MITRE

MTR210601
MITRE TECHNICAL REPORT

The AI Relevance Competence Cost Score (ARCCS) Framework

McLean, VA

Author(s):
Mike Hadjimichael, Ph.D.
Anne Townsend
J. Cory Miniter
Andrew Hand

November 2021

Table of Contents

List of Figures

1 Introduction

Artificial Intelligence (AI) has become increasingly prevalent in the public imagination, and as a result, a variety of commercial enterprises have begun marketing products as "AI-enabled." However, we see that "AI" has become a marketing buzzword, and not every product claiming to be AI-enabled actually uses AI in a meaningful way [1]. Since AI is such a broad area of study and because vendors often hesitate to share details on underlying algorithms, it is often difficult to determine the effectiveness of these implementations.

This document proposes the AI Relevance Competence Cost Score (ARCCS) framework, an evaluation methodology and metrics to assess the degree and effectiveness of the AI component of a commercially offered, AI-enabled tool. The framework guides the assessor to organize the available evidence, evaluate its strength, and determine whether a product performs as advertised in a technically relevant manner. Additionally, ARCCS serves as a guide for further investigation once such a determination is made.

ARCCS provides a method for any organization considering the acquisition of AI-enabled tools to make more informed, rigorous, and consistent assessments and final decisions.

2 Methodology

The ARCCS assessment is divided into 3 major dimensions and two modifiers:

- **Relevance** –Is the AI component is necessary and appropriate?
- **Competence** – How well does it do what it says it does?
- **Cost** – Consider the risks and benefits of the product, including organizational and material.
- **Confidence** – This modifier of the 3-dimension score is a measure of the overall quality of information available, providing a metric for determining how confident an assessor can be that the overall assessment is correct.
- **Strength** – This reflects the amount of knowledge available, as indicated by the percentage of answered questions in the Relevance, Competence, and Cost sections.

Each ARCCS dimension and the modifier is composed of a set of features, expressed as questions, and split into logical categories. Individual feature assessments are conducted on an ordinal scale of 1 to 5, with 5 being the optimal value. Some of the assessment metrics are binary questions (e.g., "Does the system allow users to provide feedback on results?"). In binary cases, a negative response will be scored a 1, and a positive result will be scored a 5. Guidelines for responses are outlined in Section 5, as well as in the supplemental "Assessment Guidelines" spreadsheet. These guidelines are intended to guide scoring rather than being rigidly prescriptive; the ARCCS assessment features are often subjective and may rely on a subject matter expert's judgement to assign a value.

The general ARCCS assessment is designed to apply across a broad range of categories, but there is room for more focused versions of this assessment. To that end, separate ARCCS "profiles" can be defined. Profiles are more specialized versions of the assessments which are geared toward a particular domain and provide greater guidance to remove some subjectivity in the answering process. In this paper, we provide a "cyber profile," along with examples, geared toward a cybersecurity application of ARCCS.

3 Applying the Assessment

3.1 Assumptions & Considerations

Assessors using the ARCCS framework should begin by outlining their particular use cases and any assumptions they may have before beginning to apply the assessment. Particularly, it is important to understand the basic product under assessment. How the assessment is applied to various products may yield different scores depending on the perspective of the assessor. Consider a hypothetical vendor that offers a holistic platform with wide ranging cyber defensive capabilities, a component of which claims to use AI to identify threats. If assessors decide to assess the platform as a whole, they will likely end up scoring the "centrality" metric lower than a team that decided to assess the individual component on its own. Our team does not make a judgement on the validity of either of these approaches, rather, our recommendation is that the assessor determines the approach most suited to their particular use case and need and applies the framework consistently along those lines.

In cases where the answer to one of the assessment questions is unknown, it can be scored N/A. This will prevent that metric from being included in the dimension score but will result in a lower confidence score. A team undertaking an ARCCS assessment should understand their own tolerance to risk and low-confidence assessments.

3.2 Assessment Template

The first version of ARCCS is distributed with a Cybersecurity Profile as a Microsoft Excel spreadsheet intended for assessors to score features on a 1-5 scale, with 1 being the lowest (poor) score and 5 being the highest (good) score. Each field additionally has a "Notes" section so that assessors can provide additional context to their reasoning, such as links to relevant documentation/papers/patents or the assessors' thought process.

3.3 Sample Questions for a Vendor

It is recommended that assessors engage with product vendors where possible. Exploratory applications of this assessment by the MITRE team based on publicly available marketing materials frequently encountered situations where some of these questions were unanswerable (for example, many vendors made no mention of underlying algorithms). In this section, we include common questions our test assessors had during the process:

1. Do you have any data indicating how effective your system is at detecting threats, images, etc.?
2. How does your system compare to your competitors?
3. Can you go into detail about the underlying models that drive your system? What technologies enable your system to learn (if applicable)? What techniques allow it to make autonomous decisions (if applicable)?
4. Is your system retrainable?
5. Can you outline the process for the end user to retrain the system?
6. Are/How are results provided to the end user?
7. Is the end user able to influence system performance by providing feedback (e.g., by identifying false positives)?

8. Does your system keep track of its historical performance once deployed? If there is a sudden degradation in performance, how will the system and operators "know"?
9. Does your system take any steps to mitigate adversarial techniques? How do you prevent poisoning or evasion attacks?
10. Beyond your publicly accessible whitepapers, do you have any technical information regarding the AI component's design that you'd be able to share with our team? Any academic publications, architecture diagrams, etc., that you would like to put in front of us?

4 Example Evaluations

Section 5 will discuss the metrics comprising the ARCCS evaluation framework. To help contextualize select metrics, three hypothetical ARCCS evaluation cases will serve as examples, using the Cybersecurity profile. Examples will be referred to as Case Study One (CS1), Case Study Two (CS2), and Case Study Three (CS3).

4.1 Case Study One (Anti-malware)

For CS1, the assessment team is in search of an anti-malware solution capable of defending against zero-day attacks while preferably matching or out-performing their company's current anti-virus software for known threats. In their search, they come across several product vendors claiming to use AI to defend against malware.

CS1 is one such cybersecurity product. This capability is deployed as an endpoint agent in communication with a vendor-homed cloud. The endpoint component primarily collects and forwards potentially malicious samples, while the cloud-based capability takes these samples and analyzes them using Deep Neural Network (DNN) classifiers to determine whether they are malicious. The product vendor is widely known and has a track record of providing competent security solutions.

4.2 Case Study Two (Network monitoring)

In CS2, the assessment team is looking for network logging solutions with built-in analytic capabilities that can flag suspicious traffic for network operators to follow up on. They currently have no on-site capability to compare against but have set a notional selection threshold of 10,000 events/second for throughput.

