B genaAI systems

Suppliers of genAI system components, such as hardware or infrastructure, are not in scope for this HRA.

Some entities straddle multiple categories. For instance, OpenAI is a supplier, a foundation model developer, and a deployer

The double arrows between foundation model / downstream developers and suppliers indicates that these actors work closely, particularly during the training and evaluation process. Both types of developers may go back to suppliers multiple times for additional data to satisfy data needs for model fine-tuning.

Other actors, such as regulators or investors, may be "value-chain adjacent." This means that their activities have impacts across the value chain. They are excluded from this image for simplicity.

Diagram by Hoh, J. Y., Andersen, L., & Darnton, H. (2025).

The assessment tool:

- Focuses on requirements to explore when selecting an AI vendor or partner. Sometimes the vendor or partner will have a specific product they are marketing, sometimes they will be offering bespoke AI-enabled services. The assessment tool aims to partially address both scenarios. It focuses particularly on the need for [‘explainable’ AI](), error detection and validation processes, and mechanisms for human review and override.

[2] Diagram created by BSR. See Hoh, J. Y., Andersen, L., & Darnton, H. (2025). Human Rights Across the Generative AI Value Chain. BSR. Accessed March 2025. https://www.bsr.org/en/reports/human-rights-across-the-generative-ai-value-chain

[3] The diagram does not include evaluation of underlying AI models, as these processes are largely inaccessible to the international development and social impact sectors.

- Can be used to help organizations have a conversation with AI Vendors about what exactly a tool or product can and cannot do. The assessment tool's two dimensions and associated criteria could serve as a rudimentary rubric to be expanded upon, as well as a spring board for an internal conversation to decide which criteria are most important to your context.
- Assumes either some in-house IT expertise or small teams willing to engage with technical aspects around security.
- Surfaces core practical and ethical issues that are within the control of an AI Vendor to alter.
- Can be a useful document for your potential vendor to understand your needs and your ethical requirements for using AI.

The MERL Tech Initiative is developing a list of more extensive frameworks to further support your selection process. Visit https://merltech.org for updates.

What this assessment tool is *not*

We have not developed this tool serve as:

- A guide for assessing adoption or use of 'all-purpose' GenAI chatbots and tools like ChatGPT, Claude, Copilot, Perplexity, Deep Seek, etc.
- A set of technical implementation details and specific capabilities that would need to be incorporated (these will be specific to the terms of services).
- A tool adapted to every service provider type or audience.
- A checklist for which every criteria must be 'ticked' in order for procurement with the vendor to proceed.
- A checklist for you/your team to prepare for internal training needs. Before entering into a procurement process, we do encourage teams to reflect upon what kind of training is needed as well as how many users will be brought into the vendor relationship and/or users of an AI tool.
- A guide for identifying and addressing bias in models and outputs. While you may be able to influence a vendor, it is unlikely you will have the ability to influence a foundational model's construction. If you are a developer, MTI's NLP-CoP is exploring areas such as ethical data annotation, environmental impact of AI, bias in AI models, AI Governance, and AI and children's data.
- A tool for understanding structural issues baked into how AI is built and sustained, and who profits from this. As a starting point, please consider Tony Roberts' Ten reasons not to use AI for development and ten routes to more responsible use.

Using This Assessment Tool Effectively

This assessment tool is a starting point for conversation, not a definitive checklist. It contains technical terminology that may be unfamiliar. We've tried to balance technical precision with accessibility. When vendors use terms you don't understand, ask them to explain in non-technical language. Reputable vendors will be happy to translate technical concepts. A few key terms:

- **LLM** (Large Language Model): AI systems like GPT-4 or Claude that generate human-like text.
- **API** (Application Programming Interface): How different software systems communicate.
- **PII** (Personally Identifiable Information): Data that could identify specific individuals.
- **Explainability**: The ability to understand and explain how an AI system makes decisions.

Making Judgements

Many of the criteria in this assessment tool require judgment calls. When uncertain about how to assess a response, ask for examples, request documentation, speak with current customers, and consult with technical advisors (or a search engine!)

> By working through this document while assessing a potential AI Vendor, we hope that you'll be able to identify and then request certain standards and good practices from the Vendor and to raise any red flags or concerns that need to be resolved before entering into a contract.
>
> The assessment tool aims to highlight important terms for you to listen out for and also learn about. There will be terms and processes that are unfamiliar to you. The Core Resources list and footnoted sources offer further material to enrich your learning.

