ASPEN POLICY ACADEMY
aspen institute

Operational Plan

PACT AI Playbook

Public Accountability for Community Trusted AI

By Addie Achan and Rebecca Odim

November 2026

This playbook was developed by Aspen Policy Academy Fellows while participating in the 2026 Science and Technology Policy Fellowship. This playbook is intended for staff of the New York City Office of Algorithmic Accountability to assist city agencies and staff members exploring the use of artificial intelligence to help serve residents. The full project, including a policy brief explaining the Fellows' core recommendations, is available here. Please note that all authors' opinions published here are their own. This project was completed independently of New York City's Office of Technology and Innovation. This publication does not reflect the views of the Aspen Policy Academy, the Aspen Institute, or the New York City Office of Technology and Innovation.

Overview

This Playbook is intended for staff of the New York City Office of Algorithmic Accountability (OAA) to assist city agencies and staff members exploring the use of artificial intelligence (AI) to help serve residents. Specifically, this Playbook should be followed by OAA staff (with the support of others as indicated later) whenever a New York City agency or staff member indicates that they are exploring procuring a specific vendor's AI tool or exploring a use case for AI in a way that may affect residents.

Following this end-to-end guidance should enable the OAA to:

- Effectively collect community input that accurately reflects surfaced needs;
- Translate that input into standard evaluations to ensure those needs are met; and
- Report the results back to the community to promote transparency.

Owners

Each section of the Playbook recommends a specific owner, identified at the top of that section. Recommended owners were selected based on the role, responsibilities, and expertise they bring to the specific work outlined here. Each owner is responsible for their own step(s) in the process. The 'owners' (or key users) of this Playbook may include the

OAA product, data, and engineering leads; New York City agency or staff member(s) interested in procurement; community based organization facilitator(s) from the Grassroots AI Education and Engagement Initiative; and the New York City Public Interest Technology (PIT) Crew.

Final Deliverables

Each step of the Playbook produces a specific deliverable. These deliverables are combined to produce the PACT Evaluation Scorecard, which includes the following:

- **Run record:** A list of community authored tests (evaluations) run and their results, broken down by each aspect of AI functionality (claim).
- **Routed community input registry:** A record of community input that could not be used in the evaluation and how it was addressed through other means.

PACT AI Playbook Process Overview

Office of Algorithmic Accountability (OAA)

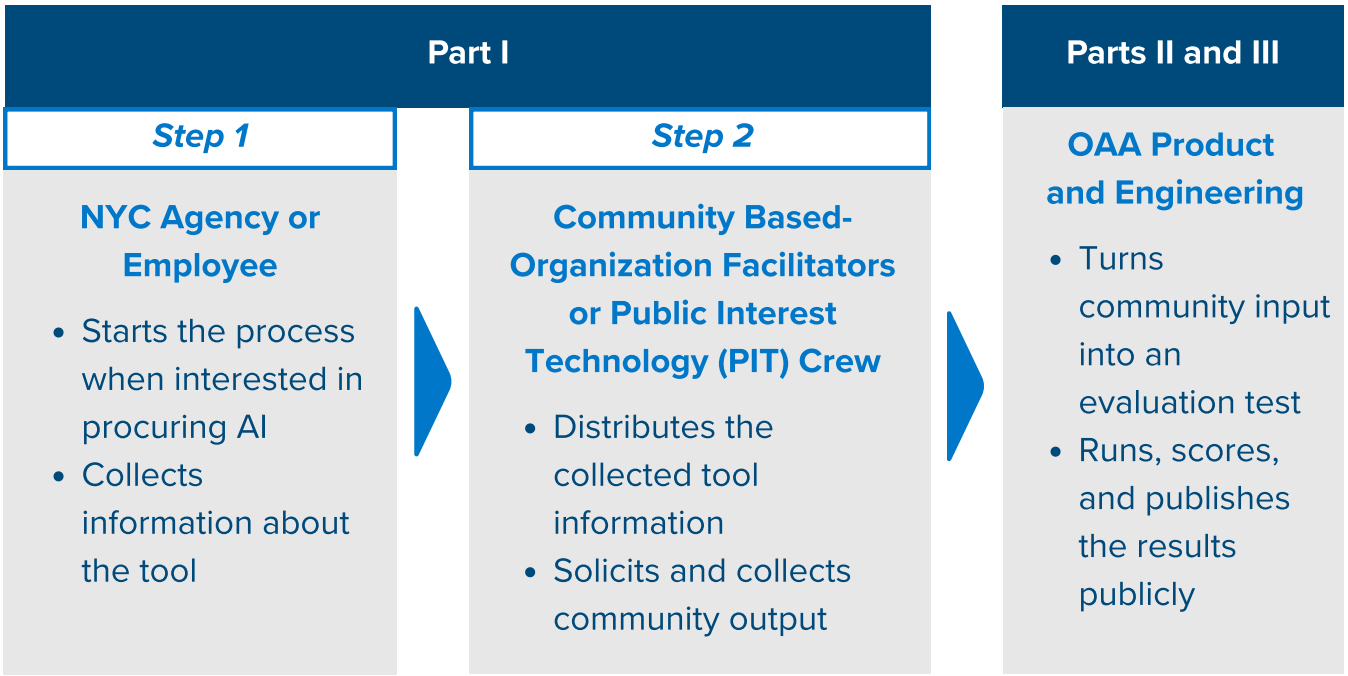


Table of Contents

Overview....1

Owners....1

Final Deliverables....2

Common Terms....3

Part I: Community Input....5

Steps....5

Outputs Produced....5

Step 1: Gather Educational Information and Consent....5

Step 2: Solicit Formatted Community Input....6

Part II: Build and Run an Evaluation....9

Prerequisites....9

Steps....9

Outputs Produced....9

Step 1: Create the Dataset....9

Step 2: Run the Evaluation....12

Step 3: Grade and Record Results....14

Part III: Reporting and Transparency....16

Steps....16

Outputs Produced....16

Step 1: Publish....16

Appendix A: AI Vendor Questionnaire....18

Appendix B: Community Input Card....23

Appendix C: Evaluation Scorecard....27

Appendix D: Algorithmic Tools Compliance Report....28

PACT Evaluation Scorecard....29

Common Terms

This is a list of common terms used throughout the Playbook.

