Review of Project Telemetry Data Collection and Usage

The following is meant to assist with a review of the project in connection with the project entity's Telemetry Data Collection and Usage Policy. Participants in the project are requested to provide responses to the following questions, regarding telemetry that is collected by the open source project and for use by the open source project community.

Project: MLflow

Completed by (name and email): Serena Ruan serena.ruan@databricks.com
Date: Jun 4, 2025 (last updated Dec. 17, 2025)

1. Specific data proposed to be collected

- Please fill in the following table with details on the specific data elements to be collected.

Data element *e.g., software version; operating system; etc.*	**Could be personal info?** (Yes/No)	**Could be tracking or unique identifier?** (Yes/No)	**Could be end-user / sensitive / business data?** (Yes/No)	**Notes**
Unique installation ID **(added 2025-11-11)**	No	Yes	No	A randomly generated ID, uniquely identifying the MLflow installation and attached to the telemetry usage events across sessions. The ID will be used only for telemetry purposes and with no other uses within an MLflow installation.
Unique session ID	No	Yes	No	A randomly generated, non-customer/non-personally identifiable UUID is created for each session—defined as each time MLflow is imported; a new session (and thus a new UUID) is generated if MLflow is reloaded or the REPL is restarted.
MLflow version	No	No	No	Version of MLflow in use, assuming users are using the public release with no customization (e.g. 2.22.0)

Python version	No	No	No	Version of Python in use (e.g. 3.10.16)
Operating System	No	No	No	The operating system on which MLflow is running (e.g. macOS-15.4.1-arm64-arm-64bit).
API name	No	No	No	Record the API name if those APIs are invoked (e.g. log_model, autolog, etc.). See below table for a full list of API names that'll be recorded
Metadata about GenAI functions usage	No	No	No	See below table for what metadata is logged
Backend store	No	No	No	Record the name of the backend store that's used (FileStore, SqlAlchemyStore, RestStore)
Component ID of interactive UI elements **(added 2025-12-06)**	No	No	No	Interactive UI elements (e.g. buttons, switches, form fields) in MLflow's frontend are currently tagged with a "component ID" (example). These are strings that indicate the component's function. A log record may be generated when users view or interact with these UI elements. These logs will also have some associated metadata. End users can turn off UI telemetry via a settings page in the UI, and a landing page banner will be implemented to notify users of the change.
Metadata about componentID interaction **(added 2025-12-06)**	No	Yes	No	See table below for full list

Full list of possible API names that will be logged:

log_assessment
log_expectation
log_feedback
trace
start_span
search_traces
MlflowV3SpanExporter.export
MlflowV2SpanExporter.export
InferenceTableSpanExporter.export
OtelSpanProcessor.on_end
autolog
tracing.enable
tracing.disable
set_active_model
clear_active_model
create_external_model
initialize_logged_model
set_logged_model_tags
log_model_params
mlflow.evaluate
mlflow.models.evaluate
mlflow.genai.evaluate
mlflow.genai.optimize_prompt
log_model
load_model
load_prompt
register_prompt
_is_signature_from_type_hint
spark_udf
mlflow gateway start

Metadata for APIs that are logged:

API name	Data element	Type	Possible values
log_model	flavor	Enumerated categorical value	catboost, diviner, dspy, h2o, johnsnowlabs, keras, langchain, lightgbm, llama_index, onnx, openai, paddle, pmdarima, promptflow, prophet, pyfunc, pytorch, sentence_transformers, sklearn, spacy, spark, statsmodels, tensorflow, transformers, xgboost
	model	Enumerated categorical value	string, PythonModel,

			ChatModel, ChatAgent, ResponsesAgent, object
	pip_requirements	Boolean	True, False
	extra_pip_requirements	Boolean	True, False
	code_paths	Boolean	True, False
	params	Boolean	True, False
	metadata	Boolean	True, False
	status	Enumerated categorical value	success, failure
autolog	flavor	Enumerated categorical value	anthropic, autogen, bedrock, crewai, dspy, gemini, groq, keras, langchain, lightgbm, litellm, llama_index, mistral, openai, paddle, pydantic_ai, pyspark.ml, pytorch, sklearn, smolagents, spark, statsmodels, tensorflow, transformers, xgboost
	disable	Boolean	True, False
	log_traces	Boolean	True, False
	log_models	Boolean	True, False
genai.evaluate Scorers	scorers	List of enumerated categorical value	Possible values for the list element: answer_correctness, answer_relevance, answer_similarity, faithfulness, relevance, custom_scorer
	predict_fn	Boolean	True, False
	status	Enumerated categorical value	success, failure

