Verkada

AI-powered search at Verkada

Search at Verkada

Building a hybrid-cloud foundation for AI-powered investigations

Introduction

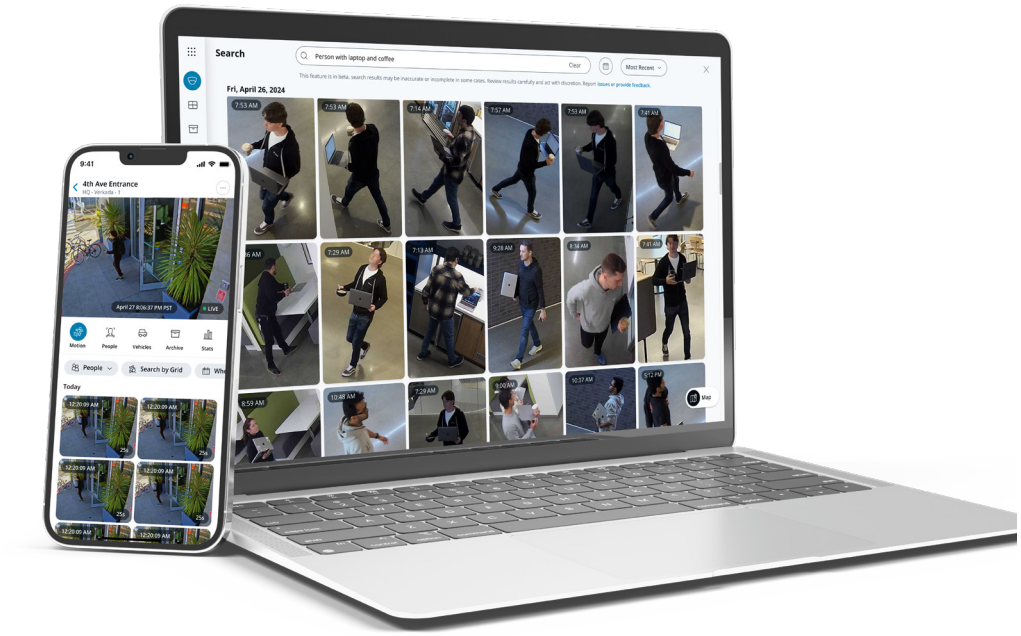
At Verkada, we often speak of our hybrid cloud approach to physical security and its benefits for storage, bandwidth utilization, processing, security, and data privacy. This foundational architectural decision — one where we store and process data both on our cameras and in backend data centers — enables us to offer advanced AI-powered analytics that otherwise wouldn't be possible. Our camera search functionality today — including AI-powered search for granular attributes of people and vehicles[1] — has come a long way since Verkada's inception.

In this whitepaper, we'll begin with an exploration of our hybrid cloud architecture and illustrate how processing both on cameras at the edge and at AWS' backend data centers has enabled us to evolve our computer vision (CV) analytics from basic person detection to today's advanced AI-powered analytics.

We will then dive deeper into AI-powered search, including our original choice of foundation model, advancements we've made since originally launching the feature, and how we've improved on open source foundation models in important ways to meet customer needs.

Like our edge-based CV capabilities, our AI-powered search functionality has its own set of limitations and challenges. We will discuss how building this feature on our hybrid cloud architecture positions us to take advantage of the latest advancements in AI and continuously improve its functionality.