CS2 is a logging capability used to monitor network traffic within an enterprise. Among its features is an AI capability that claims to use AI to detect malicious traffic among network data. This capability is intended to be deployed and configured by the prospective user, and the assessment team has made the decision to evaluate the product holistically as opposed to only assessing the component containing purported AI (Refer to Section 3.1 for discussion on this decision). The product vendor is relatively new to this market but can point to ten other large companies that have adopted its tool. This platform does not require expansion of existing network infrastructure.

4.3 Case Study Three (Web-based phishing prevention)

In CS3, the team is evaluating tools to prevent phishing attacks against people using their corporate email. During their evaluations, they come across a product claiming to use AI to prevent phishing.

CS3 is a browser security solution developed by a new startup company intended to protect users from phishing attacks. The core idea behind the product is that it uses machine learning based image processing techniques to determine the legitimacy of websites by checking whether web pages visually match the images the vendor has on file. The product is installed as a web-browser plugin which claims to block malicious sites.

5 ARCCS Metrics

This section will describe the metrics of ARCCS, as well as suggested guidelines for scoring.

5.1 Relevance (how necessary and appropriate is the AI component)

5.1.1 Goodness of fit

In the ARCCS framework, goodness of fit determines whether an AI application is appropriate to the problem being solved. It is not good enough to simply have an AI function; any such AI must contribute meaningfully toward the overall performance of the system. Some questions an assessor should ask:

How appropriate is the use of AI in this context?

Cyber Profile:
Do the technologies listed in the documentation fit with the capabilities claimed by the vendor? If the system is attempting to sort events into categories (e.g., malicious, benign files for an anti-virus), the underlying technology must have some sort of classification mechanism (decision trees, random forests, expert systems, etc.). If the system is attempting to identify deviations from normal patterns, for example, in network data flows, anomaly detection approaches may be appropriate. These may include supervised approaches like **Support Vector Machines (SVM)** or unsupervised methods like **k-nearest neighbor (KNN)**. It is also possible that some of these approaches may be used in tandem, where a system may use a first pass approach to flag an anomalous sample, and then use (for example) **deep neural networks** (**DNNs**) to identify the type of anomaly or threat. Conversely, if a piece of antivirus software claims to use AI but the only listed technology is Latent Dirichlet Allocation (LDA), it may warrant some skepticism as that technique is typically used for text document classification.
Please note: the technologies outlined in this section are included to illustrate some applications of AI technologies in the cyber domain and are not exhaustive. Section 5.2.3 includes an additional list of keywords that one may encounter while evaluating an AI product.

For example:

- Using classification for natural classification problem, regression for regression problem (y/n)?
- Does data that the system uses line up with the data available? Matching data types, data directly related to output? (y/n).
- Is this a problem typically addressed by AI in other products within the same domain? For example, using anomaly detection approaches to identify suspicious patterns in network data is common in AI enabled cybersecurity platforms.

1	Approach is counter to best practice. Academic research/other publications/evaluator experience indicates that approach has serious flaws or is arbitrary.
2	Approach is unproven, hypothetical. No supporting documentation can be found but product can be demonstrated to do what it claims.
3	Approach has some 3rd party evidence supporting claims of functionality.
4	Approach is widely accepted and has been frequently applied in similar or related contexts.
5	Approach is widely accepted and has been frequently applied in the specific use case.

CS1: The AI-enabled antivirus tool claims to take a deep learning approach to classify new files as benign or malicious based on behavior. Rather than relying on previously encountered hashes of malicious files, the system uses a complex ensemble of neural networks to classify samples based on information collected when it is detonated in the tool's sandbox, including separate models for data collection, process information, and an image classifier for any GUI features that are opened. All these information streams work in concert to determine whether a file is classified as malicious. This is an appropriate application and would be scored a 5.

CS2: The AI-enabled component takes an unsupervised anomaly-detection approach to data entering the system to flag data deviating from the baseline behavior as suspicious. This is a common approach to this problem and is a reasonable use of AI. Like CS1, this would be scored as a 5.

CS3: The assessment team can only find vague allusions to machine learning techniques being applied. Upon review of the product, the assessors discover that while there is an image processing step, the final decision process for determining a malicious site is conducted by a human analyst, assisted by a confidence score generated by the ML component. These results are propagated to the client as a domain blacklist. For this reason and since this approach requires the agents to build a comprehensive list of domains to be effective, the assessment team scores this a 1.

5.1.2 Centrality of AI component

AI centrality seeks to uncover the degree to which the overall system under assessment relies on its AI algorithm(s) to accomplish its task. Note: When discussing centrality on large, multi-role platforms, it is important to discuss whether assessing centrality from the perspective of the overall system, or from individual components is more worthwhile to your understanding. Some example questions assessors may ask:

- Is the AI component central to the solution such that removing it removes core functionality?
- Is AI component foundational?
- Is it sold as an AI tool, or tool with AI support?
- Would system lose all functionality without the AI component?

Guidelines for grading centrality:

1	AI is secondary to operation of the system; tacked on and not directly relevant to the rest of the system or the problem the system is attempting to solve.
2	
3	ML/AI component is an important feature of the overall system but is not required for successful system operation (e.g., a module) but in direct support of the mission
4	
5	ML/AI component is a keystone feature of the system; System has been designed around this component.

CS1: The AI component of this system forms the bulk of the functionality, and there are few features beside it. The assessors scored this as a 5.

CS2: Since the system is being evaluated as a whole and the AI component is an optional module, the team scored centrality as a 1.

CS3: The AI component could hypothetically assist analysts but is not a vital part of the solution. The team scores this a 3.

5.1.3 Proportion of overall functionality

This metric seeks to assess the proportion of AI and non-AI functionality of a system. Some example questions an assessor may ask:

- How much of total work/functionality depends on the AI component?

- Proportion of AI components to non-AI components.

Guidelines for grading:

1	Component provides less than 10% of the system's overall functionality
2	Component is a small proportion of overall system functionality (<25%)
3	AI component is a significant portion of overall system functionality (<50%)
4	AI component is a large portion of overall system functionality, but there are additional features that can function in its absence (>50%)
5	AI component encompasses most of the functionality of the system (excluding infrastructure e.g., data storage and transport). The system provides no functionality in its absence.

CS1: As alluded to in the previous metric, the AI component forms the bulk of overall system functionality. As such, the proportion metric is also scored a 5.

CS2: The AI component of this product is a minor portion of the overall functionality. Since the team made the decision to assess the platform wholistically (see section 4.2), proportion is scored a 2.

CS3: Between the client blacklist and human evaluators, the AI component comprises a very small portion of overall functionality. The team scores this a 2.