Preparing to enter into a conversation with a potential AI vendor

Be explicit about your capacity constraints when engaging vendors and prioritize those who demonstrate an understanding of your organizational context. Some areas to consider before conversing with a vendor include:

Budget

In today's funding landscape, particularly following drastic reductions in aid budgets and changing donor policies, organisations are facing increased financial constraints. When facing severe budget constraints, organizations may be tempted toward suboptimal approaches. Even with limited budgets, maintaining transparency and governance around AI adoption is essential for managing risks. We recommend:

1) Establishing a clear budget ceiling before approaching vendors.
2) Prioritizing flexible pricing models that allow for piloting before full implementation.
3) Considering total cost of ownership, including training, maintenance, and potential exit costs.
4) Examining open-source or locally deployable options that may have lower long-term costs than subscription services.
5) Creating clear policies for staff who are using free consumer AI tools for organizational purposes.

Building Internal Consensus

Below is a set of questions for discussion amongst your team. Document your reflections to guide your vendor selection process and create clear parameters for acceptable AI implementations.

1) How does AI adoption align with your organization's mission and values? What are your non-negotiables?
2) How might AI adoption shift power dynamics with the communities you serve?

3) What internal capacity do you need to build to responsibly oversee this technology?
4) What specific use cases or applications would your organization consider off-limits?
5) What risks would be unacceptable in your specific implementation?
6) What AI tool usage already exists across the organization?

Learning from failed AI projects

Our sector has experienced numerous AI project failures. Transparent vendors will openly discuss past challenges and how they've adapted their approach. When assessing vendors, ask about their failures and what they've learned from them. Common patterns include:

1) Many projects fail because the complexity of data preparation, integration, and maintenance was severely underestimated
2) Vendors often oversell AI capabilities, leading to systems that cannot perform as promised
3) Many projects successfully pilot but fail to transition to sustainable long-term operations
4) Systems designed for high-resource environments often fail in non-profit contexts
5) Initial resource estimates rarely account for the full lifecycle costs, leading to abandoned projects when financial and human resources run out

Consider issues options for 'data sovereignty' at the outset

It is worth noting the growing movement around "data sovereignty', you can research:

1) Options for data storage in specific geographic regions
2) Compliance with local data laws beyond just GDPR & understanding of regional regulations beyond US/EU frameworks
3) Clear policies on cross-border data transfers
4) Flexibility on data hosting location requirements
5) Options for local deployment without data leaving your infrastructure

Assessment Tool for Potential Vendors

RULES OF THUMB

1. The best vendors will be transparent about both the capabilities and limitations of their specific product *and* well-documented structural issues embedded in GenAI more broadly and will have clear, documented processes for all critical aspects of their service.

2. While responses are presented in binary terms, conversations will most often sit somewhere on a spectrum. What matters most is a vendor's willingness to engage and discuss clearly and transparently.

3. Not all the questions in this tool will be relevant to your priorities. First, determine what your priority questions are, why, and agree with colleagues why some questions are *not* a priority. A helpful starting point is to consider high and low risks for your particular organisational context.

Provider Integrity

Vendor Stability, Experimentation, and Exit

Can you provide references for current/previous clients in our sector?

✅ Good Response:

- □ Multiple relevant references available
- □ References attest to how well the tool delivered on its promise, available features, and vendor responsiveness
- □ Sensitivity to needs of users of the tool
- □ Case studies with measurable outcomes
- □ Long-term client relationships
- □ Industry-specific expertise demonstrated

⚠️ Concerning:

- □ No relevant references
- □ Marketed features under development
- □ Only pilot projects
- □ High client turnover
- □ Poor user experience
- □ Limited industry experience

What opportunities do you provide to test your tool, product, or service before full deployment?

✅ Good Response:

- □ Offers a trial period
- □ Provides a sandbox environment for team to test with sample data
- □ Flexible contract terms for testing before full financial commitment
- □ Provides support during trial
- □ Has established processes for incorporating user feedback into product improvements

⚠️ Concerning:

- □ No trial period
- □ No sandbox environment
- □ Requires significant upfront investment before proving value
- □ Dismisses the need for trial/testing
- □ Vague about support resources during trial
- □ No processes for incorporating feedback for product improvement

