

○○×



Google Workspace

AI Classification

for Google Drive



Contents

○	Executive Summary	03
○	Introduction	04
	Use cases for data classification in Google Workspace	
○	Problem Statement	06
○	Solving the Data Classification Conundrum with AI	07
	AI classification for Google Drive	
	Understanding model performance	
	What's the model score?	
	Improving your model performance	
	Changes to labeled files in auto-apply scenarios	
	Privacy protection & model compliance	
○	Conclusion	15



Executive Summary

As an [industry leader](#) in data security and protection, we continuously enhance zero-trust and data protection capabilities in Google Workspace to meet customer needs. At the heart of our strategy is to provide customers with the ability to tailor data protection and governance policies to their specific requirements and security preferences. This customization is facilitated by our comprehensive data classification capabilities, which provide an integrated

toolset for identifying, classifying, managing, and securing sensitive information.

This whitepaper dives into Google's approach to data classification, which utilizes our advancements in artificial intelligence (AI) to address a persistent challenge faced by systems and security administrators: identifying and classifying data at scale with accuracy.

Disclaimer: This whitepaper applies to Google Cloud and Google Workspace products described at [cloud.google.com](#) and [workspace.google.com](#). The content contained therein is current as of July 2024 and represents the status quo as of the time it was written. References to forthcoming features are annotated as such and do not constitute a commitment to a specific release schedule. Google's security policies and systems may change going forward, as we continually improve protection for our customers. The availability of the product features and capabilities described in the paper are subject to license availability of various [Google Workspace editions](#) product offerings.

Introduction

Given the collaborative nature of how organizations operate today, the rate at which data is created, shared, replicated, and communicated is increasing exponentially. Unilateral access and management policies that don't consider file sensitivity or other content-related attributes can be too rigid when used on a large scale. This creates a challenge for organizations when it comes to data protection and management.

Classification labels, a data-classification solution within Workspace, provide an adaptable framework for identifying and categorizing information to enable granularity in data lifecycle management, data protection, auditing, and discoverability-and-search scenarios. With classification labels, organizations can specify up to 150 unique labels, each with flexible metadata structures and configurable permissions, through an administrator-defined taxonomy.



Use cases for **data classification** in Google Workspace

When classification labels are applied to files stored in Google Drive, they are recorded as metadata. This metadata is subsequently accessible by various Workspace systems, allowing the labels to act as triggers for specific policies. Practical examples of such use cases include:



Data Protection

Workspace data loss prevention (DLP) capabilities allow administrators to control the sharing of sensitive information with configurable content rules. These rules identify sensitive content and apply policy enforcements, such as restrictions on external sharing and download, copy, and print operations. DLP rules can use admin-defined or pretrained content detectors and classification labels as rule conditions to trigger policy enforcement. [Learn more.](#)



Auditing

For customers who analyze Drive log events in BigQuery, the events are enriched with classification label metadata. This allows administrators to monitor activities, such as file access, edits, and sharing of sensitive files, by filtering on specific label conditions. All label application, modification, and deletion events are logged to the Drive log event data, which also includes information about whether the action was performed by a user or system. [Learn more.](#)



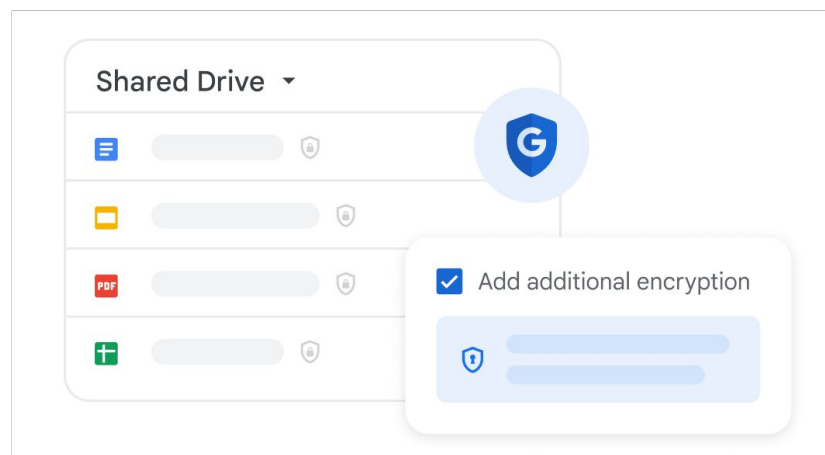
Search & Discoverability

Classification labels, which serve as file metadata, are accessible as an Advanced Search parameter when searching in Drive. End users can use label values to search for files in their My Drive folder, shared Drives that they can access, and files shared with and opened by them. [Learn more.](#)