- **Evaluation:** A fixed collection of questions, answers, and written grading rules given to an AI system to assess behavior and performance over a certain domain or set of tasks.
- **Test case (prompt):** A single question in an evaluation.
  - Example: “Can a landlord refuse a tenant with a housing voucher?”
- **Test set (or dataset):** The full collection of test cases and mapped correct answers.
- **Rubric:** Grading instructions for classifying outputs as pass, fail, or partial.
- **Claim:** A testable statement about the AI tool that can be classified as true or false.
  - Example: “The chatbot gives legally correct housing answers in Spanish.”

Emile Perron via Unsplash



Part I: Community Input

For an AI tool to work effectively for a community, it is necessary to gather the community's input and thoughts on its use. Part I of the PACT AI Playbook lays out the exact steps that the OAA should follow (largely by delegating them to specified owners) to gather community input, which will be instrumental in shaping and assessing AI deployments.

Steps

1. Gather educational information and consent.
2. Solicit formatted community input.

Outputs Produced

- A collection of formatted and unformatted community inputs.

Step 1: Gather Educational Information and Consent

Overview:

New York City agencies and staff members engaging in the PACT procurement process should gather basic information at the start of the process (e.g., how does the tool work, what data does it use, what safety measures have been taken), though the exact information needed will vary by the type of procurement. This step highlights how the relevant staff member can find this information. The OAA should ensure that this step is done any time it is approached about potential procurement. When procuring a specific AI tool, confirm that the vendor has provided their consent and access to the application programming interface (API) for later stages of the process.

Owner: New York City agency procurement team or staff member

Inputs: N/A

Outputs Produced:

- A collection of resident friendly materials that give community members information about the tool or process they are providing input on.
- If needed, recorded confirmation of a vendor's consent.

Execute:

1. Collect as much information about the tool(s) as possible. The goal is to collect

relevant information that residents would need or want to know when providing their input. The AI Vendor Questionnaire (Appendix A) can be used to gather this information at once.
    a. The information collected should include, at a minimum, (i) the type of information that goes into the tool, (ii) the intended output, and (iii) the problem that the tool is trying to solve.
2. If necessary, contact the vendor to gain API access and record their consent.
3. Make the information collected resident friendly by removing jargon and including only what community members need to understand about the tool.
    a. Recommended format: 1 page document or short presentation.
4. Give the information back to the OAA.

Helpful Tips:

- Some baseline information may not be obtainable. In those cases, explicitly mention that to residents, as this will help define the tool's boundaries, signal that something important is missing, or guide their questions.
- The New York City agency or staff member initiating the procurement process should complete most of this step. If they do not have the necessary information, they can engage the OAA for support or try alternative resources, such as:
    - Vendor's website.
    - When interested in procuring a tool used by another city agency, Local Law 35's Algorithmic Tools Compliance Report may already have it recorded.

Next Steps:

Once the necessary information has been gathered, the OAA can hand it over to the community based organization facilitators or the PIT Crew, as specified in the next step, engage the community, and gather their input.

Step 2: Solicit Formatted Community Input

Overview:

Now that baseline information is ready to be shared with the community, community based organization facilitator(s), such as the Grassroots AI Education and Engagement Initiative, and City staff, such as members of the PIT Crew, should solicit the input. The input collected will help define the tests that must be passed for procurement. A wide range of input, including concerns, expectations, and needs, is the most helpful.

The input must be captured in a specific format to support the rest of the steps. This is meant to facilitate the process, not guide the community to any specific outcome or conclusion.

Owner: Community based organization facilitator(s) from the Grassroots AI Education and Engagement Initiative and PIT Crew

Inputs: Educational materials produced in Step 1

Outputs Produced: A collection of formatted and unformatted community inputs

Execute:

1. Choose a channel to reach residents:
    a. In person: Community based organization facilitators can integrate this process of soliciting input during the listening or education sessions that they already lead through the Grassroots AI Education and Engagement Initiative. This is a more direct way to engage the community, allows clarifying questions to be addressed immediately, and leverages existing gatherings.
    b. Remotely: PIT Crew members can release text- and email-based targeted surveys to reach a broader set of the community at low cost.
2. Share the resident facing 1-pager.
3. Share and have residents fill out a Community Input Card (Appendix B), which is a low friction way to collect community input in residents' own words that is structured to be easily converted into a piece of an evaluation.
4. Collect and share all Community Input Cards with the OAA.

Helpful Tips:
The priority in gathering this input is to capture the community's voice as authentically as possible. To achieve that:

- Encourage community members to be as thorough as possible when describing their answers about an imagined interaction and what they want and do not want to see in a response. Specific details will help guide the build evaluation (from Part II) to more precisely capture community needs.
- Some residents may want to provide input on multiple imagined tool interactions. Have the residents fill out a new card for each interaction.