What is your pricing structure and how do you prevent unexpected costs?	
✅ Good Response: □ Clear, predictable pricing model (fixed, tiered, or usage-based with caps) □ Transparent about all costs, including implementation, training, and maintenance □ No hidden fees for standard features □ Ability to set spending limits or caps □ Cost projection provided	⚠️ Concerning: □ Vague or complicated pricing structure □ Usage-based pricing without caps □ High costs for basic features or functionality □ No hidden fees for standard features □ History of unexpected charges with other clients
What is the process for transitioning to another provider?	
✅ Good Response: □ Documented data export procedures □ Standard data formats □ Transition assistance included in contract □ No data hostage situations □ Clear timeline and process	⚠️ Concerning: □ Proprietary data formats □ Export fees □ No transition support □ Long lock-in periods
Explainability & Transparency	
What level of model explainability can you provide?	
✅ Good Response: □ Feature importance rankings □ Clear confidence scores □ Decision path visualization tools □ Detailed logging of model inputs/outputs □ Provision of explainability reports[4]	⚠️ Concerning: □ "The model is too complex to explain" □ Black box approaches without any visibility □ No monitoring of decision patterns □ No explanation of where and how AI reasoning & judgement occurs

[4] European Data Protection Supervisor (2023) [TechDispatch: Explainable AI.]

Which commercial LLM provider(s) do you use and how?[5]	
☑ Good Response: □ Clear disclosure of LLM providers (e.g., OpenAI, Anthropic, etc.) □ Specific model versions used □ Detailed architecture showing where LLM sits in the processing pipeline (this is dependent on solution type, off-the-shelf or custom build □ Pros/cons, knowns/unknowns regarding data privacy, changing political contexts & unstable terms and service agreements □ Version-controlled prompt library □ Regular prompt testing and optimization □ Security review process for prompts □ Monitoring of prompt effectiveness	⚠ Concerning: □ Unwillingness to disclose LLM provider □ No version control for LLM integration □ Lack clear conveyance of LLM in processing architecture □ "The company's terms and conditions say the data will be secure"; no mention of changing political contexts □ Ad-hoc prompt creation □ No prompt version control □ No security review of prompts □ No monitoring of prompt performance
What guardrails would you advise we build together around LLM output?	
☑ Good Response: □ Clearly explain the process, options, and any current guardrails in their offering □ Willingness to learn and adapt and open to considerations they may not have thought of before □ Content filtering systems in place □ Ringfence LLM use for specific functions (e.g. opt-out features □ Output validation against business rules[6] □ Human monitoring for hallucinations or incorrect responses □ Clear processes for handling LLM errors □ Regular testing of output quality	⚠ Concerning: □ Unable to discuss options for implementing guardrails and what is possible in the current offering □ Unwilling to learn or consider particular needs & concerns of the development sector □ Raw LLM output without validation □ No monitoring of response quality □ No system for detecting hallucinations □ LLMs 'black box' integration does not distinguish between functions for opt-in or customization

[5] This assessment tool~~question rubric~~ assumes a vendor is using commercial LLMs. There are a plethora of open source, and small language model options emerging for GenAI. We believe this a promising alternative to the data privacy security issues facing AI in Big Tech. Not all open source models are created equal, however. Some are known to have security vulnerabilities and fewer guardrails. The NLP-CoP intends to further this discussion in the future. In the meantime, this paper is a starting point. For an overview of small language models see here.

[6] A validation rule ensures value entered is legitimate for the context of its field (e.g age value = 5, valid vs. age value = -5, invalid). A business rule ensures values which passed validation adhere to policies and procedures of the business.

Performance Monitoring

How do you measure and maintain response quality?

✅ Good Response:

- ☐ Regular **human-centered review process** for sample outputs, clearly documented
- ☐ Clear quality thresholds and alerting system
- ☐ Source verification methods so that AI outputs can be traced back to specific source material informing AI judgments
- ☐ Quality scoring system with clear criteria
- ☐ Regular stakeholder reviews
- ☐ Root cause analysis for quality issues
- ☐ Details an improvement cycle that extends beyond the initial benchmark setting

⚠️ Concerning:

- ☐ **Missing human oversight**
- ☐ No defined quality metrics
- ☐ No mention of source verification for AI outputs
- ☐ Manual or ad-hoc quality checks
- ☐ No systematic improvement process
- ☐ Unclear quality standards
- ☐ Poor feedback integration

What is your approach to error detection and handling?

