



OFFENSIVE SECURITY

# State of Application Security

Our View from the Front Lines of Technical Assessments

# What's Inside

03 | Foreword

04 | Key Insights

05 | Section 1  
The Three Trends We're Seeing

16 | Section 2  
The Data: Slicing the Vulnerability Landscape

36 | Section 3  
Looking Ahead

37 | About the Firm

# Foreword

For 15 years, our practitioners have been breaking into applications for a living. They do so with permission and rigor, and for the specific purpose of helping the organizations we work with understand what adversaries can find before those adversaries find it.

In that time, we have tested thousands of applications across every architecture that has defined an era. We were finding vulnerabilities in monolithic web applications when that was the entire attack surface. We watched the shift to multiplying microservices and APIs creating attack surfaces that expanded faster than security programs could track them. We tested throughout the mobile explosion, when the application that lived in someone's pocket became the front door to some of the world's most sensitive systems. We documented the cloud revolution's security debt as it happened (and continues still). And now, we are assessing AI-infused applications at scale, meaning systems for which the model itself is the attack surface and the vulnerabilities do not yet have years of CVE history behind them.

We have done this work in the United States, Canada, the United Kingdom and across Europe. We have tested Fortune 100 institutions and early-stage companies. We have worked inside financial services, energy, healthcare and retail. These industries share little beyond being targets for attackers.

What has surprised us most over 15 years is not how much has changed but how much we have learned from our clients. The best engagements have never been purely transactional. They have been collaborations that included conversations about architectural choices, the trade-offs between deployment velocity and security, and what the findings actually meant for a business beyond just representing a scorecard. Our clients have taught us how technology evolves at the speed of commercial pressure. In turn, we provide them with insight into what that evolution looks like to an attacker.

This report shares what we have seen expressed in the data. Through five years of engagements, we have standardized how we classify findings, tracked remediation over time, and looked for patterns that hold across industries, architectures and geographies. What emerges is a picture of a security landscape that is genuinely improving in some dimensions and persistently struggling in others.

We are publishing now because the moment warrants it. The emergence of autonomous AI testing, the maturation of regulatory frameworks on both sides of the Atlantic, and the arrival of AI-infused applications as a mainstream attack surface all represent inflection points. The security industry needs data to navigate them.

So, it is time we put our data to work.



**Krish Raja**

Managing Director,  
Global Head of Offensive Security

# Key Insights

Kroll analyzed five years of penetration testing data. Of the trends that emerged, we focus on three in this report: the static application security testing (SAST)/software composition analysis (SCA) plateau, the need for more attention around authentication and authorization, and the security health divergence, which shows that regulation is not a reliable predictor of attack surface health.

A snapshot of the three trends is provided, followed by a detailed overview in Section 1, which also points to what these trends mean for security leaders. A more extensive look at the data behind these trends can be found in Section 2.

## The SAST/SCA Plateau

Modern SAST/SCA tools remain foundational security controls. Organizations that have not yet implemented them still have significant opportunities for risk reduction. However, among organizations already operating mature shift-left programs, designed to move testing earlier in the software development life cycle (SDLC), our data indicates that the security gains available through additional investment in the same tooling category are becoming limited. The industry's remaining risk is concentrating in places that automated scanning cannot easily reach at this time.

## Authentication and Authorization Deserve More Attention

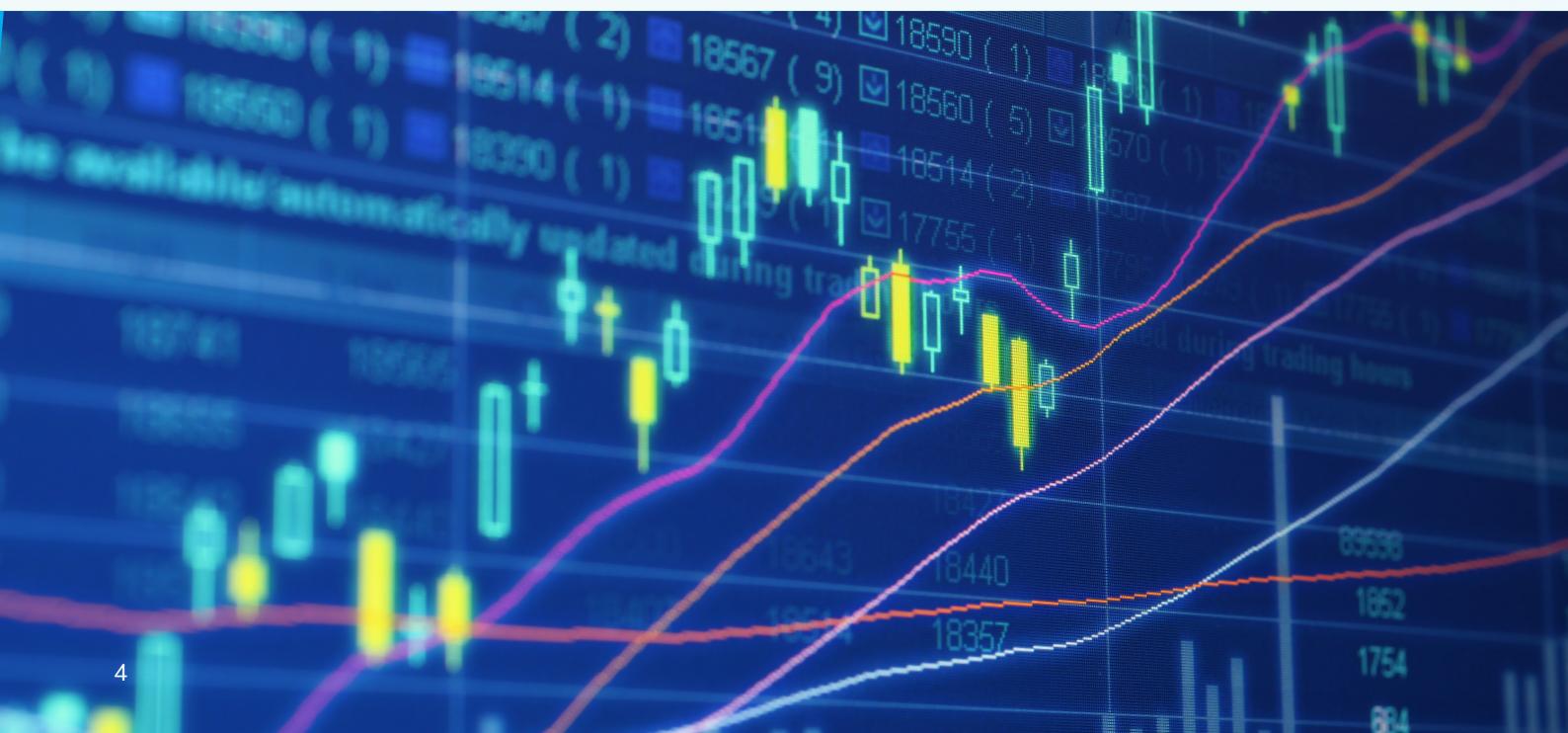
The most serious vulnerabilities in modern applications are not primarily injection flaws, dependency issues or configuration mistakes; they are authentication and authorization vulnerabilities. Broken Access Control issues alone account for more than 40% of all critical- and high-severity findings in the dataset. When this category is combined with the Identification and Authentication Failures category, these issues represent nearly 60% of the serious vulnerabilities that were observed. These issues are also the least detectable through static analysis and are disproportionately introduced during the design phase of software development.

**Nearly 60% of serious vulnerabilities were either Identification and Authentication Failures or Broken Access Control issues.**

## The Security Health Divergence: Regulation Doesn't Dictate Security

Over the observation period, some sectors demonstrated measurable improvements in vulnerability density, severity and overall attack surface health. Others remained stagnant or declined despite operating under similar regulatory pressures.