- If there is public documentation on the project site describing this data, please also provide a URL to that documentation:
    - **Updated 2025-11-11**: Documentation on MLflow telemetry is available at https://mlflow.org/docs/latest/community/usage-tracking/
    - **Original**: N/A at the moment. We will add a page for this prior to the release of the telemetry collection on mlflow.org website.

Metadata for UI events logged **(added 2025-12-06)**

Data element	Could be tracking or unique identifier?	Type	Possible values	Notes
Remote server status	No	Boolean	True, False	Whether the UI is served from a local server (e.g. localhost:5000), or a remote server (www.example.com)
Browser family	Yes	Enumerated categorical value	Safari, Firefox, Chrome, Other	What type of browser the user is viewing the UI from
Mobile status	No	Boolean	True, False	Whether or not the user is viewing the UI from a mobile or desktop device
Event type	No	Enumerated categorical value	onClick, onView, onValueChange, onSubmit	What type of interaction (click, view, etc) happened with the element identified by the associated component ID
Component type	No	Enumerated categorical value	button, alert, banner, modal, radio, input … (36 types)	What type of interactive component the interaction happened with. This list may extend in the future, but will always be a static categorical value
Component View ID	Yes	String	-	A random UUID associated with a component. This ID is regenerated whenever a component renders in the UI, and serves as a way to link view and click events
Timestamp	No	Date	-	The time at which the interaction occured

2. User notification and opt-in

- Please describe how users are *notified* (1) that telemetry will be collected; and (2) which specific data elements will be collected:
    - MLflow will publish clear documentation on its official website (mlflow.org) outlining:
        - How users can enable or disable telemetry collection
        - The specific data elements that are collected

- If there is public documentation on the project site or in the project source code with the particular notices, please also provide a URL:
    - The documentation will be published on MLflow's official website at mlflow.org
    - The source code for the telemetry instrumentation will be publicly available in the MLflow GitHub repository at https://github.com/mlflow/mlflow, excluding any security-sensitive configurations necessary for the secure collection and processing of telemetry data
- Is the telemetry only collected and shared if the user *voluntarily opts into* collection? (As opposed to, collecting data unless the user opts out.)
    - Telemetry collection will be enabled by default; however, users will have the option to disable it at any time—before or after MLflow is imported—by setting an environment variable. This configuration mechanism will be clearly documented in the official MLflow documentation.
- Is the user able to select between only sharing certain data elements, but not others?
    - No, at this time, all data described in the table above will be collected and transmitted as defined.
- How does notification and opt-in function if the software is installed and runs in a *fully-automated* installation (e.g., where there is no user who sees the notice and affirmatively clicks the "I consent" button)? Would telemetry data ever be collected in this type of scenario?
    - MLflow will automatically detect common automated environments (e.g., GitHub Actions, Jenkins) using standard environment variables and will disable telemetry collection in those contexts
    - Release notes and product documentation will inform users about the telemetry collection, and users will be able to opt out by following the guidance provided in the MLflow documentation.
- **(Added 16 Dec 2025):** For UI telemetry, data collection will respect the above environment variables if set on the server side. However, end-users still have the ability to independently opt out via a settings page in the MLflow UI. Additionally, a banner will be implemented on the MLflow UI landing page to inform users of the change.