5.1.4 Necessary vs Gratuitous AI

Assessing the gratuity of AI in the systems allows an assessor to understand whether the AI component of a system is accomplishing a vital task, as well as whether that task could be accomplished more efficiently through non-AI means. An example question:

- Can the AI component be replaced with an equally functional non-AI-component?

Guidelines for grading:

1	AI is unnecessary, functionality could be equivalently accomplished with non-AI component.
2	AI provides a slight advantage (some minor improvement) over non-AI solution.
3	AI provides a notable improvement over a non-AI solution (one of accuracy, speed, etc.).
4	AI Provides a significant improvement (more than one of accuracy, speed, etc.).
5	AI is irreplaceable. Task could not realistically be accomplished without AI.

Cyber Profile
A system for monitoring network traffic that uses anomaly detection to identify threats that a human or traditional signature-based system might not catch (Necessary), versus software that monitors network traffic and has an NLP based "assistant" that can query the system the same way a human at a keyboard could (Gratuitous). The latter system would be scored a 1, while the former would likely be scored 3 or higher, depending on claimed advantages.

CS1: Using AI to detect malware is not the only possible approach to this problem, but it is by no means gratuitous. Compared to other, non-AI approaches, it may allow the system to spot malicious behavior that has never been encountered before. The team scored this a 5.

CS2: The AI component of this system allows for malicious dataflows to be detected on a large scale. The assessment team's existing non-AI-enabled tools allow them to accomplish similar detections without AI, but not nearly as effectively. The team scores this a 4.

CS3: The AI component claims to use AI to compare images of website landing pages, but there are non-AI techniques that could be used to accomplish the same goal. The product documentation is very vague about what techniques are being applied, and the product sales team does not make a convincing case for why AI vs these other algorithms. The assessment team scores this a 1.

5.2 Competence (how well does it do what it claims)

5.2.1 Needs Alignment

It is important to understand how the performance of the system compares to, or aligns with, the user's needs. This knowledge can inform further decision making and help in understanding whether an AI component is a useful addition to a system.

1	Performance of the system does not satisfy project requirements or does not function as advertised.
2	Performance of the system is slightly worse than project requirements.
3	Performance of the system is on-par with the project requirements.
4	Performance is a significantly better than expectations and needs.
5	Performance represents a breakthrough advancement in capability, well beyond the stated needs of the project.

Cyber Profile

The system under assessment claims to be effective against malware samples. In this case, it is useful to look closely at the accuracy of the system, especially error rates. Assessors must determine their acceptable accuracy thresholds and tolerance to risk and assess along those lines. AI approaches that significantly outperform in-place solutions without generating excessive false positives would likely be scored highly on this scale, though "excessive" comes down to the assessor's tolerance.

CS1: The assessment team requires an anti-malware solution that will detect all malicious test samples without excessive false positives (<10%). Upon review of vendor statistics and benchmarks, performance of the system yields more frequent, accurate detections than their project requirements. Under a controlled test, the system displays no false negatives and only 3% false positives. The team scores this a 4.

CS2: The team requires a logging and analytic solution that can process maintain a throughput of 10k events per second. Under test, the system can handle slightly over that amount while still generating useful alerts. Performance of the system is found to be in line project requirements. The assessors score this metric a 3.

CS3: During testing, the team notes a false-negative rate of 12% for the malicious websites under test. This is slightly worse than the project imposed 10% requirement, so the team scores this a 2.

5.2.2 Errors

5.2.2.1 Detect model drift

This is a simple binary metric which scores whether the system can detect model drift. Normal behavior changes over time and requires the system model to follow that "drift" instead of generating an increasing number of false positives and false negatives. A system which monitors its performance, or continuously retrains on new normal data, will better manage model drift.

1	The system has no ability to detect drift.
2	
3	
4	
5	The system has the ability to detect drift.

CS1: The assessment team was unable to determine whether this is possible with the information available. This is scored an N/A.

CS2: The system can detect and account for model drift, so it is scored a 5.

CS3: The team is unable to determine whether this is possible with the information available. The team scores this an N/A.

5.2.2.2 Retrainability

In an AI-enabled system which uses a model developed using ML, the model is developed by a training process on a set of input data. This data may become invalid over time or may not be appropriate to the desired use. The "retrainability" metric tracks how easy it is for a user to update the underlying ML models, if appropriate. If the system does not have a machine learning component but has another form of AI that functions without retraining, this metric should score 5.

Example question:

- Can the machine learning component be retrained for more relevant or improved performance?

Guidelines:

1	Not retrainable.
2	System retraining can only be done by the product vendor and has an excessive turnaround time or cost.
3	Retrainable by vendor along with consistent updates regularly provided by vendor, e.g., through automatic model pushes.
4	Retrainable by system operators on their own schedule, but may be difficult to do (e.g., arcane process, onerous downtimes).
5	The system requires training that can be easily undertaken by the operators through a simple process or is automatic or the system functions without the need for training.

Note: Depending on the specific use case, a team may decide that automatic updates from a vendor are preferable to a difficult training process done in-house. In that case, the score may be increased from a 3.

CS1: The system requires occasional retraining but is under tight control by the vendor. These are provided as routine software updates. The team scores this a 3.

CS2: The machine learning algorithm is easily retrainable with several “wizard” style helpers that walk the user through the process. The team scores this a 5.

CS3: The assessment team is unable to determine whether the system can be retrained. This is scored an N/A.

5.2.3 Technological Maturity

The Technological Maturity metric helps guide ARCCS users to understand the maturity of the underlying AI algorithms in a system. Note that while novel approaches will be scored low on this metric, that is not necessarily an overall judgement on the system. Novel approaches may prove to be effective, but adopting an unproven approach represents a potential risk.

Cyber Profile
Various anomaly detection approaches operating on network data. In such approaches, the network defender builds a baseline "normal" network activity profile, and then checks for traffic that varies significantly from that baseline. This is a commonly applied approach and providing the product under assessment has been reasonably well-developed, would be scored highly on this scale.

Look for the following Keywords (not an exhaustive list):

Deep learning	Generative Adversarial Network (GAN)
Anomaly detection	Adversarial Machine Learning
Clustering algorithms	Generative model
Latent variable models	Adversarial network
Expert system	Multi-task learning Neural Network
Semantic web	Machine Learning
Fuzzy logic	(Un)supervised Learning
Bayesian optimization	Deep Learning
Evolutionary algorithm	Ensemble Models
Genetic algorithm	K-means
Gradient Descent	Gradient Boosted Trees
Active learning	Decision Trees/Forests
Feature extraction	Q-learning
Adaptive learning	

Is the AI technology well-proven (as opposed to recently developed) when applied to the intended use case?