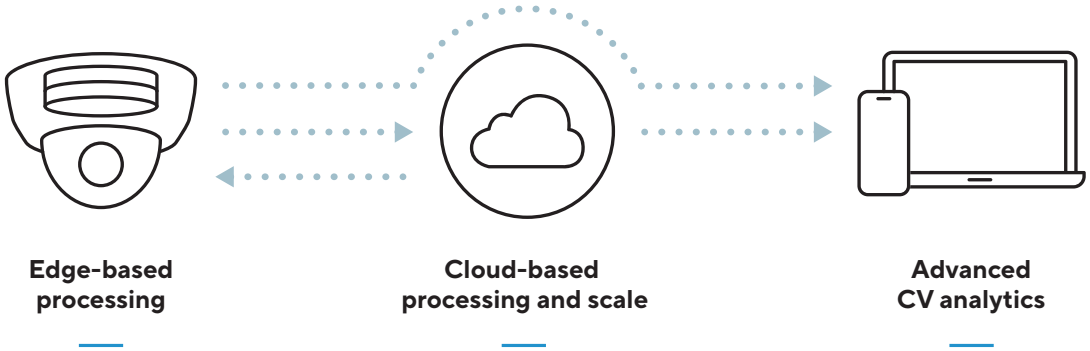
Finally, we will conclude with an overview of our moderation techniques and how we're tackling AI-powered search with social responsibility, respect, and anti-bias.



1. Check out our AI-Powered Search User Guide for example queries and methods for constructing your own queries relevant to your organization's requirements.

Hybrid-cloud architecture

A future-proof choice for AI



When we set out in 2016 to explore a new approach to physical security management, several architectural choices existed. A traditional, closed-circuit NVR/DVR-based system held the promises of airgapped security, but fell short on scale and distributed management. A pure cloud-based system offered the necessary scale and tools for remote management, but constantly streaming video to the cloud presented serious bandwidth issues that threatened to compromise video quality and retention.

We opted instead for the best of both worlds: a hybrid cloud system where we could store and process video both on the cameras themselves and over the cloud. We specifically chose AWS because of its notable data protection and privacy, security, and compute at scale.

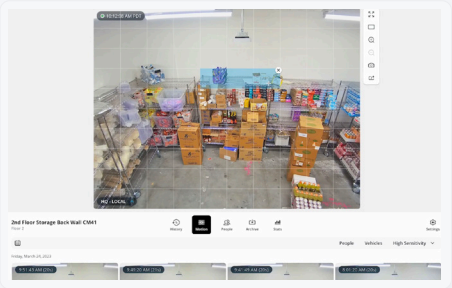
The benefits of a hybrid cloud security system are now well-understood and numerous security system vendors utilize this approach today. Our early decision to ground our systems in a hybrid cloud architecture has enabled us to innovate quickly and incrementally improve our analytics at scale — results that would have proved difficult or impossible otherwise.

Verkada cameras include on-device Ambarella processors and can leverage additional processing and computing power within the AWS cloud. As a result, our innovation has focused on two primary goals: delivering advanced edge-based analytics that optimize speed and privacy alongside powerful cloud-based analytics that take advantage of the tremendous computing power available in the cloud.

The timeline below depicts Verkada's initial years as we explored the limits of edge-based processing and backend computing power. Off the bat, our primary goal was to leverage hybrid cloud processing to help customers search for motion events — typically the most relevant types of events when investigating security incidents. During this time period, our analytics relied mainly on backend processing — limiting the power and effectiveness of video search and attribute detection.

Our early days

Exploring the limits of edge-based processing and backend compute

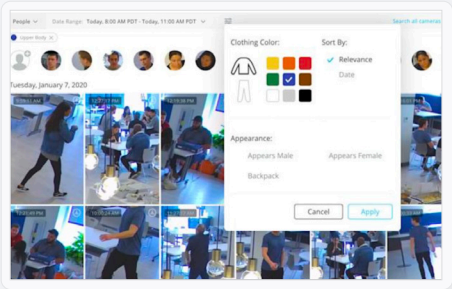


First iteration of motion detection in Command

2016-2017

Verkada's founding and the first iterations of motion detection in Command

- Like today, our cameras uploaded thumbnail images to AWS once every 20 seconds to optimize bandwidth consumption. Unlike today, however, where we support near real-time motion detection and trajectory tracing, our camera thumbnails were divided into a 10x10 grid (consisting of 100 sub-blocks). If we detected pixel changes moving from one sub-block to another, we would capture a thumbnail of that motion change (in a "motion thumbnail") and send it to AWS for additional processing. Limitations existed as "motion thumbnails" could only be sent every 10 seconds and only one "motion thumbnail" could be sent per 20 second interval between traditional camera thumbnails. Given that motion in a camera's field of view can occur multiple times within these 20 second intervals (i.e., a second object enters that camera's field of view during the period that the first "motion thumbnail" is undergoing processing), this functionality came with obvious limitations for our customers.