3. Storage and use of collected data

- Please describe where the telemetry data is collected and stored (e.g., on which servers / repos; where they are physically located, if known):
    - Telemetry data is stored in a Unity Catalog (UC) table within a Databricks workspace owned and maintained by the MLflow open source maintainers. The physical location of the underlying infrastructure depends on the Databricks cloud environment in use. Ultimately, the data is stored in Amazon S3, in accordance with Databricks' standard security and compliance practices. Access to this workspace is restricted to authorized MLflow maintainers.
- Who administers and has access to the servers where data is stored?
    - Access to the telemetry data (the UC table) is limited to a small group of MLflow maintainers who are authorized and approved to administer the workspace. These individuals are responsible for maintaining the integrity and security of the data.
- Are all participants in the project community permitted to view and use the collected telemetry data? Or only particular participants / community members?
    - MLflow will publish dashboards on its official website mlflow.org to share aggregated insights derived from telemetry data, making the analysis results accessible to all community users. These dashboards will be updated on a regular cadence to ensure timely visibility into usage trends. However, the raw telemetry data itself will not be publicly available for direct access or querying.

4. Security mechanisms

- Is there a documented way that an organization could block the telemetry data from being collected from their systems, even if one of their employees inadvertently approves it?
    - Organizations may choose to block telemetry collection by restricting network access to the known endpoint used by MLflow to transmit telemetry data. However, MLflow does not plan to provide detailed firewall configuration instructions for the wide range of public and private cloud environments in which it may be deployed.
- Is there a reasonable possibility that including telemetry functionality opens up security vulnerabilities?
    - No. Data is transmitted using standard HTTP POST requests to a fixed, well-defined endpoint. The implementation adheres to security best practices to minimize risk, and no executable or untrusted code is received or run as part of telemetry collection.
- If so:
    - What steps are taken to mitigate this?
        - N/A
    - If a user does not opt into telemetry data collection, would this risk be fully mitigated?
        - N/A

5. Future changes

If the project plans to extend the scope of telemetry collection in the future (e.g. to begin collecting new types of data), or if the answers given above would change, please update this form and notify us so that we can quickly review the updated proposal.

2025-06-24 comments from LF review

The proposed telemetry data collection and usage is generally fine, and does not appear to raise the sorts of concerns described in the Telemetry Policy. We had just a couple of follow-up questions; please let us know your thoughts in a separate section below.

- In section 1, regarding the metadata about APIs / GenAI function usage that gets logged: It appears that this is set up to collect just the enumerated values for broadly-available AI tools. If a user were to use MLflow in coordination with their own internal, proprietary AI tooling, would any details (even e.g. its name) be included in the telemetry that gets sent back to the project?
- In section 3, it sounds like the raw telemetry data will be stored in a Databricks hosted environment, with access to raw data limited to a few specific MLflow maintainers.
    - Apologies as I'm not familiar with who the project maintainers are. Just to ask: will any non-Databricks employee maintainers have access? Or if all current maintainers are Databricks employees, then if and when there are additional external project maintainers in the future, can we ensure that they would also have equivalent access?
    - We just want to make sure that access to the telemetry data will be available to a broad set of project participants. Or at least, access to a copy of the data should be provided upon request by any project participants (even if not available for immediate download by anyone). From the Telemetry Policy, it states that any approved telemetry data collection "must make the collected data available to all participants in the project community."

2025-06-26 Responses from MLflow team

- Section 1: No, MLflow does not collect or transmit any identifying information about internal or proprietary AI tooling.
    - The telemetry system only logs usage of MLflow-tracked APIs. For features like genai.evaluate, the telemetry captures enumerated values corresponding to MLflow's built-in classes and tools. If users provide custom or internal tools (e.g., a proprietary scorer), the system replaces any such identifiers with a generalized value like 'custom_scorer', ensuring that no sensitive or specific internal tool names are included in telemetry data.
- Section 3
    - Yes, all maintainers of the MLflow GitHub repository—regardless of whether they are Databricks employees—will have access to the raw telemetry data. Currently, there is one maintainer who is not a Databricks employee, and he will have the same level of access. While the telemetry data is hosted in a Databricks-managed environment, it is stored in a standard Databricks account (not an internal account reserved for employees). Access is granted to the full maintainer group, and future maintainers from outside Databricks will receive the same level of access as internal maintainers.
    - To promote transparency and community participation, we will publish aggregated dashboards summarizing the telemetry data on the MLflow website, ensuring it is accessible to all project participants. If a participant requires a copy of the underlying data, they may submit a request to the MLflow team. Upon approval, we can export the requested data to a destination of the requester's choice (e.g., S3, Azure Data Lake Storage, etc.), provided they grant us the necessary write access.