1	Technology is based on recent research.
2	AI technology has been in development for two or more years, preferably with proven application.
3	AI technology is well known and has been actively developed and researched for 5+ years.
4	AI technology is mature (7-9 years).
5	AI technology is mature and widely accepted as a standard, in development or practice for over a decade.

CS1: The system uses a complex mixture of novel algorithms and well-developed AI

technologies. Because of the presence of the novel pieces, the team decides to drop the score from a 4 to a 3, citing the unproven approaches.

CS2: The documentation of specific algorithms within the system is sparse, but the approaches that are referenced are well established and have been areas of active research dating back at least 8 years. The team scores this a tentative 4.

CS3: There is only a high-level description of the AI component, with no greater detail than "image processing." The team scores this an N/A.

5.2.4 Effectiveness

5.2.4.1 Used by other organizations?

Understanding how well-adopted a system is can help boost confidence in an assessment. Widely adopted systems will have third parties who are able to answer questions, while information may be limited with new, recently introduced, products. Thresholds may need adjustment according to the domain of application.

Used by other organizations, either commercial, government, or other?

1	No other organizations are reported as using the system.
2	System has been adopted by a few organizations.
3	System enjoys some industry adoption, at least one of which is using it for a similar use case to the evaluators.
4	A wide array of organizations currently uses the product for a diverse set of use cases, including the evaluators' use case.
5	System is considered an industry leader and is adopted by many organizations.

Cyber Profile
A few examples of broadly adopted technologies with AI components (scored 5 on this scale): Windows Defender, Splunk, Elasticsearch. Cybersecurity is an area important to any organization, and therefore will have higher thresholds for this metric. Suggested Cyber Thresholds:

1 = No adopters
2 = < 10 organizations
3 = 10 - 50 organizations
4 = 50 -100 organizations
5 = > 100 organizations

CS1: Upon review, the assessment team finds an extensive list of other organizations utilizing this AI tool. In total, there are over 500 organizations using the product, so the assessment team grades this metric a 5.

CS2: The product vendor provides a list of 15 other organizations using their product. They claim a higher number, but this cannot be independently verified. The assessment team scores this metric a 3.

CS3: The product vendor lists 12 different originations that are using their product. The assessment team scores this a 2.

5.2.4.2 Provides transparency and explainability

The transparency metric tracks how much of the system operation is revealed to a user. Explainability reports the reasoning behind system output. To score well in the explainability metric, the system should highlight why a decision was made in one way or another, in a way that is understandable and useful for humans. Understanding this level of information can help better understand how the underlying algorithms are being applied, as well as how effective the system is.

Does system output provide:

- Transparency into system operation?
- Explanations which help the user understand the system's reasoning and conclusion.

1	AI component is a complete black box with no transparency into system operation, and no algorithmic or domain explainability.
2	AI component is partially transparent
3	AI component provides either transparency into system operation or explainability, but not both.
4	System is transparent and at least partially explainable, or vice versa.
5	AI component is both transparent and explainable

CS1: The system is highly controlled by the product vendor and allows virtually no transparency into operation at the user level beyond a results screen. The assessment team scores this metric a 1.

CS2: The system provides statistics about ML jobs and allows users to drill down into records and provides information on what triggered the anomaly detector. It doesn't provide any information on the underlying algorithms though, so the assessment team scores this a 3.

CS3: The AI component of the system is completely controlled by the product vendor, with no explanation of how a rating is determined when a site is blocked. The team scores this a 1.

5.2.4.3 Historical tracking of results/performance

Does the system provide the ability to track results for retraining purposes?

1	System does not track historical results and performance
2	
3	System does not track historical results, but does not require retraining
4	
5	System tracks historical results and performance in a sufficient manner to facilitate retraining

CS1: The assessment team is unable to turn up any information on tracking of historical records. They initially rank this N/A due to lack of information. An interview with the product vendor reveals that they keep samples of malicious files, but they do not reveal any more information regarding record keeping. The team adjusts the score to a 1.

CS2: As a logging solution, this product provides robust, detailed historical tracking. The team scores this a 5.

CS3: The vendor does not indicate any capability for tracking historical records. The team scores this a 1.

5.2.4.4 User feedback mechanism

This is a simple binary metric that serves to track whether, during operation, a system allows its users (system operators, cybersecurity personnel, etc.) to influence AI behavior, e.g., through identification of false positives.

A question the assessor should ask:

Does the system provide the ability for users to override results or provide other feedback on computational results?

1	System has no facility to users to override or provide feedback on accuracy of results.
2	
3	
4	
5	System allows user to override results and provide feedback to system on accuracy of results.

Cyber Profile
Example: A product that automatically blacklists domains based on some learning algorithm could inadvertently flag a known-good domain. Does the system allow a user to overrule it? Do these corrections inform future decisions beyond just the false positive that was flagged? If the system is unable to accept these types of adjustments, it would be scored 1. Similar false positives are possible in other cybersecurity use cases, such as a benign file being flagged as malicious and prevented from running. The system should ideally be able to learn from its own mistakes when humans are forced to intervene.

CS1: The user can override malware detections and force execution of the sample, but the system does not take these manual overrides into account. The team scores this a 1.

CS2: The user can provide feedback to the system, tagging and overriding false positives. The system incorporates this feedback into future results. The team scores this a 5.

CS3: The end user can provide no feedback on the client side for suspected false blocks. The team scores this a 1.

5.3 Cost of AI usage (cost/benefits)

5.3.1 Vulnerabilities (Unaddressed/Unmitigated) introduced

Does the addition of an AI component introduce vulnerabilities to the system? Have the system's developers considered adversarial machine learning techniques, such as those that attempt to generate samples that evade detection or those that attempt to poison models at the training step? Are models trained on publicly available or accessible data? Could the underlying models of the system be easily reproduced or licensed from a different source? Are there any mitigations for adversarial techniques in place?

1	Vulnerabilities are introduced and not mitigated.
2	
3	Vulnerabilities are possibly introduced, possibly mitigated.
4	
5	There are no obvious vulnerabilities, or vulnerabilities positively mitigated.

CS1: This product introduces the potential for poisoning or evasion attacks, but the vendor can demonstrate that they have algorithms in place designed to mitigate these types of attacks. The team scores this a 3.

CS2: This product introduces the potential for poisoning and evasion attacks but does not take any action to mitigate against this possibility. The team scores this a 1.

CS3: The machine learning approach may be vulnerable to poisoning or evasion attacks, but since the final determination is made with a human in the loop, there is also some potential to mitigate this issue given relevant expertise on the analyst side. The team scores this a 3.