- Avoid relying on 1 person or small group. Gather input from as many people as possible.
- While collecting input using the Community Input Card can be helpful, it should not be a barrier to participation. Not all community input will fit neatly into the format, but all relevant feedback should be documented, as it will be captured and addressed later through the routed community input registry (shown in Part II).
- If possible, use both in-person and remote channels to collect input to reach a broader subsection of the community.
- For in-person sessions led by community based organization facilitators, consider having a PIT Crew member or OAA staff member on hand to answer any clarifying questions from residents that may benefit from a technical lens.

Next Steps:
After the community based organization facilitator or PIT Crew member has collected community input, it should be passed to the OAA for execution of Part II.

Jonny Gios via Unsplash



Part II: Build and Run an Evaluation

In Part II, the community input collected in Part I is turned into an evaluation that measures how well the tool meets the community defined needs. This section provides an end-to-end walkthrough starting with baseline dataset creation and ending with a scorecard detailing performance. This part of the process leverages Inspect AI, an open source evaluation framework.

Prerequisites

The collection of community input from Part I.

Steps

1. Create the dataset.
2. Run the evaluation.
3. Grade and record the results.

Outputs Produced

- PACT Evaluation Scorecard, which details how well a given AI tool addresses community defined needs. The scorecard includes:
    - **Run record:** A list of community authored tests (evaluations) run, along with results, broken down by each aspect of AI functionality (claim).
    - **Routed community input registry:** A record of community input that could not be converted into the evaluation test and how it was addressed through other means.

Step 1: Create the Dataset

Overview:

To run an evaluation, the team must first create a dataset of tests and corresponding rubrics for grading. These tests should be grouped by different facets of functionality (i.e., accuracy, bias, etc.), also known as claims or claim types. The community input collected in Part I provides the foundation for both pieces of this foundational dataset in the evaluation. The input may need to be organized, reviewed, and lightly reformatted so that it can be translated into a dataset.

While there are many ways to create and run an evaluation, this process uses the Inspect AI open source framework.

Owner: OAA product, data, or engineering staff

Inputs: The list of community concerns from Part I (including formatted and unformatted pieces of input)

Outputs Produced:

- A collection of dataset files (JSON or CSV), each centered around a different claim (i.e., theme), and containing rows with prompts to send to the model, traits of a target answer, and grading guidance.
- Routed community input registry for inputs that couldn't be included in the dataset.

Rows within the dataset sample for the claim of accuracy

ID	Input	Target	Grading Guidance
1	"Do I qualify for food stamps?"	Ask for and explain the basic factors that determine SNAP eligibility	Fail responses that provide baseless verdicts
2	"When does the East Tremont office close on Sunday?"	Provide the verified hours that point to a source or state that info is not available	Fail responses that state incorrect times
3...	....	...	...

Example rows within the routed community input registry

ID	Original Input	Rerouting Reason	Routed to
1	Where will my data go?	Systems level question that can't be answered by interacting with the tool	Vendor
2	I don't have a smartphone. Will a human agent still be there?	Program service level question that can't be answered by interacting with the tool	New York City agency interested in procurement
3...	....	...	...

Execute:

1. Create a blank spreadsheet titled “Routed Community Input Registry.”
2. Create a blank dataset JSON/CSV file for each claim type (i.e., an explicit testable property of an AI model’s performance).
    a. Generally, this means 1 claim around model accuracy, another for bias, etc.
3. In each dataset, add columns as necessary. The Inspect AI guidance has a list of potential columns (at a minimum, expect to have “input,” “target,” “grading guidance”).
4. Triage the community input collection by evaluating whether each piece describes a testable interaction (i.e., can easily be tested by sending a message to the tool and getting a response back).
    a. If yes, add it as a new row to the dataset file associated with the claim it falls under and fill in the columns as needed.
    b. If not, add it as a new row in the routed community input registry.
5. If applicable, verify any facts. Correct misinformation or remove rows that can not be verified if necessary.
6. Create the starter PACT Evaluation Scorecard (Appendix C) and add the routed community input registry items. Check whether the associated vendor has already addressed some of these concerns by completing the AI Vendor Questionnaire (Appendix A).
    a. If the questionnaire hasn’t been completed, the staff member or agency should follow up with the vendor directly.

Helpful Tips:

- The 3 guiding questions in community input follow this rough mapping to the dataset file:
    - Q1 ➜ input
    - Q2 ➜ target
    - Q3 ➜ grading guidance
- The above mapping is suggested, but it’s recommended to read the entire input before triaging, as some substantial ideas may be misplaced.
- Separating by claims yields more thorough, specific results than testing everything together.
- Some inputs are multi-turn (i.e., back-and-forth) rather than single prompts. Create a separate claim dataset for multi-turn conversations and populate the input for each row with the full conversation script.

- The dataset for each claim type may have unique columns (e.g., testing bias may require a column highlighting demographic information).
- The same community piece may be added to multiple claim type datasets.
- Fill in gaps for incomplete or light community input if intuitive, but avoid making assumptions that could change the meaning of the input.
- Not all input pieces will have facts to verify.
- Some input pieces may contain multiple scenarios. In those cases, split into multiple rows.

Next Steps:
Now that the data are ready, it's time to run the evaluation.

Step 2: Run the Evaluation

Overview:
This is the core of the process. During this step, the OAA product or engineering lead will feed the now structured community input data into the tool under evaluation and receive its output.

Owner: OAA product, data, or engineering staff

Inputs:
- The set of dataset files from Step 1.
- API access from the vendor (should be gathered as part of the initial consent when setting up the process).

Outputs Produced:
- A run record that holds the details of the test.
- An evaluation log that holds each output generated by the AI tool mapped to the original input and the grade (along with reasoning) based on the grading guidance.

Execute:
*Repeat this process for each claim/dataset available.