**One discovery stands out:** Regulation is not a reliable predictor of better outcomes. Two heavily regulated industries in our dataset followed markedly different security trajectories despite sharing comparable compliance obligations.



## SECTION 1

## The Three Trends We're Seeing

## TREND

## The SAST/SCA Plateau

Static analysis is essential. However, for programs that already use it, the returns on extending it further are diminishing (i.e., its reach has plateaued). The remaining attack surface needs to be improved with different methods.

Across this study's observation window, the health of application security has been slipping. This decline appears specifically in the classes of findings that static analysis is not typically expected to cover. If your organization already runs a shift-left program, additional SAST/SCA investment is not the lever that will further reduce your attack surface. The empirical case for that claim and guidance on the sources of residual risk are detailed in the Exposition section.

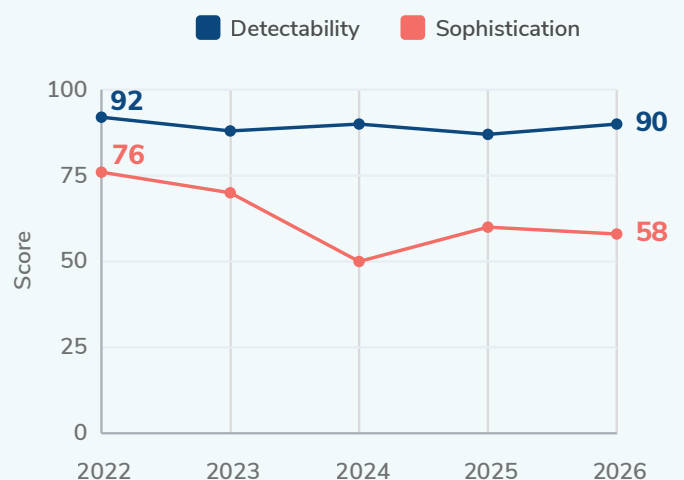
## What This Means for Security Leaders

Modern SAST/SCA tools do their job well and haven't plateaued themselves. The plateau separates marginal security investment that moves risk earlier in the SDLC from investment that does not. For programs that already run shift-left practices effectively, the analysis suggests focusing improvements on the residual attack surface, which has concentrated in design-stage authorization and authentication risks. Static tooling, by its nature, cannot reach these areas.

## EXPOSITION

Our composite Attack Surface Health Score (§2.9) is a 0–100 measurement of how a certain population of findings appears regarding two component maturities. First, **detectability** measures how easily SAST/SCA tools could detect the issue. A high detectability score lowers an Attack Surface Health Score because easily detectable findings have not been handled and remediated. Second, **sophistication**, which is weighted more heavily, measures how difficult it is for attackers to discover the issue and further exploit it. A higher sophistication score raises the Attack Surface Health Score because it is less likely that an attacker will have the ingenuity to find or exploit the issue. The Attack Surface Health Score has declined from around 81 in 2022 to around 67 in 2026.

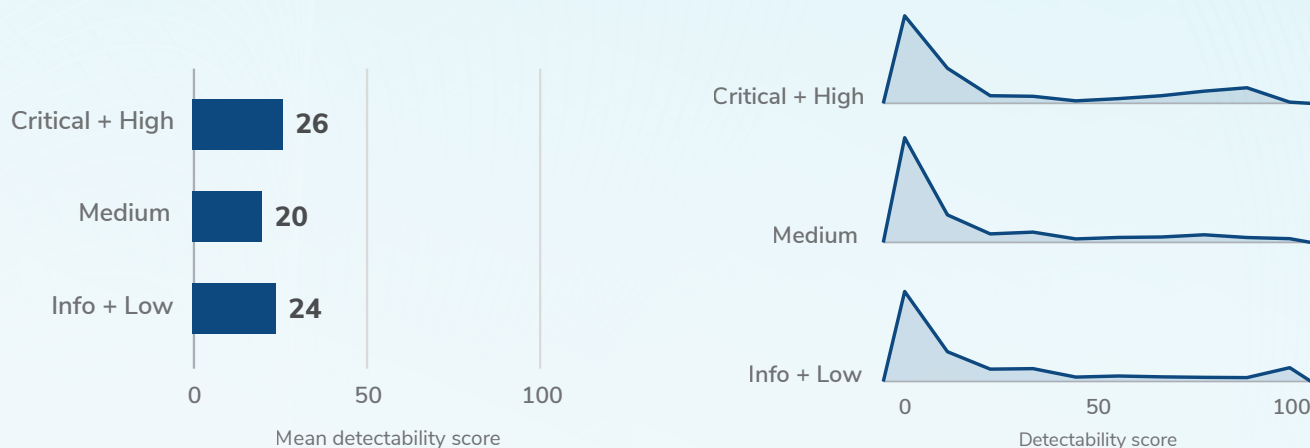
Over the five-year window, while detectability has broadly held stable, the sophistication score has decreased from 76 to 58. The Attack Surface Health Score drops as the sophistication score of found or exploited issues trends downward.



Detectability (blue) and sophistication (orange), the two components of the Attack Surface Health Score, across 2022–2026.

**Vulnerabilities are becoming easier for attackers to exploit.**

**Severity and detectability are essentially independent (§2.8.1).** The seriousness of a finding tells you almost nothing about whether SAST would catch it. Across the three severity groupings, the mean detectability score sits within a six-point band, and the underlying distributions broadly follow the same shape. A well-tuned static tool does not prioritize catching the findings that matter most for risk; it catches those with a specific code structure and format, regardless of a finding's exploitation impact.

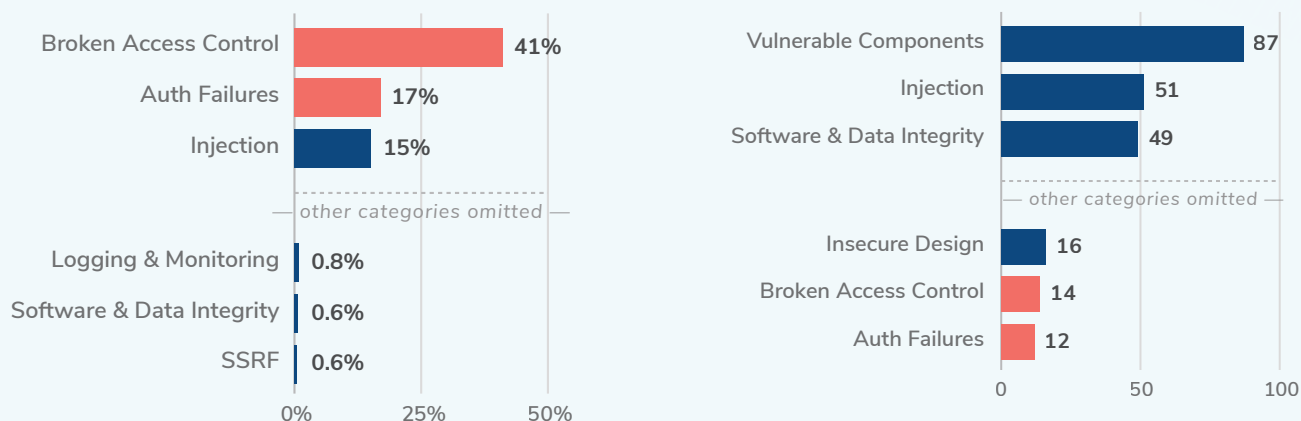


The mean detectability score for findings of three severity ranges (Critical and High, Medium, and Info and Low) are similar. The severity of a finding does not indicate how likely it is to be detected by SAST tools.