5.3.2 Cost of implementation/specialization

Cost of implementation refers to effort or funding required to prepare the system for the organization's use. Does the system come pre-trained on some dataset, or does it rely on user data?

Cyber Profile
Does the system rely on in-house data? How does the system get data for training or testing? Does data collection infrastructure exist, and if so, is the system compatible with existing logging infrastructure (e.g., network sensors)?

1	AI component needs specialization, training, or other work.
2	
3	AI component needs minor configuration costs.
4	
5	AI component is essentially ready out of the box.

CS1: All configuration is handled at the vendor side. Users require no training for use. Score as a 5

CS2: Operators of the system need to understand a vendor-specific query language and require some ML domain knowledge to fully take advantage of the system. The team scores this metric a 3.

CS3: This product is a simple blacklist from the user/client perspective and requires no training to use. This scores as a 5.

5.3.3 Solution Efficiency loss/gain

Solution efficiency assesses whether the AI of the product results in any gains compared to similar systems that do not employ AI. This metric is tracked on a 3-dimension scale, where we look at efficiency, accuracy, and speed. Sum the scores of each category and refer to the scoring chart to determine the final score.

- Efficiency: How effective is the system when considering hardware footprint and human intervention requirements?
- Accuracy: True Positives/Negatives versus False Positives, False Negatives.
- Speed: How quickly can the system process incoming data and generate results?

	Loss	Neutral	Gain
Efficiency	-1	0	1
Accuracy	-1	0	1
Speed	-1	0	1

Sum of Efficiency, Accuracy, Speed	Solution rating
-3	1
-2 or -1	2
0	3
1 or 2	4
3	5

CS1: The tool provides results within minutes of encountering a novel sample, and seconds for known samples. This is slightly slower than other tools, so the team scores the speed score a small loss. However, the system is extremely accurate and provides a service that the AI aids greatly in (detecting novel malware based on its behavior). Compared to an older, hash-based antivirus system, CS1 correctly flagged nearly 50% more test samples. Further, the methods used by the tool suggest a reasonable possibility that the system may be able to catch some zero-day attacks. The team labels both accuracy and efficiency a gain. The sum of the scores is 1, leading to a Solution score of 4.

CS2: The team finds that the AI module provides a significant boost in malicious traffic identification using a smaller hardware footprint than their existing solution. This, combined with a 30% increase in True Positive results compared to their non-AI analytics leads the team to score all 3 categories a gain, resulting in an overall Solution score of 5.

CS3: When put to the test, the team finds a similar speed to their existing solution, scoring that category a 0. The hardware footprint is virtually non-existent because most of the product is maintained by the vendor, so that category is scored 1. However, the accuracy rate is lower than the requirement, so the team scores that a -1 for a total of 0. The overall Solution score is 3.

5.4 Confidence (in the assessment)

5.4.1 Transparency

5.4.1.1 Data

- Is there transparency about the data used to develop the system?
- What types of data were used to create (train/test) the model?

1	No info about training data (if used).
2	Vague allusions to the type of data used, no significant detail.
3	Unannotated training data is available.
4	Training data is available with some supporting detail: an outline, description, etc.
5	Training data is available and accessible, with annotations, explanations, and a clear methodology for selection and inclusion.

CS1: The vendor provides no information about their training data, therefore scored as a 1.

CS2: The system uses an unsupervised process for detecting anomalies and does not require training data as such. The vendor provides several demonstrations of appropriate use cases for their AI module. The team scores this a 3.

CS3: The vendor describes scraping web images but does not make their core dataset available. The assessors score this a 2.

5.4.1.2 Methods

Is information provided about the underlying AI methodology and/or description of how the model was created?

1	No information about the underlying algorithms is provided.
2	High level descriptions of methods but no information on specific algorithms, e.g., "Unsupervised machine learning in order to…"
3	Specific algorithms are mentioned without context for their use within the system.
4	Specific algorithms are mentioned along with a high-level description of how they apply to the problem and system.
5	Detailed description of the underlying AI component, with specific technical detail justifying the application.

CS1: Despite the lack of transparency in other areas, this vendor provides detailed descriptions of the underlying machine learning algorithms, including academic publications outlining some of their novel approaches and public talks at security conferences. The team score this a 5.

CS2: Vendor documentation and interviews name drop several specific algorithms and techniques (clustering, various types of time series decomposition, Bayesian distribution modeling, and correlation analysis), but with no context about how they are employed within the system. The lack of context leads the team to score this a 3.

CS3: There is very little information provided about underlying algorithms. The team scores this a 1.

5.4.2 Documentation

5.4.2.1 ABOUT-ML (cards/sheets)

Does standard machine learning model documentation exist, such as Model Cards [2] [3], Data Sheets [4], or ABOUT-ML [5] data? These documents provide information about the training data, such as type, distribution, biases, etc., and the learned model, such as training method, domain of application, etc.

1	No Model Cards, Data Sheets, or other standardized model descriptor.
2	Model Card (or other) exists but with little usable information.
3	One non-standard but sufficiently detailed description of the model.
4	One standardized description of the Model with usable information.
5	Model is well documented across 2+ standards, with useful information in each.

CS1&2&3: There are no model cards for these products.

5.4.2.2 White papers

Are detailed technical whitepapers provided which have adequate detail for validation? ARCCS assigns a maximum score of 3 for the Whitepaper category. While vendor information is important for assessing a solution, an emphasis should be placed on independent verification.

1	Nothing found.
2	Whitepapers exists, but with little information sufficient for evaluation.
3	Whitepapers exist with sufficient technical information to assist the evaluation.
4	
5	

CS1: Whitepapers are easy to find and plentiful, at varying degrees of detail. Third-party evaluations can be found as well, so the assessors scored this a 3.

CS2: No first-party whitepapers were found. This scores as a 1.

CS3: No whitepapers found. This scores as a 1.

5.4.2.3 Publications

Are detailed technical conference or journal papers provided which have adequate detail for validation?

1	No publications of note found.
2	Publications on the overall system exist, but no information about the AI component
3	Publications describing the AI component exist, but are low-quality and/or lack sufficient detail for evaluation of the product
4	Publication in a journal in the relevant field with sufficient technical information for evaluation
5	Multiple publications in journals in the relevant field with detailed technical information

CS1: The assessment team discovered several detailed academic journal articles in the process of their review describing some of the algorithms used by the system. The team scores this a 5.

CS2: Third-party publications utilizing the products were found, but none from the vendor, and the publications were more focused on the applications of the ML techniques rather than going deep into detail about how they worked. The team judged this metric a 3.

CS1: No academic publications can be found. The team scores this a 1.