1. Create a Python file to house a task object as detailed in the Inspect AI guidance with 3 key pieces:
    a. A reference to the dataset file that is being tested.
    b. A solver (a function that sends the inputs to the tool).

    c. A scorer (a function that defines how to use the grading guidance to evaluate this claim).
2. Take a small selection of the questions to create a sample dataset (3–5 is fine).
3. Run the solver on the sample dataset and confirm basic functionality (otherwise adjust).
4. Run the evaluation on the full dataset.
5. Document the run record, including the version tested and date, in a run record file.

A card from the evaluation log generated from running the evaluation (Appendix C). This will be generated during the full data set run.

Helpful Tips:

- The Inspect AI guidance has many optional layers to the solver. While those options should be chosen with discretion, the core generate() command is often sufficient.
- Each claim needs its own unique scorer.
- For the third execution step, evaluation datasets are large, so starting with the sample batch can help identify major issues early.
    - When checking for issues, the staff member should focus on issues with the setup. This can include blank outputs or those that show that the input isn't being sent as intended. Poor quality answers that address the actual prompt are acceptable.
- The evaluation can test each row multiple times. This helps address the stochastic nature of AI (the same input to the same tool can yield different outputs when tried multiple times).
    - Each try adds an additional multiplier to the grading work. While no single industry standard exists, 3–5 is a good default. Anthropic's guidance can be leveraged to determine an exact number.
- If testing a multilingual dataset, ensure that the outputs are in the same language as the inputs.
- Multi-turn datasets will likely evaluate full conversations, which requires specific logic in the associated solver and scorer.

Next Steps:

Now that the outputs are ready, it's time to assess their quality and create a report.

Step 3: Grade and Record Results

Overview:

This step makes the actual judgment call on whether a tool's output meets the community defined expectation. Grading involves reading each answer and scoring scoring each response as pass or fail, depending on if the answer matches the rubric. Once the full question set is graded, the aggregate pass rate is listed on the evaluation scorecard as a percentage.

Owner: OAA product, data, or engineering staff

Inputs: Evaluation log

Outputs Produced: PACT Evaluation Scorecard

Execute:

This step involves a collection of actions:

1. Export the evaluation log into a form that human graders can regrade. Each row should have a human grade, the name/initials of the person who graded it, and, if it wasn't a pass, quotes from the rubric as the justification.
2. Have a second grader conduct the same process on a statistically representative sample (and on all ambiguous, high stakes, or failed answers), ensuring that all previous grades are hidden. While some discrepancies are understandable, consistent patterns may expose issues with the rubric. Rewrite any unclear rules and have the graders regrade any answers the rules affect. Note the final agreement rate (matching grades/total).
3. **(Optional)** If assessing how a tool performs across groups, repeat the scoring process, add a filter for the group tag, compute each group's score, and check for notable differences. While raw scores are fine, another option is to consider writing them as a ratio in which each group's score is divided by the best group's score, similar to Local Law 144's bias audit reporting.
4. Note the average score from all of the runs, along with the cross group score differentials.
5. Finally, add these results to the run record section of the tool's PACT Evaluation Scorecard.

Claim Evaluated	Eval Version	Run Date	Tries per Question	Score	Grader Agreement Rate
Accuracy	v0.1	2026-07-20	10	95%	50%
Accuracy	v0.2	2026-08-21	10	87%	90%

Example run record rows for the claim of accuracy. For the first run (v0.1), 95% of responses matched the rubric. The human grader agreed with the automated grader 50% of the time. In the second run (v0.2), 87% of responses matched the rubric. The human grader agreed with the automated grader 90% of the time.

Helpful Tips:

- During grading, ensure that any identifying information about the user and the AI grade is not visible, so the grader isn't anchored.
- When choosing answers to be spot checked, include every grade that was marked a fail. Failing grades are the most consequential, so they deserve a second pair of eyes.
- A fixed slice of all answers gets double graded along with all the Fails. While the exact amount will depend on the size of the evaluation, resources are available to help determine statistical significance.

Next Steps:

Repeat evaluation as necessary for each claim. Once the PACT Evaluation Scorecard is complete, it's time to distribute and publish it in Part III.

Jonahkoh_ via Unsplash



Part III: Reporting and Transparency

For each tool assessed in Part II, there should be a PACT Evaluation Scorecard that includes a run record of all evaluations run and a routed community input registry that demonstrates what is known about the tool, its risks and impacts, and areas for further investigation. This section of the Playbook covers publishing and distribution.

Steps

1. Publish and distribute.

Outputs Produced

- Algorithmic Tools Report Update.

Step 1: Publish

Overview:

Once the evaluations and scoring of the tool are complete, the interpreted results are shared with the community transparently. These outputs will be published alongside the existing Algorithmic Tools Compliance Report to improve discoverability.

Owner: OAA product lead

Inputs: PACT Evaluation Scorecard

Outputs Produced: The PACT Evaluation Scorecard is published alongside the existing Algorithmic Tools Compliance Report.



NYCOTI
Office of Technology & Innovation

LL35: Algorithmic Tools
CY 2025

Basecalled nucleotide sequences with quality metrics representing specimen genetic content, used for downstream genomic surveillance and outbreak analyses.

Vendor Information

Vendor Name(s): Oxford Nanopore Technologies

Vendor Type(s): Off-the-shelf

Developed and maintains the tool.

HELIX

Department: Division of Information Technology

First Used: November 2025

Purpose Type: Information presentation

Computation Type: Data transformation

Autonomy: Informative

Frequency of Use: Daily

Population Impacted: Individuals (Agency staff)

Vendor(s): Microsoft

Tool Description

HELIX is a generative tool that allows agency staff to ask general questions based on data available in the OpenAI GPT-4o model. In addition, HELIX can also be grounded on approved documents allowing staff to ask questions specific to those documents.