In each severity group, the number of findings (represented on the y-axis) with low detectability scores in comparison with those with high scores is very similar.

**A well-tuned static testing tool does not prioritize identifying vulnerabilities that matter most for risk.**

**The categories that include the highest concentration of serious findings are the ones static analysis reaches least well (§2.4.2, §2.8.2).** The Broken Access Control and Identification and Authentication Failures finding categories (together involving 58% of all critical- and high-severity findings) score at the bottom of the detectability ranking, with mean detectability scores of 14 and 12, respectively.

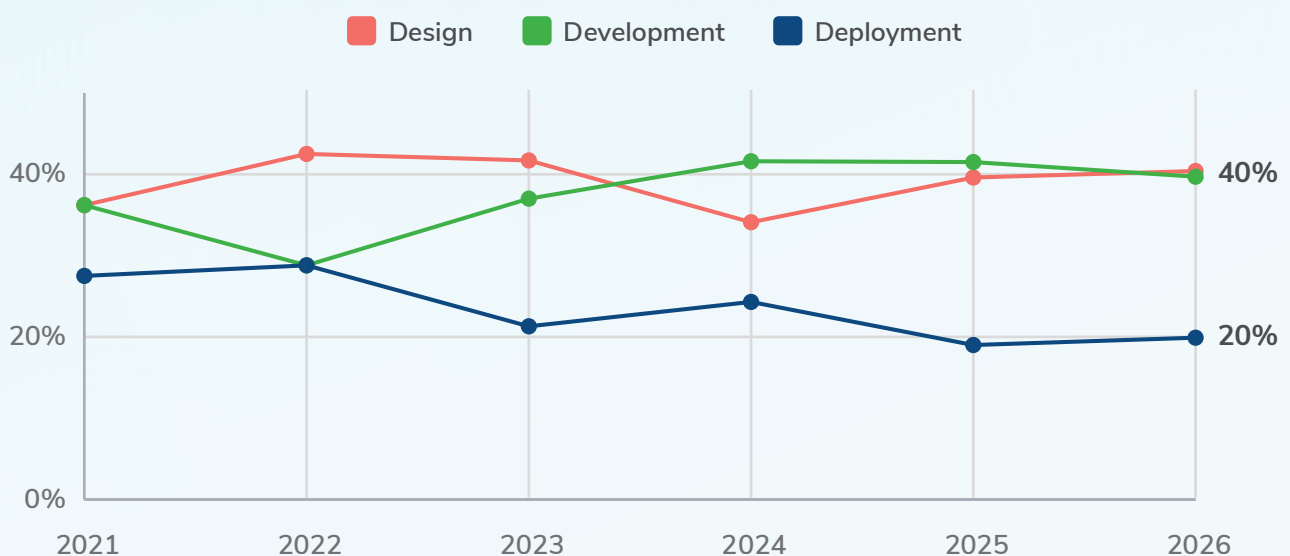


The share of all critical- and high-severity findings by Open Worldwide Application Security Project (OWASP) 2021 category. The top three highest and the bottom three lowest are shown, with the remaining four categories omitted for brevity. Broken Access Control alone accounts for over 40% of the higher-severity band. See §2.4.2 for the complete ranking.

The mean SAST/SCA detectability by OWASP 2021 category, with all severities pooled. The top three highest and bottom three lowest are shown, with the remaining four categories omitted for brevity. Note that the Broken Access Control and Identification and Authentication Failures categories sit at the bottom, regardless of finding severity. See §2.8.2 for the complete ranking.

The serious risk remaining during or after remediation programs (i.e., the residual risk) has been trending upstream in the SDLC. The share of critical- and high-severity findings introduced at the **deployment** stage, a common target for misconfiguration-catching automation, has fallen from around 28% in 2021 to around 20% in 2026 (§2.6.4). The shares of findings introduced at the **design** and **development** stages have in turn grown, accounting for roughly 40% each at the end of the study window.

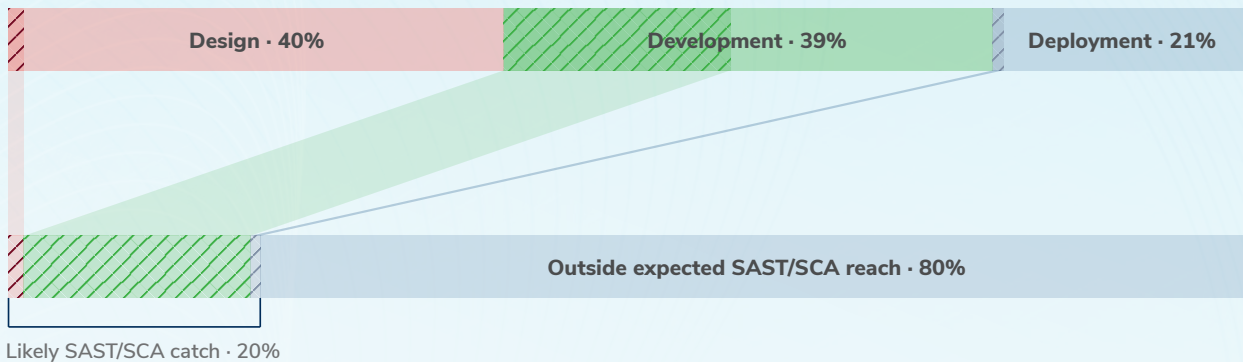
**The design and development stages are seeing more critical- and high-severity vulnerabilities, up to 40% in 2026. Fewer vulnerabilities are making it to the deployment stage, down from 28% in 2021 to 20% in 2026.**



The design and development stages are seeing more critical- and high-severity vulnerabilities, up to 40% in 2026. Fewer vulnerabilities are making it to the deployment stage, down from 28% in 2021 to 20% in 2026.

The rise in findings introduced at the development stage, at first glance, appears to provide a partial explanation of the SAST/SCA plateau. You might assume that static analysis catches development-stage bugs. This is partially true. About **1 in 5** development-stage critical- and high-severity findings score high enough with our detectability metrics that SAST would easily catch them. Another **1 in 4** score high enough that SAST would likely catch them. But that SAST reach is confined to the development stage. For design- and deployment-stage findings, only about 3%–4% score high enough in detectability for SAST to catch them.

### Critical + High findings by introduction stage

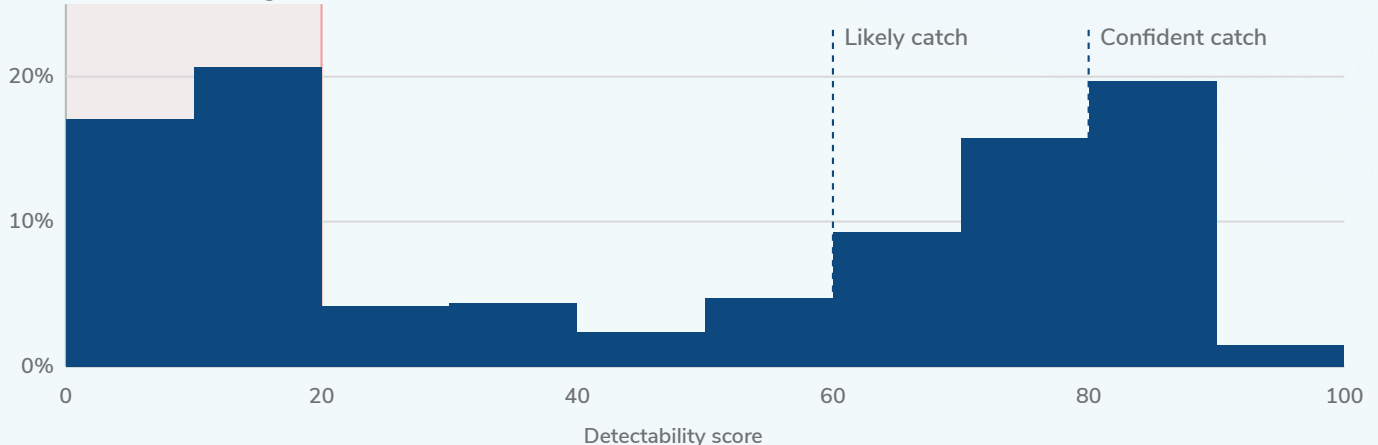


### Same findings, split by SAST/SCA reachability

The full critical- and high-severity finding population is viewed in two ways. In the top bar, the findings are split by the SDLC stage in which they are introduced, with the hatched subregion within each segment marking the portion that scores in the range that SAST/SCA is likely to catch (detectability  $\geq 60$ ). In the bottom bar, the three hatched subregions are aggregated. Even at the generous likely-catch threshold, static analysis is expected to reach only about 1 in 5 serious findings, with the majority coming from development-stage findings. Under the stricter confident-catch threshold (detectability  $\geq 80$ ), the reachable share shrinks to about 1 in 10.