Records Management

Google Vault supports custom retention rules for Drive based on classification labels. Admins can set file-level policies with configurable label conditions, such as "label is," "label is not," or "label date is before." Labels with multiple options are supported. [Learn more.](#)



Problem Statement

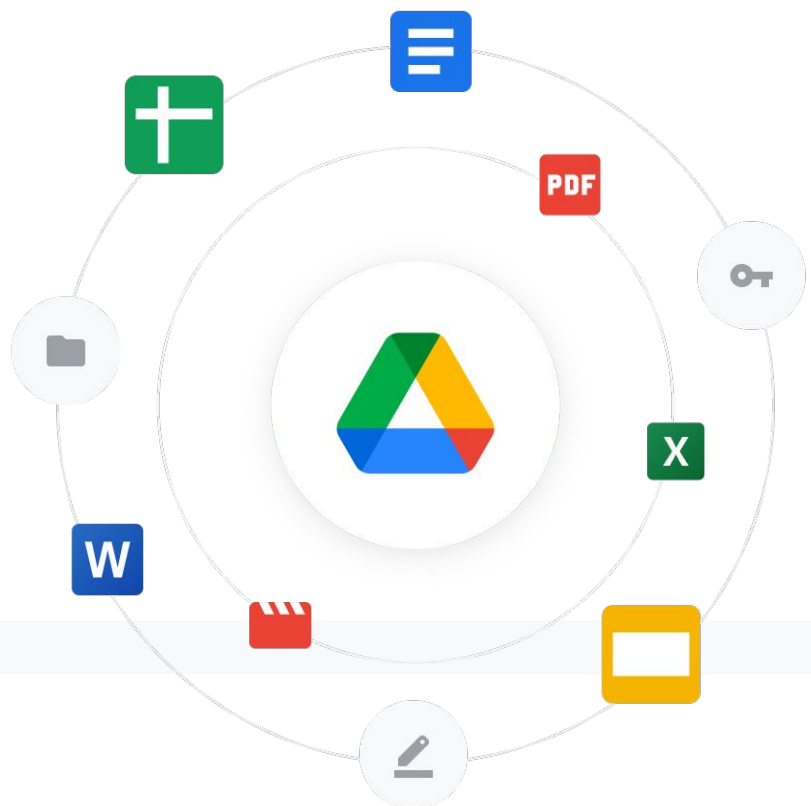
The effectiveness of leveraging data classification to achieve granular data management and protection is heavily influenced by an organization's ability to label its files at scale with precision. For many, this is an intractable problem. While the utility from file classification is high, achieving broad data classification coverage is usually a challenging and manual process.

Organizations that successfully classify a large percentage of their files employ various classification methods over protracted periods of time. These methods include manual labeling, [utilizing DLP rules to apply labels](#), programmatic labeling with the [Drive API](#), and the implementation of [default classification rules](#).

To address the challenge of achieving data classification at scale and precision, Workspace developed AI classification for Drive. AI classification enables customers to develop completely customized AI models – trained exclusively on their data – to automatically apply bespoke labels to files without the need for a developer to write a line of code.



Solving the Data Classification Conundrum with AI



AI classification for Google Drive

Overview

AI classification uses privacy-preserving, custom AI models to automatically identify, classify, and label an organization's sensitive content. After an initial training period during which the AI model learns an organization's criteria for the label, AI classification automatically applies labels to new and existing files in Drive. With models trained for specific data sets and automated application methods, organizations are able to achieve a high degree of precision and scale. The onboarding and model development process are as follows:

Phase 1

[Prepare for training](#)

Classification label creation: To get started, a classification label is created in the Label Manager, which the AI model will automatically apply to files once the initial training process is complete.

Training label creation: A separate training label that mirrors the aforementioned classification label is also created. The training label, as the name implies, is used for training purposes only; these are applied to files with characteristics that best represent the label classification for the model to learn from.