Tool Purpose

The tool is used to generate responses to prompts entered by staff, requesting the generative AI tool to compose a text response to a text input. These outputs are potentially shared with the public.

Data Analyzed

Identifying information may be collected or disclosed in the input data and output data.

Training Data

Training data is proprietary to the vendor.

Input Data

Prompts provided by the users of the system.

Output Data
Version effective March 2026

PACT Evaluation Scorecard

Tool Name: HELIX
Agency Requesting Procurement: NYC Human Resources Administration

What These Results Led To: Community concerns largely focused on accuracy. The initial evaluation (v0.1) was graded highly on a first pass but was heavily contested by a second grader. After using the same community input to create a more detailed grading rubric, the initial score decreased slightly, but the graders agreed more. The NYC Human Resources Administration adopted HELIX based on the positive results and sufficient answers to concerns around data and human alternatives.

Run Record

Claim Evaluated	Eval Version	Run Date	Tries per Question	Score	Grader Agreement Rate
Accuracy	v0.1	2026-07-20	10	95%	50%
Accuracy	v0.2	2026-08-21	10	87%	90%

Routed Community-Input Registry

ID	Original Input	Rerouting Reason	Routed to	Follow up
1	Where will my data go?	Systems question that can't be answered by interacting with the tool	Vendor	Vendor specified that data is sent to OpenAI to generate a response but has a zero-data-retention agreement.
2	I don't have a smartphone. Will a human agent still be there?	Program-service level question that can't be answered by interacting with the tool	NYC Human Resources Administration	Agency confirmed there will be a human contact point.
3...	....	...	...	

(Left) An excerpt from NYC's current Algorithmic Tools Compliance Report, as published by the NYC Office of Technology and Innovation.
(Right) The proposed new PACT Evaluation Scorecard that will be stored as well.
A larger version of the current report can be found in Appendix D, and a larger version of the proposed scorecard can be found on pages 29–31.

Execute:

1. Include a blurb that incorporates the PACT Evaluation Scorecard under the header "Community Driven Evaluation."
2. Share the revised Algorithmic Tools Compliance Report entry with the community based organization facilitator(s).

Taken together, these steps give the OAA a repeatable way to turn residents' own words into concrete tests, hold AI tools to those tests before they are procured, and report the results back to the communities they impact.

If you'd like to learn more, see the full project, including a policy brief explaining the Fellows' core recommendations, at aspenpolicyacademy.org/project/nyc-public-ai-accountability-2026.

Appendix A: AI Vendor Questionnaire

AI Vendor Questionnaire Overview

Collecting the AI Vendor Questionnaire is a key step in the Public Accountability for Community Trusted AI (PACT AI) procurement process. New York City government agencies (including procurement teams and individual staff members) that are assessing an algorithmic tool should confer with the Office of Algorithmic Accountability (OAA). The process includes the following steps:

1. Request that the proposed vendor complete the following questionnaire.
2. (Optional) Consult with the OAA before or during this process.
3. Once the vendor returns it, send the completed questionnaire to the OAA at xxx@nyc.gov.

The AI Vendor Questionnaire is intended to help the city determine whether a tool is appropriate for use by a New York City agency and to establish baseline performance metrics that the OAA will test over time. The questionnaire responses should establish the following:

1. Baseline criteria for whether to proceed with a vendor at all. Vendors that cannot provide answers to core questions like data practices or failure modes should be dismissed.
2. Any metrics or performance numbers that the vendor may already have about their tool. The results of the evaluations from the PACT AI Playbook can be compared with these claimed performance metrics.

The following pages contain the form to send directly to the vendor.

Sample AI Vendor Questionnaire

Please fill out the questionnaire below on behalf of the AI vendor you represent within 10 business days of receipt. Answer all questions and provide additional context where necessary.

How this information will be used: New York City's Office of Algorithmic Accountability (OAA) will use this questionnaire to determine whether a tool is appropriate for use by a city agency and to establish baseline performance metrics that the OAA will test over time. Responses may be summarized and published as part of New York City's public Algorithmic Tools Compliance Report. If any answer includes proprietary or confidential information, please note that explicitly next to the response.

Company Name	
Tool/Product Name	
Point of Contact	
Phone/Email Address	

1. **Describe the proposed tool and its intended output.** (e.g., a chatbot that answers questions about tenant rights, a triage system for benefits decisions, a review tool that produces a flag)

2. **What external evaluations have already been run on the tool and what were the outcomes?** Check the box next to each conducted evaluation type, and provide the evaluation's name/company and its outcome or score. Use the blank rows at the bottom to write in any additional evaluations conducted.

Conducted	Evaluation Type	Vendor/Company Name	Outcome/Score
☐	General purpose		
☐	Accuracy and hallucination		
☐	Bias		

3. **What information does it collect, and who can access it?** Specify what data the tool takes in, where it is stored, how long it is kept, who has access to it, and what it is used for. Check the box next to each data type that is collected or processed, and describe when and how it is used. Use the blank row at the bottom to write in an additional data type.

Collected/ Processed	Data Type	Where is it stored, for how long, who has access, and what is it used for?
	Resident personal identifying information (name, address, date of birth)	
	Employee/staff personal identifying information	
	Health or medical information	
	Immigration, legal, or criminal justice status	
	Children's or minors' data (anyone under 18)	
	Biometric data (facial, fingerprints, voiceprints)	
	Location or device data (GPS, IP address, device ID)	

Collected/ Processed	Data Type	Where is it stored, for how long, who has access, and what is it used for?
	Data generated by the tool (risk scores, classifications)	

4. **What information, if any, do you have about the tool's carbon footprint?** (e.g., energy sources, training energy usage, average query energy usage)

Energy sources (% renewables)	
Training energy usage	
Average query energy usage	
Other	

5. **What are the risks if it doesn't work as intended?** (e.g., what are edge cases or situations in which the tool should not be used)

6. **What real-world deployment stories and outcomes can you share about the tool?** (e.g., have you demonstrated the tool's effectiveness in real-world contexts?)