5.4.2.4 Patents

Are detailed patents provided which have adequate detail for validation?

1	No patents found.
2	
3	Patents were discovered but with little technical documentation of the AI component.
4	
5	AI component is described in the patent with specific technical detail.

CS1 & 2: Both products have several patents, but the underlying AI components are only described in high level terms. The team scores both as a 3.

CS3: A patent was discovered but the AI component was not mentioned. The team scores this a 3.

5.4.3 Other Information available

5.4.3.1 Specification of relevant use-cases

Does any documentation specify the relevant use-cases?

1	Vendor documentation does not hint at relevant use cases or claims that their tool may be applied to any problem.
2	Documentation contains a passing reference to one or two use cases with no detail on how it may be relevant (e.g., "Our tool uses AI to secure email").
3	Documentation specifies relevant use cases at a high level or in generalities with references to why their approach is relevant to the problem.
4	Documentation specifies multiple use cases with a high degree of technical specificity.
5	Documentation specifies multiple use cases with a high degree of technical specificity and includes examples of and comparisons to other relevant approaches (not strawman or toy demonstrations).

CS1&2: The vendors specify their use cases very specifically, but do not explore alternate approaches. The team scores both a 4.

CS3: The vendor specifies a relevant use case, but only talks about their approach in very general terms. The team scores this a 3.

5.4.3.2 Who developed the product?

Track record/industry reputation of vendor or system provider.

1	Vendor is a new or unknown startup.
2	Vendor is a relative newcomer but has had some success in this area.
3	Vendor is well established with a history of successful products, but not for this specific purpose.
4	Vendor is an industry leader in a similar technical area, expanding into a new market using an established approach.
5	Vendor is an industry leader with a well-established history of technical development in this area.

Cyber Profile
Some well-known vendors include McAfee, Symantec, Palo Alto, Crowd Strike, Splunk, Cisco. This is not an exhaustive list, and we make no endorsement of any of these organizations, this list is provided as an example of well-established industry leaders.

CS1: The product vendor is an industry leader with a history of developing market-defining products in this field. The assessors rank this a 5.

CS2: The vendor is a relatively new startup, but the product under assessment has been adopted by several Fortune 500 companies. Due to increased adoption, the team increases the score from a 1 to a 2.

CS3: This vendor is a raw startup and has very limited adoption in its field. The team scores this a 1.

6 Scoring

6.1 Dimension Score

The Dimension score is an average which captures the Relevance, Competence, and Cost sections of the assessment. This metric serves to give an at-a-glance measure of how well a tool performed in the assessment, but it is important to understand that because of the ordinal nature of the assessment, some context will be lost when not considering each metric on its own. The Dimension score is the result of calculating the average score across the three categories, with any N/As dropped out of the equation.

6.2 Confidence Score

The Confidence score is again taken by calculating the averages of the Confidence section features of the assessment. This metric serves to present a single score describing the overall quality of the information found during the assessment.

6.3 Strength of Assessment

The Strength of assessment is the percentage of answered questions in the Relevance, Competence, and Cost sections. Any N/As that are dropped from previous section will reduce this score. An assessment with a low strength score should be treated with skepticism, as this indicates that there is missing information or unanswerable questions. It is possible for an assessment to display a high Dimension Score and a low strength, as N/As are dropped from Dimension score calculation.

6.4 Case Study Outcomes

By entering our notional case studies into the tool, we get the final scores for each. In the figures below, each normalized score is plotted on one dimension of the spider plot.