The goal of highlighting the SAST/SCA plateau is to point toward what static code scanners don't catch. About **38%** of development-stage critical- and high-severity findings have scores at the low end of the detectability scale (i.e., findings that the industry does not typically expect SAST/SCA tooling to detect). These findings exist in our dataset regardless of whether the client that was affected by them ran SAST/SCA. Although the industry does not expect static analysis to address authorization logic, authentication flows or business logic assumptions, critical- and high-severity findings are those that matter most for risk. The SAST/SCA plateau does not point to a limit on tool quality. Instead, it details a limit on the kinds of findings these tools are typically expected to reach.

### Outside expected SAST/SCA reach 38% of findings



The distribution of SAST/SCA detectability scores for critical- and high-severity findings introduced at the development stage. The shaded left region (marking findings scoring below 20) is the range in which the industry does not typically expect SAST/SCA tooling to detect findings. This region accounts for about 38% of serious findings introduced in the development stage. Dashed lines mark the two SAST-reach thresholds. Findings at 60 or above are likely caught by SAST/SCA (about 46%), and findings at 80 or above are confidently caught (about 21%).

**The SAST/SCA plateau does not point to a limit on tool quality. Instead, it details a limit on the kinds of findings these tools are typically expected to reach.**

For an organization that has not yet implemented a shift-left program, the case for SAST/SCA investment remains straightforward. The tooling catches a real and useful portion of the finding population, and the coverage is essential. For an organization that already has SAST/SCA in place and is asking what its next security investment should target, the empirical answer offered by this dataset is that additional investment in the same class of tooling is unlikely to move the needle. It would duplicate existing coverage. **The residual risk sits in categories the tool cannot reach.**

## TREND

### Authentication and Authorization Deserve More Attention

When the industry conversation turns to serious application security threats, you likely hear the terms that dominate breach disclosures and vendor pitches (e.g., remote code execution, data leaks and ransomware). However, the results of thousands of penetration tests point elsewhere. Authorization- and authentication-class failures are consistently the most serious, most persistent and least detectable findings in the dataset. Yet they are rarely the finding classes that the conversation is centered on. The empirical case is detailed in the Exposition section.

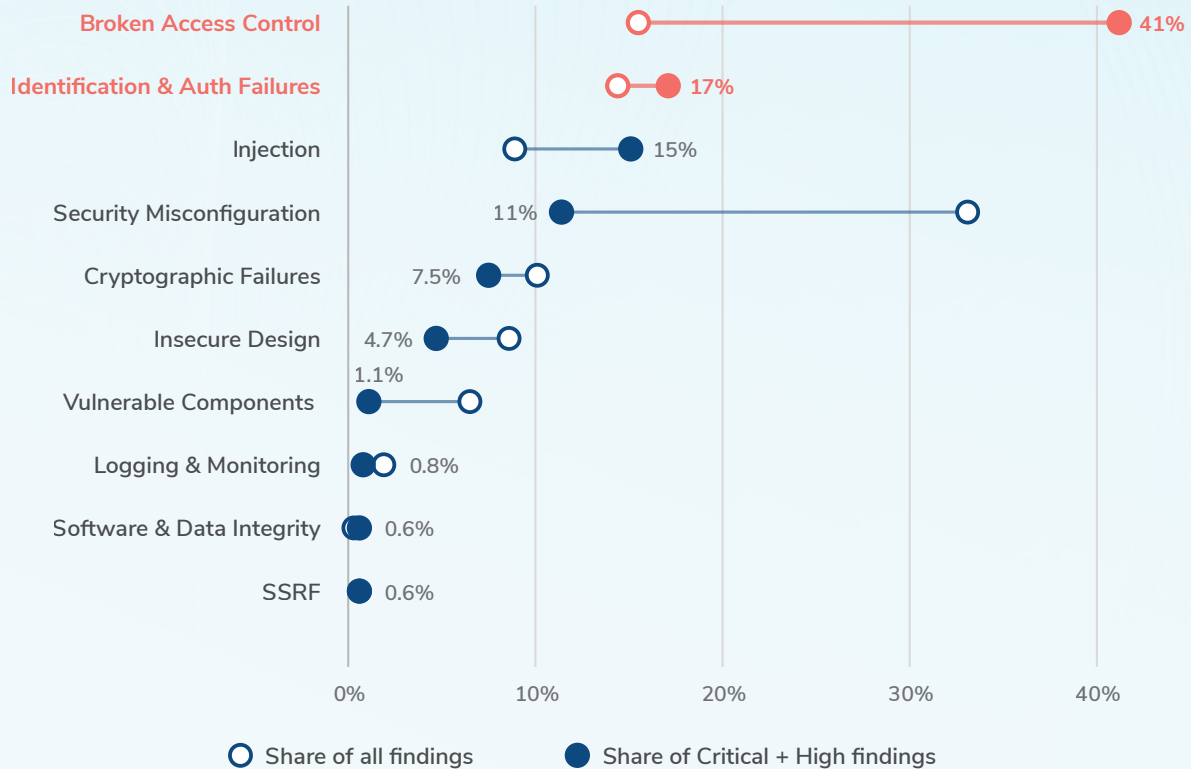
## What This Means for Security Leaders

When the critical- and high-severity findings of a penetration test are authorization and authentication issues, the answer isn't more SAST. The risk is related to how the target is designed in terms of who is allowed to do what and how they prove who they are. This insight is empirically consistent across taxonomies, severity thresholds and the years that the dataset covers, and it deserves proportional weight in any application security risk conversation.

## EXPOSITION

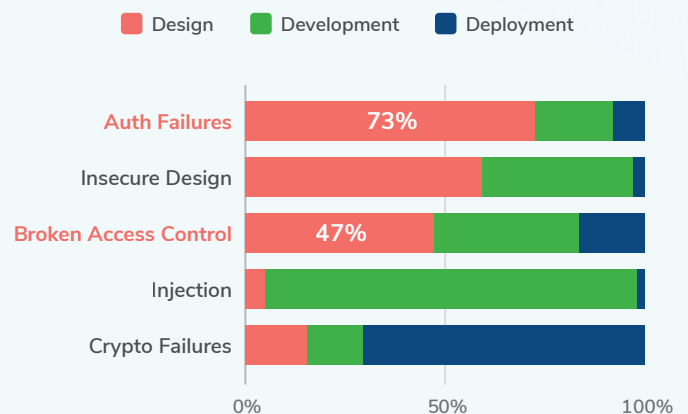
Broken Access Control findings alone account for **41%** of all critical- and high-severity findings, which is **2.6 times** the category's share of the overall finding population. If the Identification and Authentication Failures category is added, the two categories together account for **58%** of the serious findings band while representing only about 30% of findings overall. This overrepresentation holds across all three taxonomies used in this report (§2.3.2, §2.4.2, §2.5).