Third-party attribute search model

2018-2019

Integrating third-party APIs for CV analytics[2]

- On-edge object detection (single shot detector API), on-cloud attribute detection (Sighthound API), face search (AWS API)
  - » In addition to motion detection, our platform began to support more refined CV capabilities like people and vehicle attribute detection. To accomplish this, we utilized a series of third-party APIs and AWS models which primarily relied on backend processing, increasing latency.
- During this timeframe, we developed the concept of a "hyperzoom" (HZ), or a highly-detailed segment of a high-resolution image that contains only a person or vehicle. Instead of running the attribute model on an entire thumbnail of footage, we ran it instead on a HZ — reducing search latency and improving investigation accuracy and relevance. Our initial attribute model allowed users to search for specific HZs of people and vehicles using a predefined list of relevant characteristics. After the onboard Ambarella CV2x chip accelerated processing of HZs at the edge, the HZ was then sent to AWS' backend data centers which hosted our attribute model for scale, security, and data privacy — an example of how our hybrid cloud architecture began supporting advanced CV analytics.

2019

Improving CV analytics with third-party APIs

- Cross-camera person / vehicle attributes search (Sighthound API)
- Profile matching (AWS API)
- Person of interest (AWS API)

2. With edge-based processing still in its early stages, we leveraged third-party APIs that mainly relied on backend computing like single-shot detector for object detection and Sighthound for attribute search.

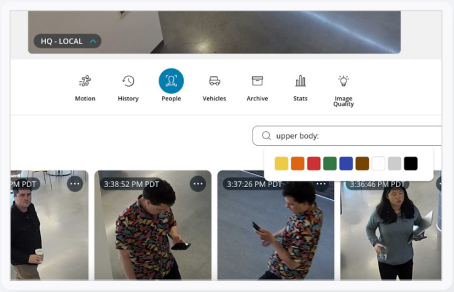
Realizing the benefits of hybrid cloud

Supporting a natural technological progression

Our early analytics provided customers with valuable insights. Conducting effective investigations, however, required near real-time, multi-object motion detection and detailed attribute search. With time, processors and backend hosting improved and the promises of hybrid compute became clearer. In tandem with these developments, we continued to invest in CV innovation internally – beginning to build attribute search, object detection, and motion models in-house, allowing us to migrate some processing from the backend to the cameras themselves.

As the timeline highlights below, we leveraged our hybrid cloud foundation to complement agile edge-based processing with scalable in-house models hosted on AWS to offer our customers new and improved features.

Establishing the modern elements for AI-powered search

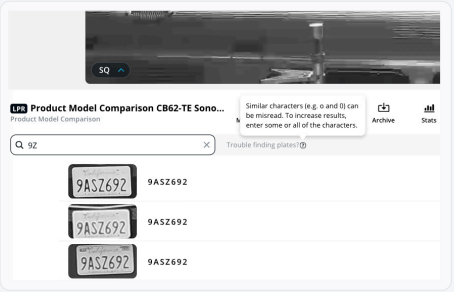


In-house person and vehicle attribute model

2021

Migrating CV models in-house

- Developed an in-house person and vehicle attribute model

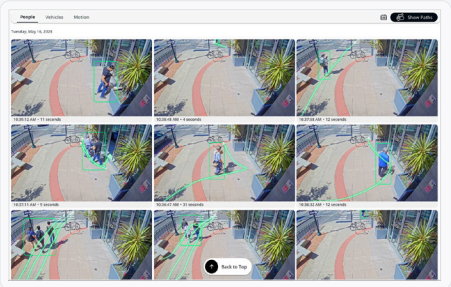


Tracking-based license plate recognition (LPR)