2025-06-26 LF Projects approval

- LF Projects has approved MLflow's telemetry collection as described in this document.

2025-11-11 Additional data review request

- Received request to add unique installation ID to collected data set.
- Questions from LF Projects:
    - Will there be any information embedded in the installation ID that would be either (a) personal information; or (b) end-user / sensitive business data? Or would this be essentially a randomly generated ID / UUID?
    - Will this installation ID have any other uses in the MLflow installation, separately from being a tracking identifier for telemetry purposes?
        - To clarify what I'm asking here: For another project, Spinnaker, each installation of the software included a unique instance ID. The telemetry being collected did not include that unique instance ID itself; instead, the telemetry included a SHA256 hash of the unique instance ID. That way, the identifier in the collected telemetry wouldn't be directly usable to obtain the internal installation ID itself. Can you please clarify which of these approaches MLflow would be taking?

    - Will it be prominently disclosed to existing users of MLflow that this additional telemetry element will be included in new versions going forward?

    - Finally -- also regarding the documentation of MLflow telemetry generally: During the review earlier this year, for the question asking about public documentation describing the MLflow telemetry, the response stated: "N/A at the moment. We will add a page for this prior to the release of the telemetry collection on mlflow.org website." Do we have a URL for that documentation yet, so that we can update those details?

2025-11-11 Responses from MLflow team

- *> Will there be any information embedded in the installation ID that would be either (a) personal information; or (b) end-user / sensitive business data? Or would this be essentially a randomly generated ID / UUID?*
    - No, this will be purely a randomly generated ID.
- *> Will this installation ID have any other uses in the MLflow installation, separately from being a tracking identifier for telemetry purposes?*
    - This will only be used for telemetry purposes. The random hash like Spinnaker's approach works well for us.
- *> Will it be prominently disclosed to existing users of MLflow that this additional telemetry element will be included in new versions going forward?*
    - Yes, we will communicate this in a release announcement and will be documented clearly in the website.
- *> Do we have a URL for that documentation yet, so that we can update those details?*
    - Yes, here is the telemetry documentation on our website: https://mlflow.org/docs/latest/community/usage-tracking/

2025-11-12 LF Projects approval

- LF Projects has approved the addition of the installation ID data element to MLflow's telemetry collection as described in this document.

- The link to MLflow's telemetry documentation has also been updated in the writeup above.

2025-12-05 Additional data review request

- Received request to add data elements relating to UI to collected data set.

- Initial email on new request:

    - Currently in the MLflow UI, our interactive components (e.g. buttons, form fields, etc) have a "component ID" which identifies their function in the UI. These component IDs are generally static strings (or string interpolations containing static parts / enums), though an audit will be conducted to make sure no user-generated content appears within them. Essentially this would be the UI equivalent of logging "API name" in the currently approved version of the telemetry proposal. Other metadata may be associated with the logs, such as a UI equivalent of "Session ID" and "Installation ID".

- If it's permissible, we'd also like to collect some information about the user's environment, such as browser family (Chrome/Safari/Firefox), and whether or not the user is on mobile / desktop, but we understand if this could be considered user identifiable and would be happy to leave it out.
- Some other relevant details:
  - **Opt-out**: Since the UI is served from a self-hosted MLflow server, there are two different groups of users we need to concern ourselves with: (1) The user who hosts and manages the MLflow server (admins), and (2) Users of the UI (end users).
    - For (1), opt-out can be controlled via an environment variable set prior to the launch of the server via the mlflow server CLI command. If set, this will disable telemetry for all users who use the UI assets served from this server.
    - For (2), end-user level opt-out is currently not planned (i.e. end users will follow the admin's settings), but if need be we can provide a rudimentary solution like setting some LocalStorage variable on the browser to disable telemetry even if the admin has enabled it.
  - **Notification**: We will announce the collection of UI usage data in a similar manner that we announced the initial launch of telemetry (in release notes and our documentation page)
  - **Version**: We will not (and can't) retroactively enable this for old mlflow versions, so UI telemetry collection would only apply in the release where this is implemented and beyond.