Of all critical- and high-severity findings, **41%** are Broken Access Control vulnerabilities, and **17%** are Identification and Authentication Failures.



The volume share (hollow dot) and the critical- and high-severity share (filled dot) are organized by OWASP 2021 category. The rows are sorted by severity share in descending order. When the filled dot sits to the right of the hollow dot, the category is overrepresented in the serious findings band relative to its overall volume; when it sits to the left, it is underrepresented. With 15% of all findings and 41% of the serious findings band, the disproportion of Broken Access Control is the largest in either direction.

Both categories of findings are also predominantly introduced at the design stage of the SDLC: 73% of Identification and Authentication Failures and 47% of Broken Access Control findings are traced to design decisions rather than coding or deployment errors (§2.6.2). Authorization and authentication design requires anticipating who users are, what they're allowed to do, and how those permissions evolve. This kind of planning can get overlooked during development, so it is more difficult to re-architect when requirements change mid-SDLC. And design-origin findings rise as issue severity climbs (§2.6.1), from 22% of informational findings to 45% of critical findings.



Vulnerability Introduction Point mix across selected OWASP 2021 categories, sorted by design share. See §2.6.2 for the full 10-category chart.

**Design decisions, rather than coding or deployment errors, account for 73% of Identification and Authentication Failures and 47% of Broken Access Control findings**

Of the five Common Weakness Enumeration (CWE) categories that most persistently recur among critical- and high-severity findings (i.e., those appearing each year at a meaningful volume), four involve authorization or authentication: Authorization Bypass Through User-Controlled Key, Missing Authorization, Missing Authentication for Critical Function and Improper Privilege Management (§2.10.1).

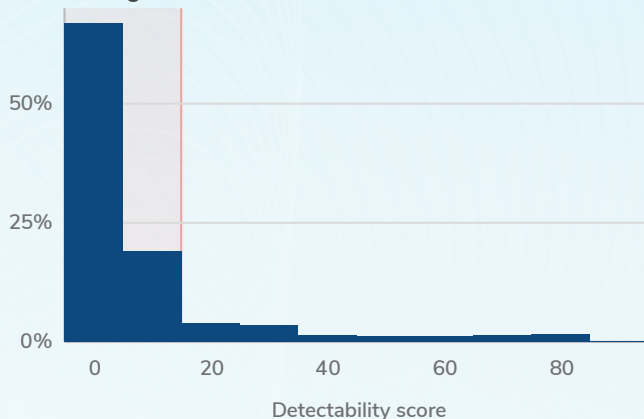
| Top persistent critical and high CWEs                      | Category       | Average annual weighted occurrence of critical-and high- severity finding |
|------------------------------------------------------------|----------------|---------------------------------------------------------------------------|
| CWE-639 · Authorization Bypass Through User-Controlled Key | Authorization  | 10.7%                                                                     |
| CWE-79 · Cross-site Scripting                              | Injection      | 8.4%                                                                      |
| CWE-862 · Missing Authorization                            | Authorization  | 6.9%                                                                      |
| CWE-306 · Missing Authentication for Critical Function     | Authentication | 6.5%                                                                      |
| CWE-269 · Improper Privilege Management                    | Authorization  | 5.8%                                                                      |

Top five CWE categories regarding the annual weighted occurrence among critical- and high-severity findings. Four of the five fit into authorization or authentication categories. See §2.10.1 for the full persistent problems table.

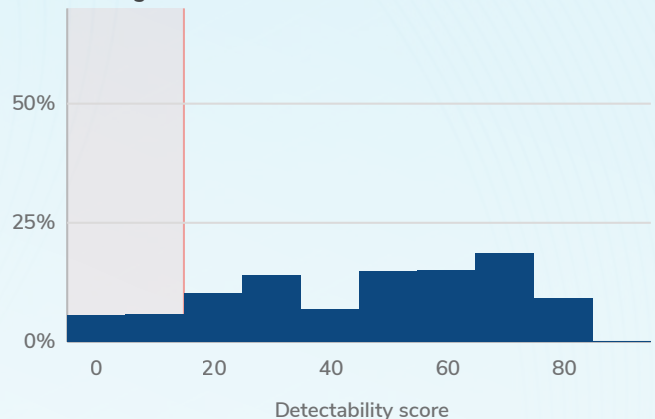
Issues within these categories are also among the least reachable by static analysis, as measured by our SAST/SCA detectability score (§2.8). Roughly 86% of Broken Access Control and Identification and Authentication Failures findings score below 20 on that scale, a range in which the industry does not typically expect SAST/SCA tooling to identify findings. For comparison, only about 11% of Injection findings sit in that range. Authorization and authentication findings remain difficult to detect because they cannot be caught in code alone. Finding them requires an architectural review or testing against a running system.



**Authorization & Authentication findings**  
86% of findings score < 20



**Injection findings (for comparison)**  
11% of findings score < 20



The distribution of SAST/SCA detectability scores for Authorization and Authentication findings (left) is compared to Injection findings (right), with all severities pooled. The same y-scale is used for both panels. The shaded left region on each panel marks findings scoring below 20, which is the range that the industry does not typically expect static analysis to reach. Different distributions have different implications. SAST is working as designed for Injection findings, but it is not designed for detecting authorization and authentication issues.

**Broken Access Control issues alone account for more than 40% of all critical- and high-severity findings.**

The industry is very familiar with injection, dependency risk and supply chain compromising. The industry is less familiar with the categories that are most consistently overrepresented in the serious findings band: **authorization logic and authentication flows. They deserve more room in the discussion.**

## TREND

### The Security Health Divergence: Regulation Doesn't Dictate Security

Not every industry is on the same security trajectory. It is difficult to state which are the determining factors regarding industries that have an improving trajectory versus a falling one. However, regulation is not a factor that can be used to predict which side of this security health divergence an industry falls on.