2022

Continuing to develop in-house CV models

- Developed our person of interest (POI) model, face search model, and tracking-based license plate recognition (LPR)

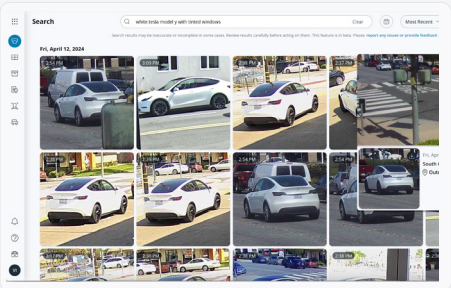


MotionX “tracklets”

2023

Improving in-house models and capabilities through added functionality

- Edge-based processing using Ambarella CV2x chips allowed us to break down video and analyze people and vehicles at a rate of 10 frames per second. From this, we could create “tracklets,” or associations of these subjects across space and time. These tracklets enabled us to extract the most informative HZ (i.e., high resolution, visible facial features, specific vehicle criteria) for more advanced analytical purposes.
  - » **MotionX:** With “tracklets,” we could recreate the entire trajectory of a person or vehicle within a camera’s field of view and trace motion of an event. MotionX trajectory analysis constitutes one of several CV features our hybrid cloud architecture makes possible.
  - » **Polygon-based motion alerts (line crossing, crowding, and loitering):** Leveraging our hybrid cloud architecture, we incrementally improved our motion detection to eliminate the multi-object blind spots customers encountered during our early years. With the CV2x, we’ve evolved from creating a “motion thumbnail” once every 10 seconds to identifying motion across 10 frames every second – opening up near real-time people and vehicle analytics at the edge. With advanced on-camera processing in place, we could now build new features on top of our in-house models including the ability to record and alert on line crossing, crowding, and loitering events.



AI-powered search

2024

AI-powered search

- The culmination of years of innovating on a hybrid-cloud architecture

The next logical step: how our hybrid cloud architecture supports AI-powered search in Command

We’ve now arrived at the next stage of this technological progression: incorporating the latest AI into our search capabilities. The combination of our Ambarella processors delivering powerful edge computing capabilities, our bandwidth-friendly upload capabilities, and the robust compute, scale, and security of AWS on the backend collectively give us the ability to leverage massive open source AI models. In addition, our team of experts coupled with the flexibility of the cloud enable us to consistently modify (and improve on) these models in-house. This approach should come as no surprise given it follows our historic product development cadence that we outlined above: improving on open source tools in-house by tailoring them to better meet our customers’ security and operational requirements.

Our original foundation model choice: one with top zero-shot performance on millions of HZs

Attribute search leverages hyperzooms (HZs) of people and vehicles and limits our users to a predefined list of search attributes, like clothing color or vehicle type. Now, with the advent of AI-based image analysis, we can run these HZs through an open source model for processing and pair them with freeform queries — expanding our users' search capabilities from a predefined list to near-limitless search criteria. Selecting the right foundation model to support these ambitions, however, required careful consideration.

We chose our model based on "zero-shot" performance, which refers to a model's ability to classify data it has never been directly trained on. This is crucial for our AI search functionality, as none of our customers' HZs are used to train the foundation model. As a result, we chose a model known for being one of the most efficient and accurate when it came to zero-shot performance, allowing us to pair one's natural language with the most likely corresponding set of HZs.

This model was trained using a massive training dataset and multi-modal training methods (leveraging image and text encoding), allowing it to outperform many other models in terms of accuracy and efficiency. The open source version of the model that we used as our foundation model was trained on over 2 billion text-image combinations and operated on a 2.5 billion parameter neural network, delivering better zero-shot results than other available open source models and previous iterations of the same model.