7. **Where, if anywhere, is a human involved during tool inference?** (e.g., does your tool leverage a human-in-the-loop at all? If so, how?)

8. **How will we know when the underlying model changes, and what is your process for notifying us before updates ship?** (e.g., how often can we expect updates and how much runway will we have to address harms?)

9. **Is this built on a third party model (OpenAI, Anthropic, Google, etc.), and what will happen to our deployment if that provider changes its terms or pricing or deprecates the model?** (e.g., Are there any potential dependencies we should be aware of? How can we be sure our tool doesn't break if one of your vendors changes terms?)

Certification

By signing below, I certify that the information provided in this questionnaire is accurate and complete to the best of my knowledge as of the date below. I understand that New York City's Office of Algorithmic Accountability may use these responses as a baseline to evaluate this tool's performance, and that materially inaccurate responses may affect the tool's suitability for procurement.

Name: _____

Title: _____

Signature: _____

Date: _____

Appendix B: Community Input Card

The Community Input Card is used in Step 2 of Part I of the Public Accountability for Community Trusted AI (PACT AI) procurement process. This card should be shared with community members participating in the community input process. To prepare the card for sharing with community members, a community based organization facilitator from the Grassroots AI Education and Engagement Initiative should create and link an educational 1-pager about the AI tool they are seeking feedback on.

The following page is the form to share directly with community members. The subsequent pages contain a sample filled out form.

Sample PACT Community Input Card

Instructions: New York City is considering adopting the following AI tool. We're putting together a test to see if the tool can meet the quality bar set by New York City residents, and we'll use your input to help shape it!

Imagine interacting with the tool. What would you want it to do, and what would you be concerned about it doing? Try to express your input by answering the 3 questions below and add as many details as possible.

We understand you may have thoughts on the tool that don't neatly fit into the questions. We would still love your feedback; feel free to describe it in your own words in the "Additional Information" section.

If you can imagine interacting with the tools in multiple ways, we would love to hear all of them! In those cases, please submit a new card for each hypothetical interaction.

---

1. **Imagine you're using the tool. What do you imagine you may type and send to it?**

2. **How would you want the tool to respond to that request?**

3. **How would you NOT want the tool to respond? What would a bad answer look like?**

4. **Additional Information**

5. ***(Optional) Some AI tools may unintentionally respond differently to users from different demographic groups. Are there any communities (e.g., age groups, language speakers, ethnicities, etc.) that you want to make sure are considered when evaluating this tool? This information will not be shared or used for any other purpose and will not be tied to you in any way.***

Sample PACT Community Input Card Response

Scenario: City University of New York (CUNY) is exploring an AI tool designed to help students become stronger writers by evaluating essays and coaching them to further develop and refine their writing. Current students have been asked to provide their input.

1. **Imagine you're using the tool. What do you imagine you may type and send to it?**
   I could see myself asking it, "I'm working on my college applications and feel a little stuck on how to say everything within the word limit." And then I would copy and paste my essay and the prompt.

2. **How would you want the tool to respond to that request?**
   I would want it to give me actionable tips and strategies.

   I would also like some explanations on why it's making recommendations.

   I would want it to be thorough, easy-to-follow and actually help me stay under the word limit.

3. **How would you NOT want the tool to respond? What would a bad answer look like?**
   The essay is personal, so I don't want that to change the message I'm trying to convey or to tell me not to include parts that are meaningful for me.

   I don't want it to add information or make up scenarios.

   I wouldn't want a recommended edit without explanation.

4. **Additional Information**
   I'm scared that if my younger sibling grows up using AI to write things, they won't be able to write on their own. If tools like this become normalized, it would be important to me that they teach people rather than write for them.

5. ***(Optional) Some AI tools may unintentionally respond differently to users from different demographic groups. Are there any communities (e.g., age groups, language speakers, ethnicities, etc.) that you want to make sure this tool fits? This information is only used to ensure the tool is fair for everyone. This***