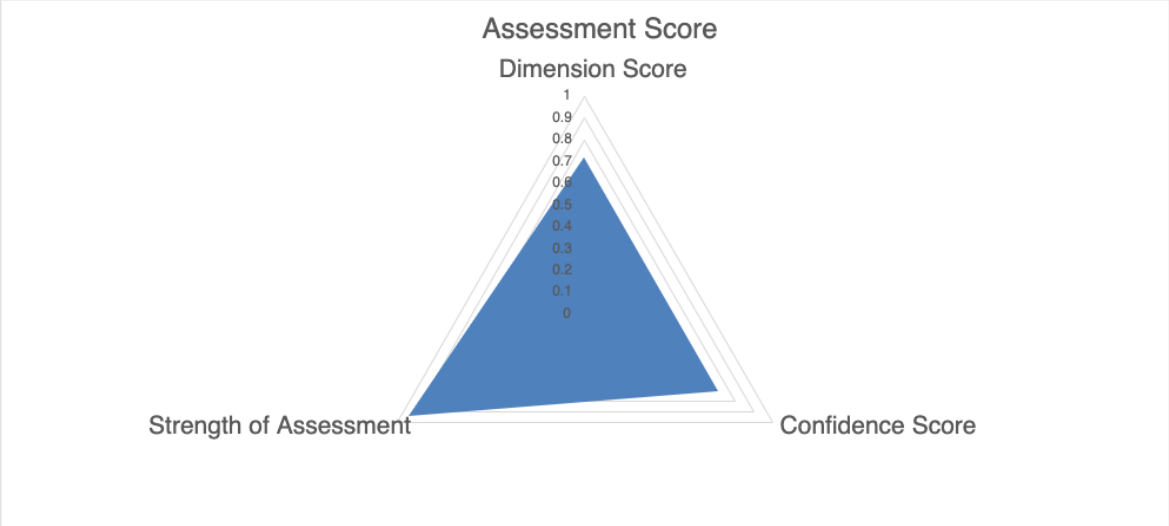

Figure 1: Spider plot Case Study 1 final scores

Dimension Score: .71
Confidence Score: 0.71
Strength of Assessment: 0.94

Figure 2: Spider plot of Case Study 2 final scores

Dimension Score: .72
Confidence Score: .53
Strength of Assessment: 1.0

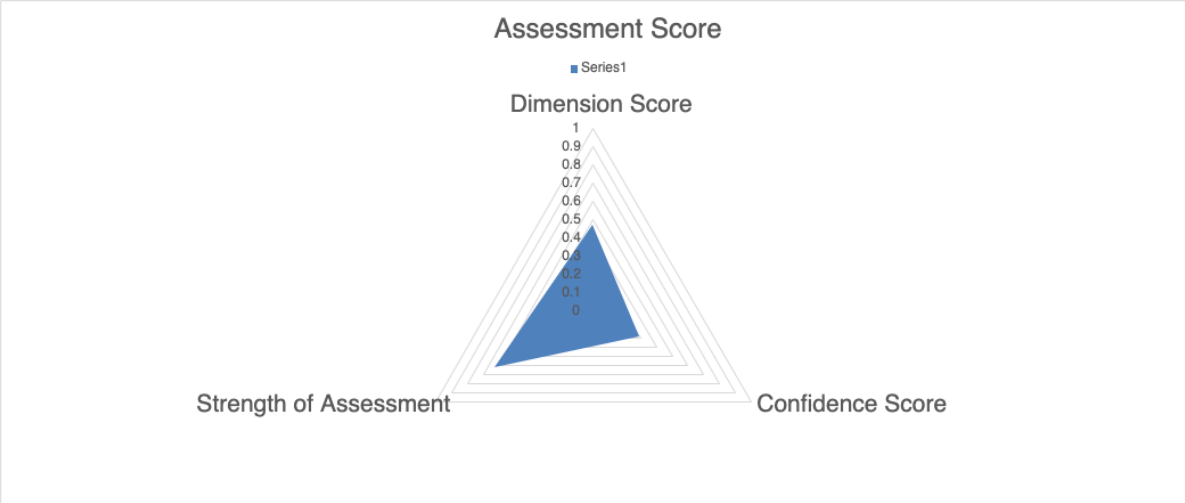

Figure 3: Spider plot of Case Study 3 final scores

Dimension Score: .47
Confidence Score: .29
Strength of Assessment: .625

7 Future Work

ARCCS currently treats every metric as equal in terms of score calculation, but in practice this is generally not the case. Some metrics (e.g., Goodness of Fit, Necessary vs. Gratuitous AI) are more technically important than others (e.g., Used by Other Organizations). This may seem obvious to an expert performing the assessment or examining a completed assessment line by line, but there can be an impulse to boil everything down to a single number. Summarizing information in this manner will generally lead to a lack of context, but ARCCS could mitigate this by assigning weights to individual metrics, which would allow important metrics to better influence the final scoring.

There is also room to extend the "Profiles" aspect of ARCCS, gearing it toward more specific use cases. We have presented a cyber profile in this paper, but AI has broad applications in many domains. Extending ARCCS to better capture, for example, biomedical or linguistic applications would give assessors examining products in those domains more confidence in their selections.

Appendix A – Mind Map of ARCCS Features

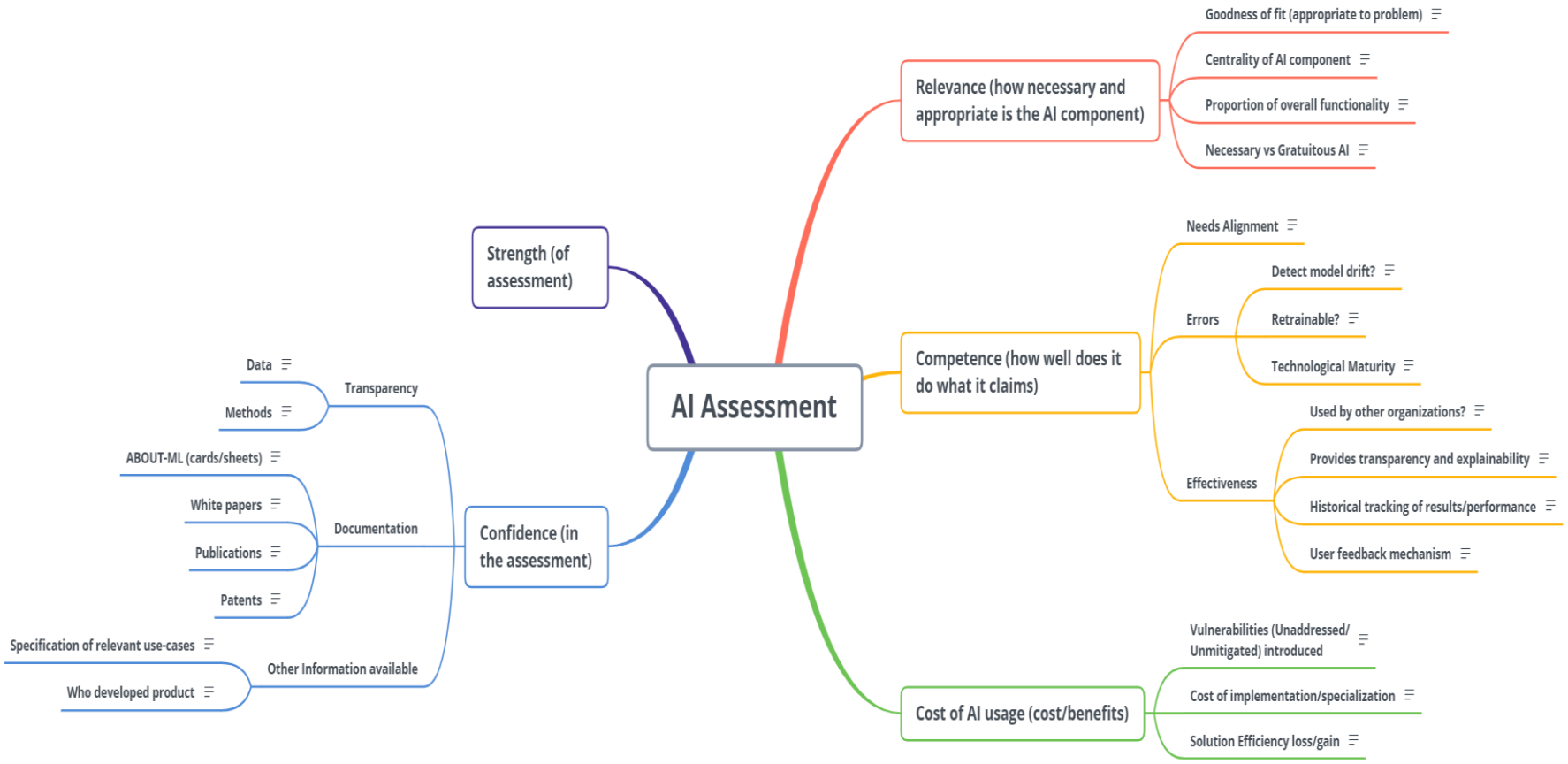

Figure 4: Mind map of ARCCS three primary dimensions along with the confidence and strength assessments

Appendix B – Assessment Summary Matrix

	1	2	3	4	5
Relevance					
Goodness of fit (appropriate to problem)	Approach is counter to best practice. Academic research/other publications/ evaluator experience indicates that the approach has serious flaws or is arbitrary.	Approach is unproven, hypothetical. No supporting documentation can be found but product can be demonstrated to do what it claims.	Approach has some 3rd party evidence supporting claims of functionality.	Approach is widely accepted and has been frequently applied in similar or related contexts.	Approach is widely accepted and has been frequently applied in the specific use case.
Centrality of AI component	ML/AI component is secondary to the operation of the system; "tacked on" and not directly relevant to the rest of the system.		ML/AI component is an important feature of the overall system, but is not required for successful system operation (e.g., a module) but in direct support of the mission.		ML/AI component is a keystone feature of the system; System has been designed around this component.
Proportion of overall functionality	Component is redundant or not directly relevant to the rest of the system.	Component is a small proportion of overall system functionality (<25%).	AI component is a significant portion of overall system functionality (<50%).	AI component is a large portion of overall system functionality, but there are additional features that can function in its absence (>50%).	AI component encompasses most of the functionality of the system (excluding infrastructure e.g., data storage and transport). The system does not function properly in its absence.
Necessary vs Gratuitous AI	AI is unnecessary; functionality could be accomplished with non-AI component.	AI provides a slight advantage over non-AI solution.	AI provides a marginal improvement over a non-AI solution.	AI Provides a significant improvement (accuracy, speed, etc.).	Task could not be realistically accomplished without AI.