Importantly, this model also outperformed many other models on a variety of granular zero-shot performance factors like text recognition (e.g., "FedEx truck"), image context (e.g., "person walking down a stairwell"), and its ability to handle more complex query inputs than other models given its extensive pretraining on textual descriptions.

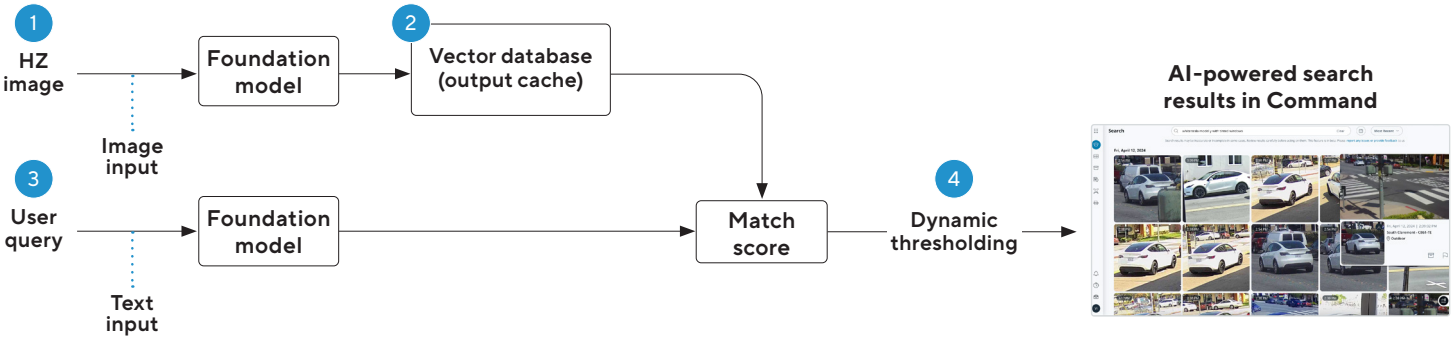
A continuation of our past: how we improved our foundation model in-house to directly address our customers' needs

It's evident why we chose our original foundation model, but we still had to adapt this foundation model's capabilities to the direct needs of our customers. In short, our foundation model is an image-to-text matching tool — not a search engine. We, in turn, created a search engineer out of natural language queries, with the foundation model as our "translator," probabilistically matching user queries to HZs.

This approach should come as no surprise — as with all of our CV capabilities, we utilize our hybrid cloud architecture to build and support added functionality that directly benefits our customers. Our foundation model was strong as an out-of-the-box matching tool, but to handle the millions of HZs from our customers on a daily basis, we needed to build on this foundation model in ways pertinent to their security and operational needs.

Each day, our systems ingest vast amounts of footage with millions of captured images. If we were to process these images through an out-of-the-box foundation model, the latency would prove too great for a seamless customer experience. Some queries would also display irrelevant or inappropriate results. To combat this, we employed a four-pronged approach to tailor the foundation model to our customers' numerous and varying requirements, as highlighted in the diagram below.

Our four in-house enhancements to the foundation model, which turned the foundation model into a low latency, scalable, and pertinent experience for our customers conducting investigations.