✅ Good Response:

- ☐ Mechanisms for humans to review AI decisions before they are finalized
- ☐ Explains how users can override or correct AI outputs when needed
- ☐ Clear remediation procedures by error type
- ☐ Escalation procedures
- ☐ Regular error pattern analysis
- ☐ Proactive mitigation strategies

⚠️ Concerning:

- ☐ No clear process for human override of AI decisions
- ☐ No error logging or error classification
- ☐ Missing remediation procedures
- ☐ No escalation procedures
- ☐ No error pattern analysis
- ☐ Reactive-only mitigation approach

What training do you provide?
How do you accommodate different levels of technical literacy in your training and support offerings?

✅ Good Response:

- ☐ Willingness to speak with client, clarify and document on an ongoing basis
- ☐ Different learning formats (videos, documentation, live sessions)
- ☐ Support staff trained to communicate effectively with non-technical users
- ☐ Examples of flexible support solutions adapted to client needs

⚠️ Concerning:

- ☐ Expect clients to have dedicated technical staff as intermediaries
- ☐ One-size-fits-all training approach
- ☐ Technical documentation only available
- ☐ Limited support mechanisms
- ☐ No examples of successfully supporting non-technical users

Responsible Data, Security & Privacy[7]

Responsible Data

How do you ensure responsible data handling across the entire LLM pipeline?

☑ Good Response:

- □ API-submitted data has opt out for training
- □ Data flow diagrams
- □ Clear data retention policies at each stage
- □ Regular audits of entire pipeline
- □ Documented data minimization practices
- □ Privacy impact assessments

⚠ Concerning:

- □ Unable to prevent data from being used for training purposes
- □ No end-to-end visibility of data flow
- □ Unclear data handling procedures
- □ No regular audits
- □ No data minimization strategy

Privacy

How do you handle data privacy?
Note: this is especially important to consider when using commercial LLMs

☑ Good Response:

- □ Clear documentation of data flow to/from LLM
- □ Preprocessing steps to remove PII/sensitive data, clearly documented
- □ Use of data privacy features (e.g., opt-out & no-storage options)
- □ Regular audits of data sent to LLM
- □ Clear understanding of LLM data retention policies

⚠ Concerning:

- □ No PII data filtering before LLM processing
- □ Unclear about LLM provider's data usage rights and retention policies
- □ No monitoring of data sent to LLM
- □ Sending raw customer/'beneficiary' data without controls

[7] UN Global Pulse. (2020). Privacy assessment tool.

Security	
What specific security certifications do you maintain and regulatory compliance do you follow?[8]	
✅ Good Response: □ Clear documentation of compliance with GDPR, EU AI Act or other equivalent regulatory measures □ Provision of ISO standards(ask about ISO27001) □ Encryption at rest and in transit □ Clear data handling procedures □ Automated compliance checks □ Contact details of the Data Protection Officer (DPO)	⚠️ Concerning: □ Minimal compliance or unclear security and data handling procedures □ No reference to GDPR or EU AI Act or equivalent regulatory measures □ "We're working on getting certified" or expired certifications □ No reference to industry-wide regulatory measures □ Basic encryption only □ No access controls □ No DPO role
What is your environmental impact assessment and mitigation strategy?[9]	
✅ Good Response: □ Water resource usage, mineral resource consumption, carbon footprint analysis □ Identifies features in model that increase energy consumption □ Describes design choices to reduce consumption and trade-offs □ Monitors carbon footprint	⚠️ Concerning: □ Minimal consideration of environmental impact □ No design choice considerations

[8] For teams with limited technical expertise, at minimum inquire into compliance with GDPR and/or EU AI Act or comparable legislation. If ISO certification is provided, research what it means.

[9] While this question may not be a priority for all, we believe climate conscious organisations and funders should consider investing in small, green, sustainable, local-first AI. For more see: Raftree L., (2025) Evidence and Learning in the Context of Climate Change: Invitation to Action.

Core Resources

- UNESCO (2023) Ethical Impact Assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence
- Future of Life Institute (2024). High-level summary of the AI Act. EU Artificial Intelligence Act
- World Economic Forum. (2023). Adopting AI responsibly: Guidelines for procurement of AI solutions by the private sector
- BSR (2025) Human Rights Across the Generative AI Value Chain: Human Rights Assessment of the Generative AI Value Chain and Responsible AI Practitioner Guides
- National Institute of Standards and Technology (2024). AI Risk Management Framework.
- 18F (2020) De-Risking Government Technology Federal Agency Field Guide
- IEEE Standards Association. Autonomous and intelligent systems (AIS) standards. IEEE. Retrieved 2025.