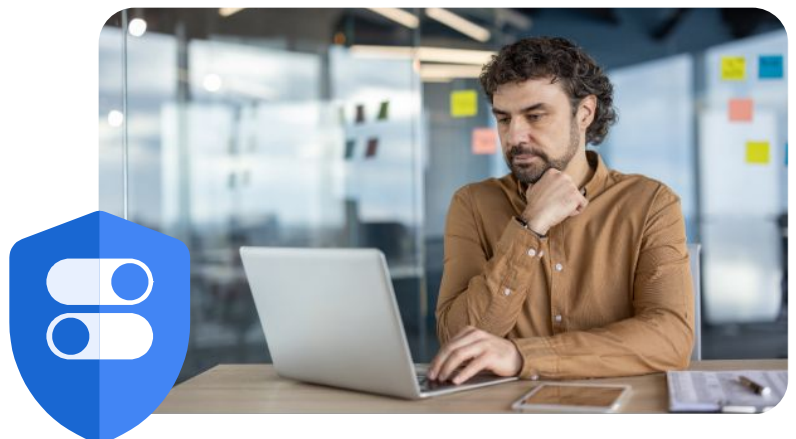
Designated labelers: Identify users in the organization that have a strong understanding of data loss prevention policies and can evaluate sensitive files. They will be part of the initial setup and ongoing process for model training and development.

Phase 2

[Train the model](#)

Files marked for training: Designated labelers in the organization begin classifying Drive files with the training label. The AI model will learn from files with the training label applied, providing a governance structure for model development. From there, the model begins learning how to similarly classify sensitive files.

Model development: Using only the files labeled for training, AI classification builds a model that can recognize patterns and characteristics within the files. This model essentially learns to identify specific features that distinguish different file categories.