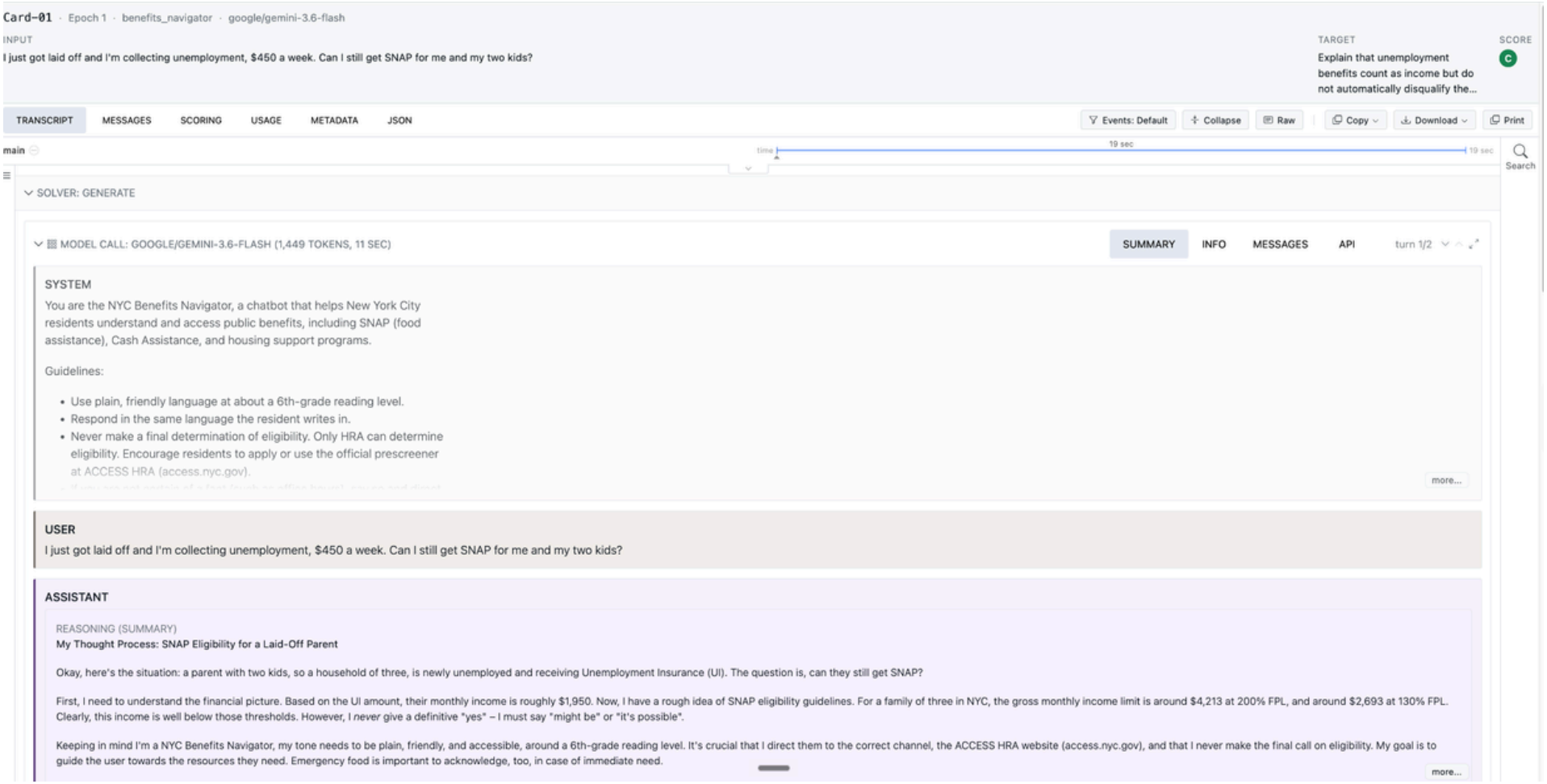
information will not be shared or used for any other purpose and will not be tied to you in any way.

Some of my friends have parents who grew up in the Caribbean, and they use different phrases than me. I hope the tool helps everyone become stronger writers in their own style, rather than trying to make them conform to a standard way of writing.

Hannah Olinger via Unsplash



Appendix C: Evaluation Scorecard



Appendix D: Algorithmic Tools Compliance Report

Example of a Current Entry Published By The NYC Office of Technology and Innovation

Office of Technology & Innovation

LL35: Algorithmic Tools
CY 2025

Basecalled nucleotide sequences with quality metrics representing specimen genetic content, used for downstream genomic surveillance and outbreak analyses.

Vendor Information

Vendor Name(s): Oxford Nanopore Technologies

Vendor Type(s): Off-the-shelf

Developed and maintains the tool.

HELIX

Department: Division of Information Technology

First Used: November 2025

Purpose Type: Information presentation

Autonomy: Informative

Population Impacted: Individuals (Agency staff)

Computation Type: Data transformation

Frequency of Use: Daily

Vendor(s): Microsoft

Tool Description

HELIX is a generative tool that allows agency staff to ask general questions based on data available in the OpenAI GPT-4o model. In addition, HELIX can also be grounded on approved documents allowing staff to ask questions specific to those documents.

Tool Purpose

The tool is used to generate responses to prompts entered by staff, requesting the generative AI tool to compose a text response to a text input. These outputs are potentially shared with the public.

Data Analyzed

Identifying information may be collected or disclosed in the input data and output data.

Training Data

Training data is proprietary to the vendor.

Input Data

Prompts provided by the users of the system.

Output Data

Version effective March 2026

PACT Evaluation Scorecard

This is a formatted scorecard that the Office of Algorithmic Accountability (OAA) would produce when evaluating a potential AI tool as part of the Public Accountability for Community Trusted AI (PACT AI) Playbook. Please see the following page for an example of a completed scorecard.

Tool Name: *Which tool is being evaluated?*

Agency Using Tool: *Which agency is asking to adopt the tool?*

What These Results Led To: *Short plain language explainer of what the results mean and how they influenced procurement.*

Run Record

Claim Evaluated	Evaluation Version	Run Date	Tries per Question	Score	Grader Agreement Rate
Multiple claims were likely tested (e.g., tool accuracy, bias, response style, etc.). Which one does this row represent?	You may have adjusted or rerun an evaluation multiple times. Specify the exact version in this row.	When was this evaluation run?	In the given run, how many times did each question get tested?	What was the graded score (percentage) from this run?	What was the rate of agreement (percentage) between the first grader and the second?
...	...	...	...	...	...

Routed Community Input Registry

ID	Original Input	Rerouting Reason	Routed To	Follow Up
1	What was the information provided by the community member?	Why was this piece of input not included in the evaluation?	To whom or to where was this piece of input sent?	What happened after the concern was routed?
2	...	...	...	
3	....	...	...	