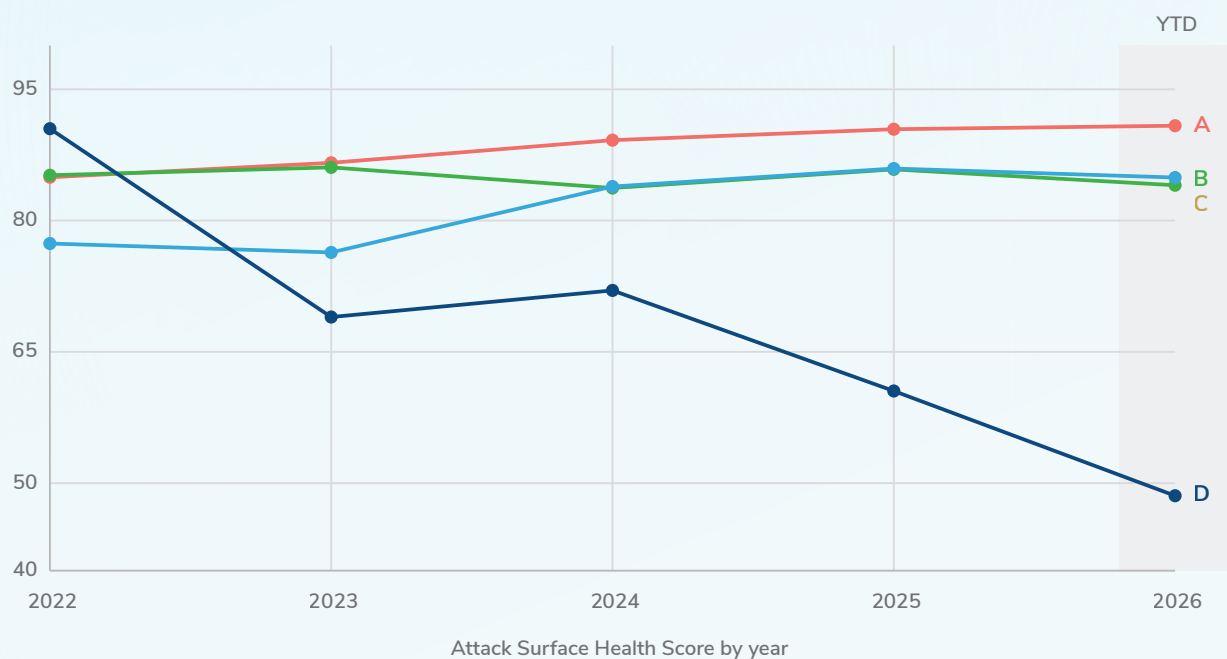
## What This Means for Security Leaders

Meeting the letter of regulatory security requirements is not, by itself, correlated with improving application security posture. Two industries under similar regulatory pressure produced opposite trajectories in our data. The regulatory floor is a floor, not a ceiling, and not, apparently, a driver. What separates programs that improve from those that stagnate is a question this report leaves open, having ruled out one common answer and offered a hypothesis worth further examination: the benefit of proactive security programs.

## EXPOSITION

Application security is best described not as a single state but as a distribution. Different industries present different health scores, and the trajectories of these scores move in different directions. This final trend examines what might explain that pattern. We begin with two empirical observations grounded in our data. We then close with a hypothesis we cannot directly measure but believe is worth the industry's consideration.

No single trajectory appears across our data. Some industries have shown clear improvement over the observation window, others have declined, and some have held roughly steady. The Attack Surface Health Score trajectories for four industries with meaningful multiyear samples illustrate the range.



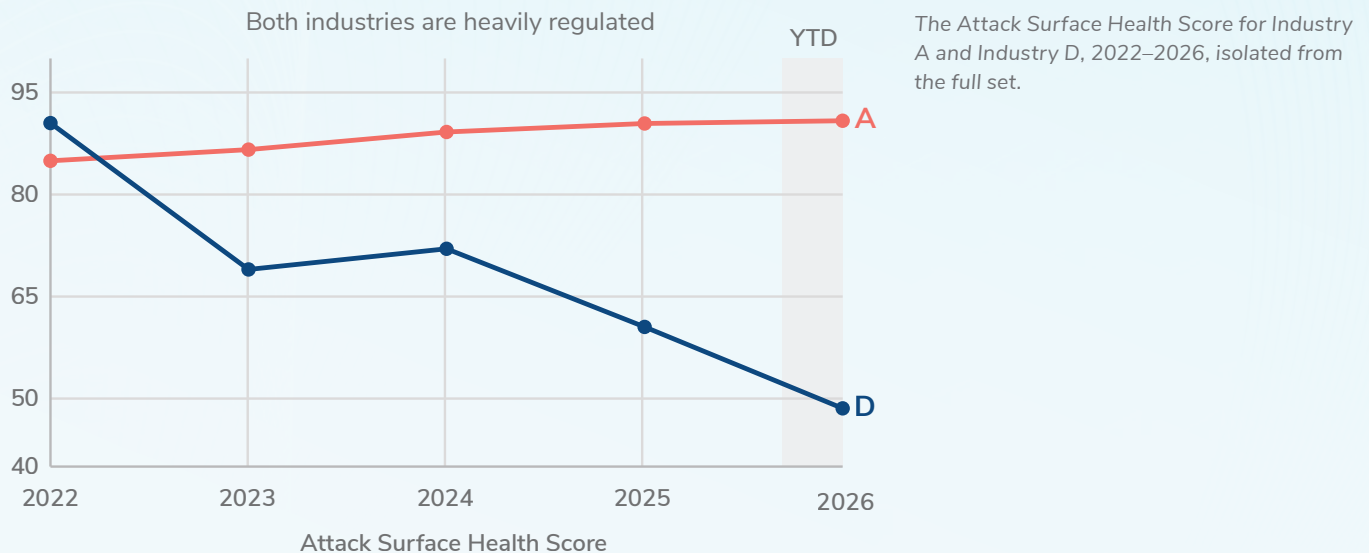
The Attack Surface Health Score by year is shown for four industries with multiyear samples. Industry labels are anonymized. Industry A is a heavily regulated provider of financial services (the largest sample in the dataset), Industry B is a large sample telecommunications services industry, Industry C is a large sample hospitality services industry, and Industry D is a heavily regulated provider of insurance products.

### Industry A's Improvement in Four Signals

- **Fewer findings per penetration test, on average:** Finding density declined by about 16% over the study window (from 4.5 to 3.7 findings per engagement).
- **Less severe findings, on average:** The average severity of findings dropped by about 23%.
- **Harder-to-find findings, on average:** Average finding sophistication rose by about 8%.
- **Broad-based improvement:** Within-category severity declined in every OWASP category that had an adequate sample size.

Although Industry A represents a clear rise in security health, Industry D represents a striking fall. The other two trajectories, with differing starting points, are roughly flat. The industries occupying our dataset are not moving in the same direction.

Regulation is not what separates improving industries from stagnating ones. You might assume that compliance forces security work and that heavily regulated industries should therefore improve most consistently. Our data contradicts that assumption. **Industry A** (a heavily regulated financial services industry) improved steadily over the observation window. **Industry D** (a heavily regulated insurance products industry) did not. They share the same regulatory environment, and both had multiyear sample sizes, yet their security health trajectories moved in opposite directions.



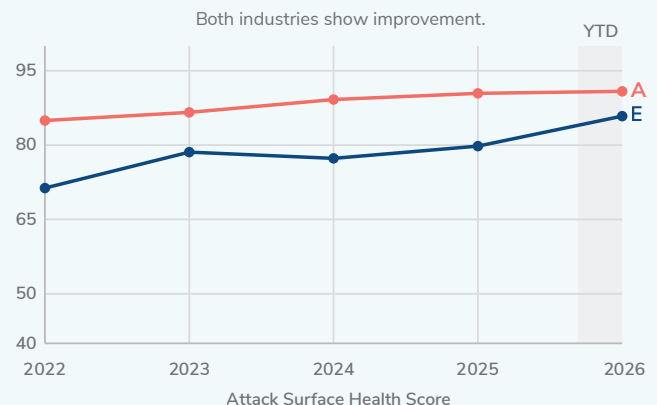
**Regulation is necessary but not sufficient. Something else separates the improving industries from the stagnating ones.**

## POSSIBLE EXPLANATIONS

What separates the improving industries from the stagnating ones? Before proposing an explanation, we considered several candidates and found each to be insufficient.

If regulation is not the differentiator, what is?

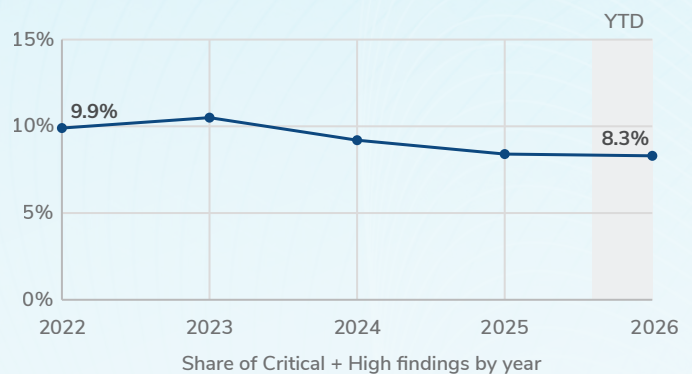
One candidate is technology conservatism. Financial services and insurance products industries typically operate mature, slowly changing technology stacks, so the reasoning would be that a slower-changing surface introduces fewer new vulnerabilities. However, this cannot be the differentiator. A technology sector industry (e.g., **Industry E** in the chart) is characterized by rapid change, yet it also shows improvement in the same metrics. A slow-moving tech stack is not required.