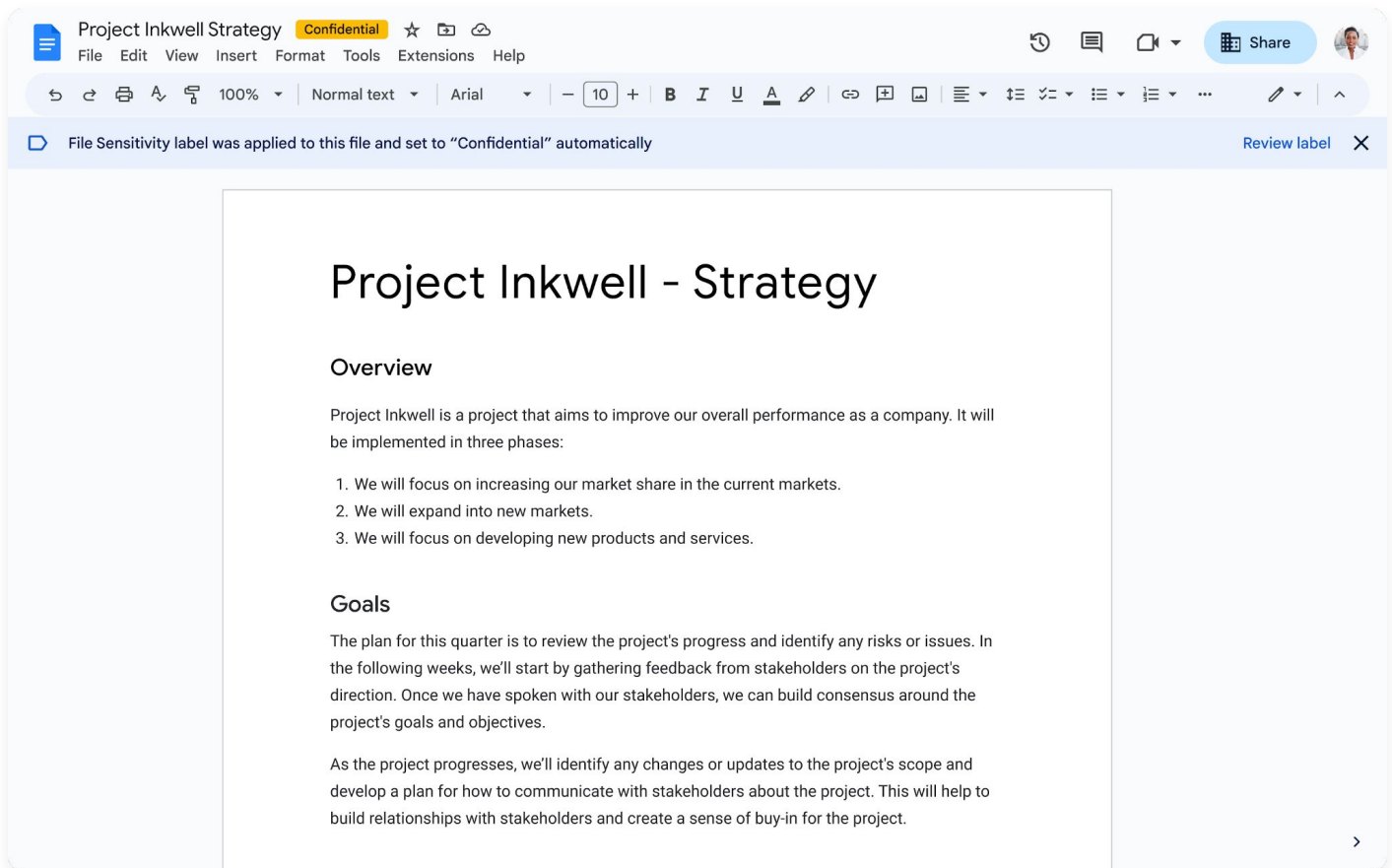


Phase 3

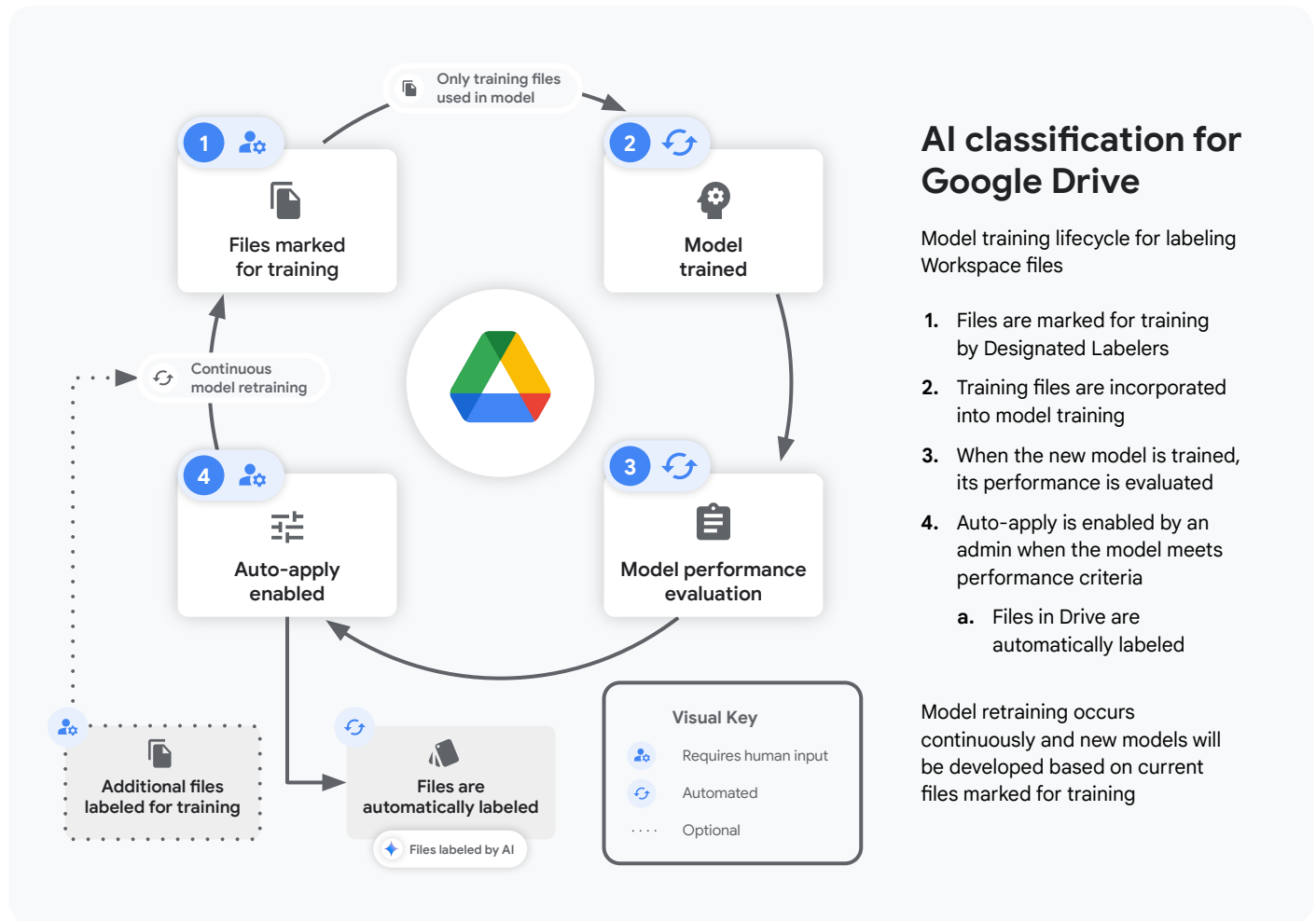
Turn on automatic classification

Auto-apply labels: Once the model is trained, the administrator is prompted to turn on auto-apply if the model attains an acceptable score. Auto-apply can be scoped to specific audiences within your organization, enabling granular control over whose files are evaluated for labeling. Once auto-apply is enabled, the model attempts to process all files owned by licensed users to see if they meet the model's criteria for labeling; newly created and edited files are continuously evaluated.

End user acceptance: When a label is automatically applied to a file, end users are prompted to review, then accept or modify the selection. User acceptance or modification of an AI-labeled file generates an audit event, and administrators can monitor how many files are classified, and how accurately, on an ongoing basis by reviewing the Drive audit logs.



An image of a Google Docs file that has a "Confidential" label applied to it automatically by AI classification. There is a banner under the toolbar prompting the user to review the suggested label.



A diagram of AI classification during the model training and development process.