01 **We only run HZs through the foundation model**	Like with standard attribute searches, when a user inputs a query into our AI-powered search, we only run the HZs of people or vehicles through the foundation model, not the entire frame. With large amounts of footage that must be ingested and sent to AWS for additional backend processing, it's imperative that we minimize latency while maximizing the accuracy and relevance of search results.
02 **Cache of model outputs**	With the foundation model's several billion parameters, running each HZ through the foundation model to return the most relevant results requires massive compute and scale. While AWS supports the compute and scale requirements for this image retrieval process, it risks taking too long to return search results — compromising our customers' ability to run investigations quickly. To combat this, we cache model outputs for the most recent 10,000 HZs on AWS so we don't need to run each HZ through the foundation model each time a user enters an AI-powered search query. We utilize our cache to return the most relevant results, enabling far quicker retrieval of results at scale.
03 **Query matching: processing & parsing**	When a user enters a search query, we do not run that query verbatim through our foundation model or supporting cache. We will, for instance, add more specificity to a query to ensure that our model returns more relevant results. Similarly, we will append certain clarifiers like "a photo of," "a photo with," "an example of," or "a photo containing" to a query for greater matching accuracy as the underlying foundation model was trained on captions with these phrases (e.g., a query for "person holding ski gear" might be amended on the backend to read "a photo with person holding ski gear"). Importantly, we've also included parsing functionality that also allows our users to search by time and space, making it possible to search for results that span specific time periods, cameras, or Command sites.
04 **Dynamic thresholding**	When we compare a user's natural language query to the results that are returned by the foundation model, we strive for accuracy. For organizations to realize the value of AI-powered search, users must consistently receive the likeliest set of HZ matches. Creating the foundation model involved grouping like images together into descriptor buckets (e.g., putting all images of dogs into a "photos of dogs" category) and "penalizing" unlike images. This process worked on a scoring system where images that more closely matched the text descriptors received higher scores and those that didn't match received lower scores. At Verkada, we assess the distribution of scores for a given query's results and dynamically set a range of scores that correspond to relevant results (i.e., an accuracy range). Users will only see results that fall into this specific, high confidence range.

Limitations and results

Large language model performance is typically measured in terms of recall and precision.[3] For AI-powered search, we generally opt for models that have higher recall than precision.[4] As a result, users will see all likely results of their investigation (inclusive of true positives and some false positives), and can determine the accuracy of their search results themselves. If we opted for a model with higher precision, some results identified by the system as false negatives would not be displayed to the user, which could be more costly than a false positive for an investigation. As designed, we put the customer in control to filter out false positive results for themselves.

Changing foundation models

Because of our hybrid cloud architecture and the flexibility of backend processing, we're able to change our underlying model and algorithms should a different open source text-image pairing model present itself as more suitable for our customers. This is important because the landscape of AI foundation models is constantly evolving.

Our team of in-house engineers has continued to explore and evaluate new foundation models as they become available. In June 2025, after extensive research and testing on new models, the decision was made to pivot to a new foundation model with the goal of further optimizing both accuracy and efficiency.

Our new foundation model is approximately one fifth of the size of our original foundation model — relying on 400 million parameters instead of 2 billion. Despite this, our engineers found that this model was substantially more accurate than our original foundation model. With noticeable benefits in terms of both accuracy and computing efficiency, it was an obvious choice to switch foundation models.

Optimizing our new foundation model for AI-powered search

Foundation models are trained on data sets containing billions of images. Importantly, these data sets typically contain more "traditional" images in the sense that they are shot head-on with minimal distortion, similar to photos an individual may have on their smart phone or digital camera.

On the other hand, our cameras are often installed high on walls or ceilings to maximize coverage, capturing images of people and vehicles from unique vantage points. This means that certain foundation models may perform well on available data sets of "traditional" images, but struggle to accurately classify HZs from our cameras.

To account for this, our team of engineers have fine-tuned our new foundation model using millions of HZs from a variety of locations. This helps ensure that the foundation model can more accurately classify images captured from non-traditional angles where people, vehicles, or objects may appear slightly distorted.

3. In machine learning modeling, "precision" is the percentage of true positive results out of the entire batch of "positive" (inclusive of true and false positives) returned results. It's expressed formulaically as [Precision = True Positive Results / (True Positive Results + False Positive Results)]. It's a measure of the model's accuracy in identifying positive results. "Recall," on the other hand, measures the percentage of true positives that are correctly identified. It is expressed formulaically as [Recall = True Positive Results / (True Positive Results + False Negative Results)]. Whereas precision measures the overall rate of correct predictions, recall is helpful to identify the rate of false negatives.
4. Note that with Large Language Models (LLMs), recall and precision statistics will differ depending on the measured sample sizes. Our model's average recall is greater than its average precision.