**Fast-changing technology stacks do not necessarily create poorer attack surface health.**

The Attack Surface Health Score for Industry A (a heavily regulated financial services industry) and Industry E (a technology sector industry characterized by rapid change) 2022–2026.

A second candidate is accelerated post-pandemic digitalization. If one industry had rapidly deployed new applications without commensurate security maturity from 2020 to 2022, we would expect a spike in critical- and high-severity findings between 2022 and 2024. We have not observed such a spike. The severity mix of findings across the dataset has stayed remarkably stable throughout the observation window, with the share of findings tagged as critical or high severity holding at roughly 8%–10% every year from 2022 onward. Whatever pressures the pandemic placed on application security, they did not manifest as a discrete event in our data.



**Accelerated post-pandemic digitalization did not cause a spike in vulnerabilities.**

## KROLL'S READING

What remains is a hypothesis that involves a differentiator we cannot directly measure: proactive security culture. The industries showing the clearest improvement are also, on average, the ones that engage in penetration testing most frequently. They test at a high volume on a consistent cadence and maintain an ongoing, multiyear relationship with our firm. If the differentiator is not compliance obligation, then this reading sees the benefit in treating security as an ongoing operational discipline.

A proactive security culture would show up in more than penetration testing cadence. It would also show up in hiring and retaining in-house security practitioners, engineering teams that treat security as a design concern, routine use of shift-left tooling like SAST/SCA as part of the SDLC, and governance frameworks that exceed regulatory minimums. Although the beneficial impact of penetration testing engagements for security health is made visible by our dataset, it is only one factor among several.

The hypothesis that more frequently testing at a high volume is always more beneficial is imperfect. Industry B (the large sample telecommunications industry) has a high volume of engagement but a flat trajectory, which stands as a counterexample to the hypothesis. Additionally, from this dataset alone, we cannot distinguish a client whose proactive culture leads to more testing from a client whose long relationship with our firm has accumulated more data points; the two are difficult to separate. What we can say is what regulation does not do. We suspect that proactive engagement matters more than compliance obligation and offer this as a reading rather than proof.

**Our data shows that regulation does not mean improved attack surface health. Our interpretation is that proactive engagement with a penetration testing mindset matters more than compliance obligation.**

The empirical piece of this trend is that **industries diverge regarding security health, and the existence of regulation does not predict which way the divergence will occur.** The interpretive piece of this trend is that proactive security engagement (i.e., penetration testing by choice rather than due to obligation) appears to align with the healthier side of the divergence, though the data cannot definitively say so. Both pieces belong in the conversation about what makes a mature application security program. Neither should be reduced to a lone cause. The first is a caution about compliance as strategy, and the second is a hypothesis we have offered for the industry to weigh.

## SECTION 2

# The Data: Slicing the Vulnerability Landscape

## 2.0 Methodology and Data Set

### The Empirical Basis Of This Report

This report is based on data drawn from **several thousand web and mobile application penetration tests conducted for Kroll clients over a five-year window ending in June 2026**. Every finding referenced in these pages originated in a real engagement, was validated by our consultants, and was delivered to and accepted by the client in a formal penetration testing report (as opposed to surveys, telemetry or public disclosures).

### Scope

The report is deliberately focused on web and mobile application penetration tests.

Network and infrastructure penetration tests, red and purple team engagements, cloud attack-path assessments and all other service lines were excluded. For engagements in which our source dataset included both infrastructure and application findings, we filtered out the infrastructure findings so that the numbers reported here reflect only the application layer.

### What We Count

The unit of measurement throughout this report is the **finding**, a distinct vulnerability that a consultant validated and documented in a client-delivered report, whether it was first discovered by a tester or professional penetration testing tooling. Unvalidated scanner findings are not included.

### Time Coverage And Comparability

The dataset spans engagements from **2020 through June 2026**. Coverage is thinner at both ends of the time window: 2020, the earliest year, and 2026, which is a partial year running through June. When we present trends over time, we treat these first and last years with care. Some charts stop at the last complete year (2025), and any 2026 figure is labeled as year to date.

We frame this report's insights as **general observations about the state of application security** rather than sharp year-over-year deltas. The dataset supports directional statements with confidence, and we resist overinterpreting single-year movements. If a chart's meaning depends on how scanner-derived findings were identified in earlier years, we note the caveat in-line.

### Characteristics Of The Dataset

This report offers a view of application security from the front line of mature security programs.

The dataset skews toward **organizations of various sizes that operate mature, recurring security programs**, with strong representation across **North America, EMEA and clients with global footprints**. This angle reflects the segment of the market that invests most consistently in independent security assessments.

Readers should interpret the findings in this context. This is not a universal survey of every application in every organization. It is a highly detailed, data-driven view of what we are finding in the applications that already receive serious security attention. Therefore, the persistence of the issues we describe should be taken as more, not less, significant.

## Client Confidentiality And How We Protect It

The data underlying this report belongs to our clients. Protecting them is an essential precondition. Therefore, we apply a small set of firm rules throughout:

- **No absolute counts of clients, engagements or findings appear anywhere in this report.** All numbers are expressed as **rates, ratios, proportions, distributions or ranks**. The dataset's scale is described in order-of-magnitude distinctions only.
- **No slice of the data that is smaller than a minimum sample size is published.** If a slice is too small to be reported safely, we either merge it into a larger bucket or omit it from that cut. If metadata was missing or ambiguous for a given engagement, we exclude that engagement from the affected cut but never from the overall totals.
- **No named outliers, signature vulnerabilities or case-specific details appear in any aggregate figure.** We exclude any finding pattern that appears in only a small number of engagements, even in aggregate results, because its presence could identify the clients involved.
- **Every chart is reviewed with a single question in mind: Could an adversary who knew we had tested a specific organization use this data point to confirm or refine that knowledge?** We redesign or remove any chart that fails this test.

We would rather be directionally right and defensibly honest than falsely precise.