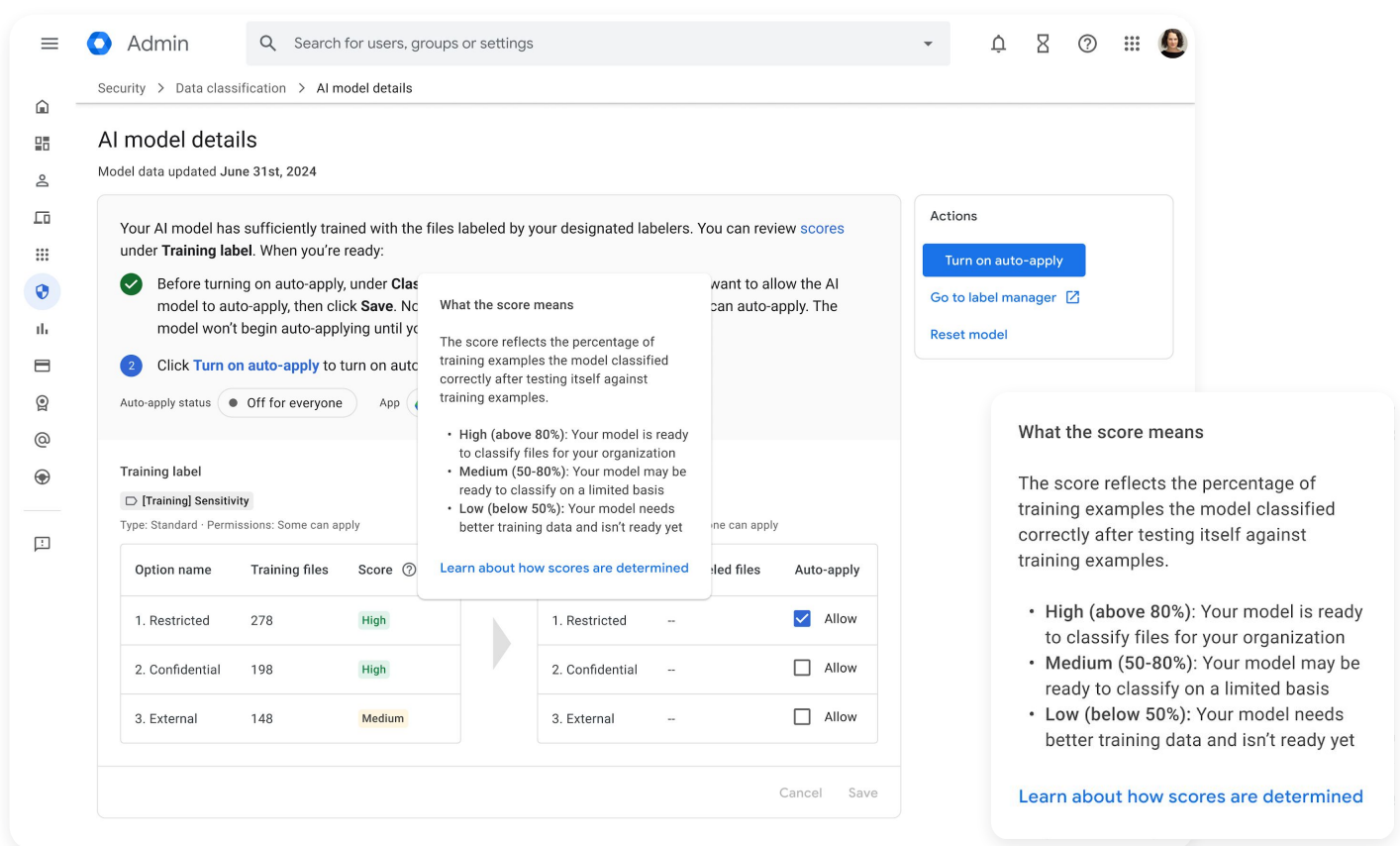
Understanding model performance

Workspace automates the model structure, training, and evaluation processes. The machine-learning model is trained on a data set composed of various file types. To ensure adequate representation of each category, a minimum of one hundred examples per class (file label) are required.

Following the training phase, the model's ability to categorize data is automatically evaluated by withholding 25% of the training data and testing the model's performance on this "hold-out" set. For example, if there are 100 files marked for the training process – the minimum required per class – 75 of those files were used for training the model; the remaining 25 files are used for evaluating the model. This rigorous development and evaluation process ensures the model can effectively generalize its knowledge and accurately classify new data.

What's the **model score**?

The training scores on the AI model details page reflect the recall of the model. The scores are calculated by comparing the model's predictions against a set of manually labeled files, which is the purpose for withholding 25% of the training data. This "withheld" testing data helps establish a baseline to determine how well the model generalizes its knowledge to unseen data.



The screenshot shows the 'AI model details' page in the Google Admin console. The page title is 'AI model details' and it indicates 'Model data updated June 31st, 2024'. The main content area contains instructions on how to turn on auto-apply for the AI model. A tooltip titled 'What the score means' is overlaid on the page, explaining the score levels: High (above 80%), Medium (50-80%), and Low (below 50%). The 'Auto-apply status' is currently 'Off for everyone'. Below this, there is a table showing training labels and their scores:

Option name	Training files	Score
1. Restricted	278	High
2. Confidential	198	High
3. External	148	Medium

Below the table, there is another table showing the 'Auto-apply' status for each label:

Option name	Auto-apply
1. Restricted	<input checked="" type="checkbox"/> Allow
2. Confidential	<input type="checkbox"/> Allow
3. External	<input type="checkbox"/> Allow

The 'Auto-apply' status is currently 'Off for everyone'. There are also 'Cancel' and 'Save' buttons at the bottom of the page.

A screenshot of the AI model details page with information on training scores highlighted in the Google admin console user interface.

High > 80%. This score indicates a well-balanced model with both high coverage (identifying most relevant files) and high precision (minimizing false positives).

Medium 50-80%. While the model performs adequately, there is room for improvement by addressing potential imbalances in the training data.

Low < 50%. This score suggests significant issues with the training data, potentially leading to inaccurate classifications.