Benefits of a more efficient foundation model

The influx of companies developing and leveraging artificial intelligence has led to a shortage of capable graphics processing units (GPUs), driving up prices and the costs associated with running a foundation model. To optimize cost and ensure scalability, we run our new foundation model in the AWS cloud, leveraging instances instead of GPUs.

Despite its smaller number of parameters, the new AI search foundation model is still too big for us to run on the edge (leveraging the computing power of individual cameras). However, we are getting closer to a point where this will be feasible.

For reference, our new AI-powered activity detection feature runs on the edge, leveraging the processing power of the Ambarella CV7x chips embedded on our latest generation of camera models. This foundation model has closer to 100 million parameters, making it approximately 1/4th of the size of the new foundation model we use for AI-powered search. With further foundation model improvements, our goal is to eventually land on an AI-powered search model small enough to be run on the edge.

Moderation

Unlike our filter-based attribute search, which has a predefined set of descriptors for narrowing search results, our AI-powered search lets customers search their footage for a wide range of attributes using their own words. Because our foundation model has been trained on publicly available data from the Internet, results may still be inaccurate, inappropriate or offensive. We have built query moderation into our platform to help reduce the risk that our AI-powered search is used maliciously or in ways that may be harmful. Moderation is a critical safeguard for this powerful feature.

At the same time, we implemented moderation in a way that isn't overly restrictive to the point that it compromises the feature's usability. We also leverage industry recognized practices, including open source data and moderation APIs, to make moderation more effective. Striking the right balance between respectful and useful searches is paramount for us.

Our approach to moderation

Just as we've chosen to use a publicly available model as the foundation for our AI-powered search feature, we developed our query moderation using publicly available practices for effective moderation. OpenAI, for instance, has published guides and blogs for developing suitable moderation techniques. We further augmented these capabilities with additional proprietary protections, which include our own list of prohibited search categories:

- Information about the race, ethnicity, or nationality of a person
- Information about the educational qualifications or religious beliefs of a person
- Subjective descriptions of people (e.g., attractive, ugly, wealthy)
- Inferred content from an image (e.g., "coolest person in the world")
- Sexual content or innuendo
- Names of specific people (i.e., public figures)
- Inappropriate or offensive descriptions of people or (e.g., "dumb person")
- Slurs equating people with animals (e.g., monkeys)

We also tapped into open source data to generate a list of banned text strings across all supported languages. When words or phrases on this list appear in a query, we reject the search and mask any results — a process known as string-based matching.