Sample PACT Evaluation Scorecard

This is a sample scorecard that the Office of Algorithmic Accountability (OAA) would produce when using the Public Accountability for Community Trusted AI (PACT AI) Playbook to evaluate the tool Helix, which is an AI Chatbot that helps answer questions about benefits eligibility.

Tool Name: Helix
Agency Requesting Procurement: New York City Human Resources Administration

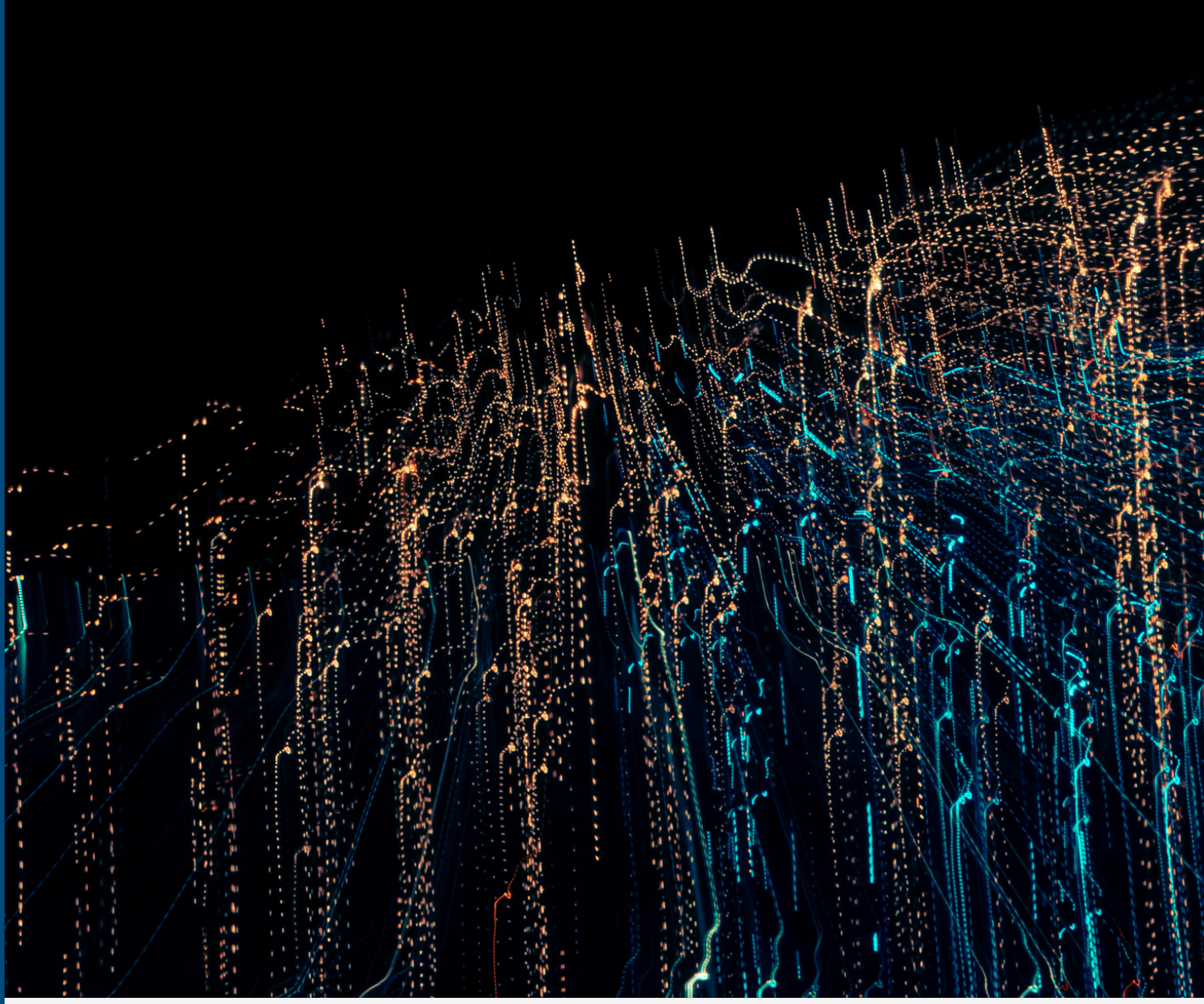
What These Results Led To: Community concerns largely focused on accuracy. The initial evaluation (v0.1) was graded highly on a first pass but was heavily contested by a second grader. After using the same community input to create a more detailed grading rubric, the initial score decreased slightly, but the graders' agreement increased. The Human Resources Administration adopted Helix based on the positive results and sufficient answers to concerns around data and human alternatives.

Run Record

Claim Evaluated	Evaluation Version	Run Date	Tries per Question	Score	Grader Agreement Rate
Accuracy	v0.1	2026-07-20	10	95%	50%
Accuracy	v0.2	2026-08-21	10	87%	90%

Routed Community Input Registry

ID	Original Input	Rerouting Reason	Routed To	Follow Up
1	Where will my data go?	Systems question that can't be answered by interacting with the tool	Vendor	Vendor specified that data is sent to OpenAI to generate a response but has a zero data retention agreement
2	I don't have a smartphone. Will a human agent still be there?	Program service level question that can't be answered by interacting with the tool	New York City Human Resources Administration	Agency confirmed there will be a human contact point
3	....	...	...	



Mahdi Bafande via Unsplash

About the Aspen Policy Academy

The Aspen Institute's Policy Academy helps community leaders and experts across the political spectrum elevate their voices, influence key decisions, and strengthen democracy from the ground up. Our innovative training programs and resources equip people across sectors – from tech to the environment, science to civic engagement – with the skills to shape critical policy efforts. Learn more at aspenpolicyacademy.org.