Improving your model performance

While automation streamlines the technical aspects of model training, data quality remains critical to optimal performance. Diverse, high-quality data examples are paramount for training the model and its success. Below are some methods to accomplish this:



Data robustness

- To improve the model's overall performance and precision, our research suggests incorporating data from all suborganizations within the organization that you intend to enable auto-apply for. This approach ensures that the model utilizes a comprehensive and diverse dataset, leading to more robust and reliable results.
- To enhance the model's learning capabilities, ensure an ample supply of examples. Provide a significant number of diversified examples within each category, enabling the model to draw insights from a wide range of data points.
- For optimal results, prioritize documents with substantial content (at least 500 words) as this aids in the model's ability to identify patterns.



Balanced labeling

- Balanced data is when all classes are presented equally in terms of number of items per class. We recommend aiming for an equal number of files per label option, with a minimum of 100 examples per category. Maintaining a balanced distribution and representation of labeled files will improve the model's understanding of that data.



Feedback loop

- When an end user accepts or modifies an AI-applied label, their actions generate an audit log. The acceptance or modification of the label does not create an automatic feedback loop to the model. This is to protect against two risks:
 - The user accepting or modifying the label may not be well trained on the data classification policy, therefore risking the model learning from inaccurate feedback.
 - The user accepting or modifying the label could do so maliciously in an attempt to bypass label-based policies, creating an inaccurate feedback loop to the model.
- To create a feedback loop, “designated labelers” are encouraged to manually evaluate files that are auto-applied. If they identify examples of inaccurate label application, they can apply the correct training label to the file, which will incorporate it into the updated model-training data set in the next model iteration.

Following these best practices can significantly improve the quality and precision of the data classification model. The model refreshes itself on a regular basis, incorporating newly added or updated files with the training label applied by designated labelers into its training data to further enhance its performance and precision over time.



Changes to labeled files in **auto-apply scenarios** ⚙️

AI classification follows specific rules when determining how to update a previously labeled file:

- If the end user accepts or modifies the AI-applied label, the model will not change the verdict in the future, even if the file is modified.
- If the end user does not accept or modify an AI-applied label, the model will upgrade but not downgrade (following the label's configured option ordinality) the label option selection if a new verdict is determined.
- If the current model is reset (deleted) and a new model is configured to apply the same label, the new model will overwrite the past model's verdicts. This means the new model may have a prediction different from the previous and can upgrade or downgrade the label value. The new model, however, will not modify labels accepted or modified by an end-user.
- If the file is previously labeled manually by the end user, the model will not modify the label value (upgrade or downgrade).
 - In general, AI classification will defer to the judgment of file owners and editors who have applied labels to a document, and will not modify those values.
 - In cases where specific sensitive data is present in the file, data loss prevention (DLP) rules can be configured to override user-applied label values.

Privacy protection

& model compliance

AI classification in Drive aligns to [Google's AI Principles](#) and [privacy and compliance](#) standards. Google Workspace does not use customer data, prompts, or generated responses to train or improve our underlying cross-customer AI models. Each AI classification model is completely bespoke and unique to you as a customer. You identify which files it considers in training, and the model is only ever used in your domain.

Upon resetting an AI classification model, it is scheduled for deletion and unrecoverable. A new model can be trained using the same designated training files. Since AI classification is trained on actual files, the model undergoes automatic retraining based on the latest designated training files according to a predetermined schedule. During each retraining cycle, the previous model instance is then scheduled for deletion. Subsequently, if training-designated files are deleted, the model that incorporated them will also be deleted in the next re-training cycle, ensuring compliance with data lifecycle management policies.





Conclusion

Google Workspace provides a comprehensive suite of data protection tools designed to facilitate the identification, classification, and protection of sensitive data – AI classification being one of them. To be effective, data security practices require granularity, precision, and the ability to identify, classify, and label documents at scale, in addition to developing well-defined policies and fostering a security-conscious culture throughout the organization. This involves securing buy-in from business leaders and employees, promoting transparency through clearly-defined policies, and maintaining ongoing communication and collaboration between IT professionals, business stakeholders, and end users.

By adopting a holistic approach that combines advanced technology, data policies, and user education, organizations can effectively protect sensitive data, mitigate potential security risks, and meet regulatory compliance needs. Google Workspace is committed to empowering organizations in their data security journey, providing innovative solutions that adapt and evolve to meet the ever-changing needs of the modern workplace. To learn more, check out this [implementation article](#) and read [Roche's story](#) on using AI classification to enhance data security and protect millions of files in Drive.