Competence					
Needs Alignment	Performance of the system is clearly worse than the standard or does not function as advertised.	Performance of the system is slightly worse than the industry standard.	Performance of the system is on par with industry average.	Performance is a significant improvement on the industry standard.	Performance represents a breakthrough advancement in capability from current industry standard.
Detect model drift?	System has no ability to detect model drift.				System has ability to detect model drift and notify operators.
Retrainable?	The system requires training and cannot be retrained.	System retraining can only be done by the product vendor and has an excessive turnaround time or cost.	The system requires retraining, but this can only be done by the product vendor.	The system requires training, and can be undertaken by system operators, but may be difficult to do (e.g., arcane process, onerous downtimes).	The system does not require training, or requires training that can be easily undertaken by the operators through a simple process.
Technological Maturity	The AI component is based on recent research or is largely unproven.	AI technology has been in development for two or more years, preferably with proven application.	AI technology is well known and has been actively developed and researched for 5+ years.	AI technology is mature (7-9 years).	AI technology is mature and widely accepted, in development or practice for over a decade.
Used by other organizations?	No other known organizations using the system.	System has been adopted by a few (<10) organizations.	System enjoys some industry adoption (10+), at least one of which is using it for a similar use case to the evaluators.	A wide array of organizations (50+) currently uses the product for a diverse set of use cases, including the evaluators'.	System is considered an industry leader and is adopted by many organizations (100+).
Provides transparency and explainability	AI component is a complete black box with no transparency into system operation, and no algorithmic or domain explainability.	AI component is somewhat transparent, but mostly obfuscated from the user.	AI component provides either transparency into system operation or explainability, but not both.	System is transparent and at least partially explainable, or vice versa.	AI component is both transparent and explainable.
Historical tracking of results/performance	System does not track historical results and performance.		System does not track historical results but does not require retraining.		System tracks historical results and performance in a sufficient manner to facilitate retraining.

User feedback mechanism – Feedback specifically for AI performance improvement.	System has no ability for users to provide feedback.				Allows users to override or change results, system accounts for these overrides during future operation.
Cost					
Vulnerabilities Introduced	System is vulnerable to poisoning and/or evasion attacks and draws from publicly accessible and editable datasets with no mitigations in place.	System has no known mitigations but draws from a protected, controlled dataset.	System implements defense(s), but these defenses are unproven or unsupported by literature.	System implements proven defenses appropriate to the problem; these defenses are consistently updated and maintained by the vendor. Alternately, system contains redundancies such that a poisoning or evasion attack could be easily detected within its system.	Due to design, poisoning and evasion are not a concern.
Cost of implementation/specialization	AI component requires a large degree of specialization and training to operate.	AI component requires some specialization and training, plus a degree of tuning and configuration to deploy.	AI component has some minor configuration cost and training requirements.	AI component has either some small training requirements OR configuration cost associated with deployment, but not both.	AI component is functional out-of-the-box with very little training required.
Solution efficiency loss/gain	There is a demonstrable loss of efficiency with the solution.	There is no appreciable gain or loss of efficiency with the AI component.	There is a notable (>10%) gain in efficiency, accuracy, or speed.	There is a large (>25%) gain in efficiency, accuracy, and speed.	The system does something that would not be possible without the ML/AI component.
Confidence					
Data	No information about training data is available.	Allusion to the type of data used without significant detail.	Training data is available.	Training data is available with some supporting detail.	Training data is accessible, with annotations, explanations, and a clear methodology for selection and inclusion.

Methods	No information about the underlying algorithms is provided.	High level descriptions of methods are provided but no information on specific algorithms etc. (e.g., "Unsupervised machine learning to...")	Specific algorithms are mentioned without context for their use within the system.	Specific algorithms are mentioned, along with a high-level description of how they apply to the problem and system as a whole.	Detailed description of the underlying AI component, with specific technical detail justifying the application. Bonus points for novel approaches published in academic journals with sufficient rigor.
ABOUT-ML (cards/sheets)	No Model Cards, sheets, or other standardized model descriptor.	Model Card (or other) exists but with little usable information.	One non-standard but sufficiently detailed description of the model.	One standardized description of Model with usable information.	Model is well documented across multiple (2+) standards, with useful information in each.
Whitepapers	Nothing can be found	Whitepapers exist, but with little information sufficient for evaluation.	Whitepapers exist with sufficient technical information to assist the assessment.	N/A	N/A
Publications	No publications of note are found.	Publications on the overall system exist, but no information about the AI component.	Publications describing the AI component exist, but are low-quality and/or lack sufficient detail for evaluation of the product.	Publication in a journal in the relevant field with sufficient technical information for evaluation.	Multiple publications are found in journals in the relevant field with detailed technical information.
Patents	No patents found.		A patent was discovered but with little technical documentation of AI component.		AI component is described in the patent with specific technical detail.
Specification of relevant use-cases	Vendor documentation does not hint at relevant use cases, or claims that their tool may be applied to any problem.	Documentation contains a passing reference to one or two use cases with no detail on how it may be relevant (e.g., "Our tool uses AI to secure email").	Documentation specifies relevant use cases at a high level or in generalities with references to why their approach is relevant to the problem.	Documentation specifies multiple use cases with a high degree of technical specificity.	Documentation specifies multiple use cases with a high degree of technical specificity and includes examples of comparisons to other relevant approaches (should not be strawmen approaches)
Who developed product	Vendor is a new or unknown startup.	Vendor is a relative newcomer but has had some success in this area.	Vendor is well established with a history of successful products, but not for this specific purpose.	Vendor is an industry leader in a similar technical area, expanding into a new market using an established approach.	Vendor is an industry leader with a well-established history of technical development in this area.