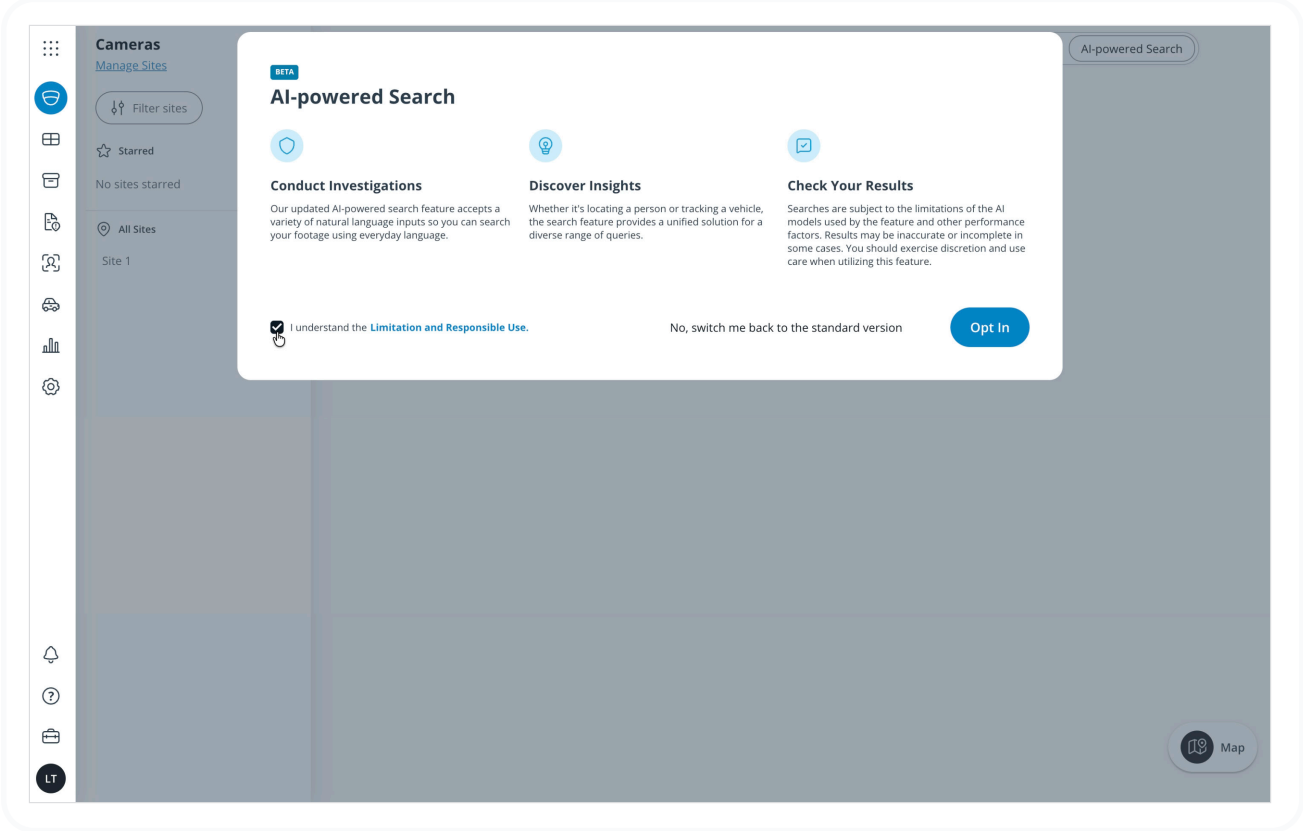
To help ensure our moderation model works as intended, we've held engineering "bug bashes" and analyzed thousands of our own engineering test queries (i.e., classified thousands of queries as "appropriate" or "inappropriate") to optimize our moderation rules for usability and to avoid bias. We can update our "blocklist" in a matter of minutes, enabling us to continuously improve our moderation techniques.

Combining these techniques yielded a moderation model that outperformed both Mistral-7B and ChatGPT 4 (two open source AI models with moderation measures) in our testing.

Moderation model testing (June 2024)

	Precision	Recall
Mistral-7B		
Inappropriate	31%	87%
Appropriate	98%	77%
ChatGPT 4		
Inappropriate	58%	67%
Appropriate	96%	94%
Our results: in-house finetuning of Mistral-7B		
Inappropriate	96%	79%
Appropriate	98%	99%

While our moderation model performs well, it is not perfect, and it is entirely possible that our users might see results that are inaccurate, incomplete, or inappropriate. We're committed to transparency and making these limitations clear to our users by including robust in-product disclosures that users must acknowledge and agree to before they can begin using our AI-powered search.



We currently offer built-in features for users to offer feedback on the quality of their search results. In addition, organization administrators can review AI-powered search queries in their audit logs to ensure that no users have attempted to input inappropriate search queries.

Conclusion

While our AI-powered search opens new possibilities for our customers, it's only the beginning. Our hybrid cloud architecture has, from the outset, enabled us to continue developing new CV capabilities and incorporate the latest technological advancements into our platform. Importantly, too, our hybrid cloud foundation affords us the flexibility to leverage a new foundation model should a more optimal one present itself, and also provides us with the tools to continuously implement new in-house enhancements. We look forward to offering our customers an easy, yet powerful and constantly-improving, search experience with privacy and respect top-of-mind.