2025-12-05 Comments from LF Projects

- In section 1, if you can add rows at the top of the table with the specific new Data elements that you are proposing to collect, that would be helpful. Please fill in the columns to clarify which types of sensitive data it could include and a brief description in the Notes column. Or if it's an update to an existing row in the table, feel free to leave a comment in the document to reflect that.
- After that, please take a look at sections 2-5 below and add comments if there are any differences to those questions from the original proposal.
- Once you've completed that, I'll take a look at the new specific details and can then circle back with any detailed feedback.
- I do have one preliminary question, as you're filling this out: I note the distinction between admin users vs. end users. In practice, would all end users be employees of a company with an admin who is controlling their installation? Or could end users ever include unrelated third parties—for example, would an admin ever administer the server on behalf of customers or other entities whose end users are individuals from a different business?

2025-12-06 Responses from MLflow team

- Thanks for getting back on this! I've added the new data element to the table in Section 1, and added a new table for some metadata that would be associated with this new data element. Please let me know if any further clarification is needed!
    - Comment in Section 1: We are in the process of formalizing enforcement that these component IDs be strictly static, but a manual audit has been conducted to ensure that no existing IDs contain user-generated data.
- Sections 2 - 5 will have no changes. To be specific:
    - Section 2 (notification and opt-in):
        - As before, the proposed UI telemetry will be opt-out (enabled by default), but we will respect the existing opt-out preferences of users, so there will be no additional steps they will have to take to disable UI telemetry.
        - The existing documentation page will be updated to reflect the new data element, and announcements will be made via our GitHub repo and release notes as before.
        - As described previously, there is a distinction between admins and end-users here. We currently only plan to implement a mechanism for telemetry disablement at the admin level, but can add an end-user level opt-out if necessary.
        - Source code for the UI telemetry mechanism will be fully open sourced and contained in the main MLflow repo at github.com/mlflow/mlflow
    - Section 3 (storage):
        - No changes, the new data element will reuse the existing data pipelines
    - Section 4 (security mechanisms):
        - No changes. As before, if an organization wants to block telemetry collection regardless of the opt-in / opt-out status, they can block traffic to the telemetry ingestion endpoint from their hosted server. The UI telemetry logs are relayed through the hosted server, and no cross-origin network calls are made.
- For the question about admins vs. end users, both of the scenarios described are possible. A company may host an MLflow server for its own employees. Additionally, a company may also take on a larger orchestration role, and run MLflow servers as a managed service for different companies (e.g. Amazon SageMaker does this). In this case, the end-users would indeed be individuals from a different business. Let me know if this creates additional concerns.

2025-12-11 Comments from LF Projects

- With the prior Python client usage, using environment variables as the mechanism for opting out of telemetry makes sense. If I understand correctly, the Python client user would presumably be able to set the environment variables in concert with when they are choosing to make use of the Python client (at least, assuming it's via a CLI environment they can control).
- However, with browser-based UI telemetry, the actual end user may not any longer have the ability to choose whether to enable or disable telemetry. This is part of what I was wondering about the admin vs. end-user questions. In the UI scenario, it sounds like the data being collected is about the end

user's own browser session and use of individual UI elements; but the end user does not in fact have any ability to opt out of telemetry from their device.

- I understand (and appreciate!) that the telemetry data is anonymized—and that is very helpful. But the individual end user should still have the ability at a minimum to say "I don't want any telemetry to be sent from my device." In light of current practices generally regarding cookies and other tracking mechanisms, I'm not sure this should be outside of the end user's control.

- Can you add a "telemetry on/off" toggle switch or similar settings option in the UI, to allow the end user to control whether or not their device sends telemetry? I'd recommend that it should be accessible to the user the first time they are visiting the UI, so that they have the ability to disable it promptly without hunting for it; and should be subsequently available in a settings pane or something similar. If that can be added, that would go a long way towards addressing the concerns here.

2025-12-12 Responses from MLflow team

- We can definitely add a toggle in a settings pane, and have a banner in the MLflow UI's landing page to notify users about telemetry collection. We can also update the documentation page to inform end users about how to opt out.

2025-12-17 Comments from LF Projects

- Assuming the opt-out for end users is implemented directly in the UI in a manner that is easy for them to access, then I am comfortable with giving the OK for the additional telemetry data collection as proposed here.