## 2.1 The Big Picture

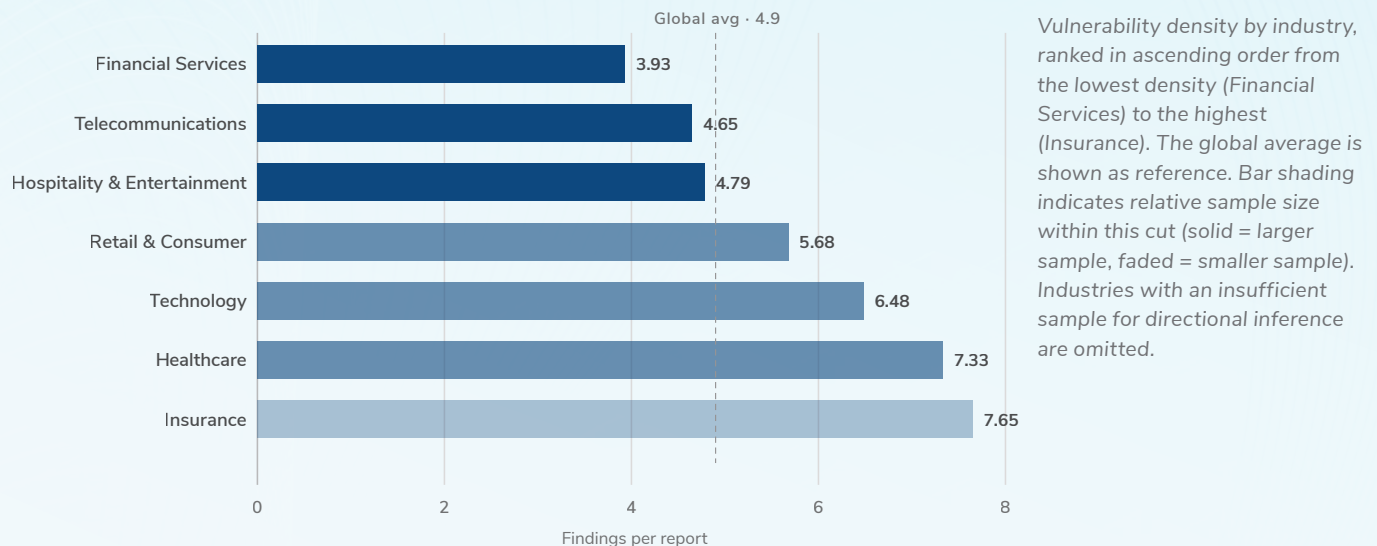
We now examine a single population of findings through nine complementary lenses. Each subsection isolates one dimension (e.g., how many, of what type, from where and how) and reports what the data shows without positing what it means. Interpretation is deliberately reserved for the subsequent sections. The subsections that follow can be read in order or approached by topic. The following questions map the concerns readers most often bring to a report of this kind to the sections that answer them.

|                                                                     |                                                 |
|---------------------------------------------------------------------|-------------------------------------------------|
| How many vulnerabilities does a typical penetration test reveal?    | <b>§2.2</b><br>Vulnerability Density            |
| What kinds of vulnerabilities are most common?                      | <b>§2.3–§2.5</b><br>The Taxonomies              |
| Where do vulnerabilities originate—design, code or deployment?      | <b>§2.6</b><br>Vulnerability Introduction Point |
| What can scanners catch, and what needs a human?                    | <b>§2.7</b><br>Finding Sophistication           |
| Would static analysis have caught these earlier?                    | <b>§2.8</b><br>SAST/SCA Detectability           |
| How is application security health overall, and how is it trending? | <b>§2.9</b><br>Attack Surface Health Score      |
| Which specific issues keep recurring year after year?               | <b>§2.10</b><br>Persistent Problems             |

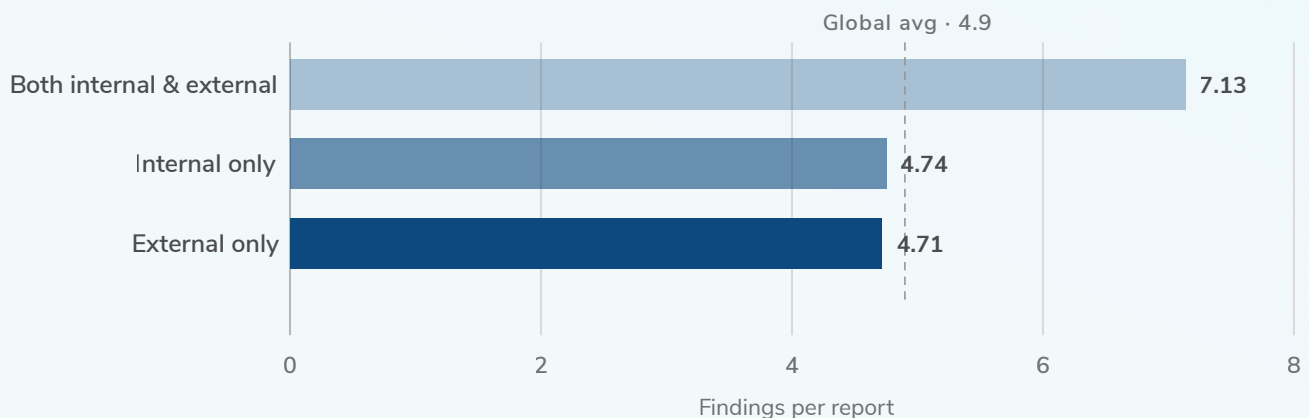
## 2.2 Vulnerability Density

Vulnerability density (i.e., the number of validated findings per penetration testing report) measures the rate at which issues surface during testing. Across the dataset, the global average sits at approximately **4.9 findings per report**. The cuts below show how that rate varies by industry and engagement scope.

### 2.2.1 Density by Industry



### 2.2.2 Density by Engagement Scope



#### Kroll Observations

- Financial services shows materially fewer findings per report than any other large industry—roughly 20% below the global average. Insurance and healthcare sit at the opposite end, each around 50% above.
- Internal and external engagements produce nearly identical density. This runs counter to the common assumption that internet-facing applications carry higher finding rates.
- The elevated density of combined-scope engagements is a scope effect, not a signal about the applications themselves. Testing both surfaces produces more recorded findings, but that does not mean those applications are inherently more vulnerable.

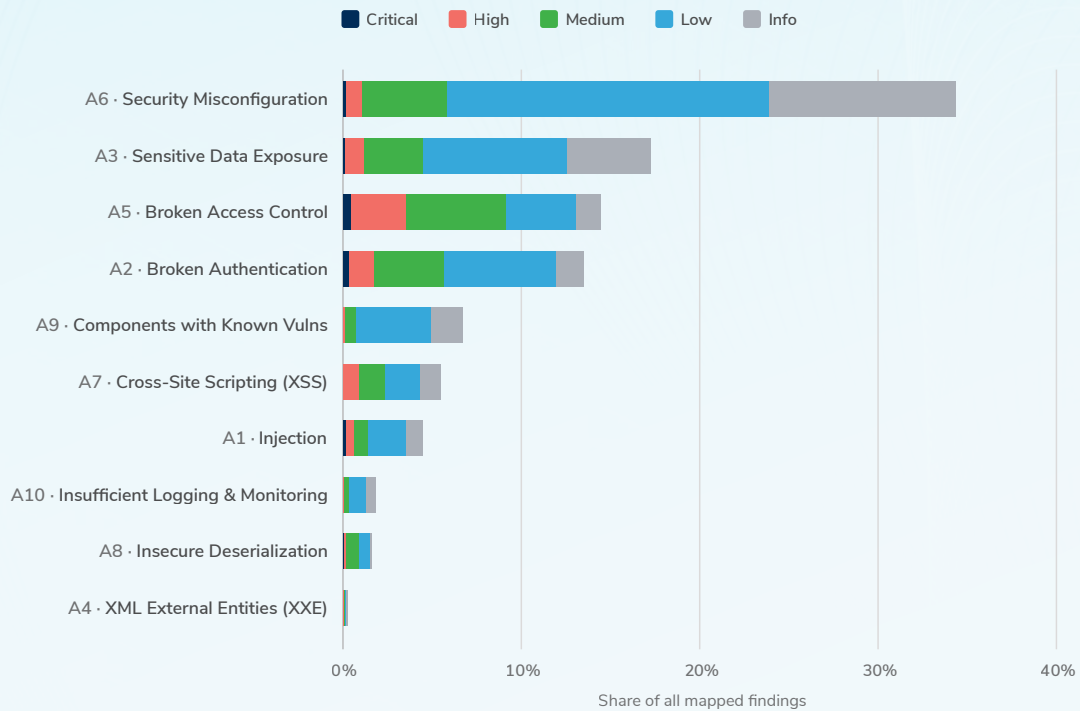
Financial services firms are affected by the least number of vulnerabilities in penetration tests—**20% less than the global average**.

Insurance and healthcare are affected by the greatest number of vulnerabilities in penetration tests—around **50% above the global average**.

## 2.3 OWASP Top 10 (2017)

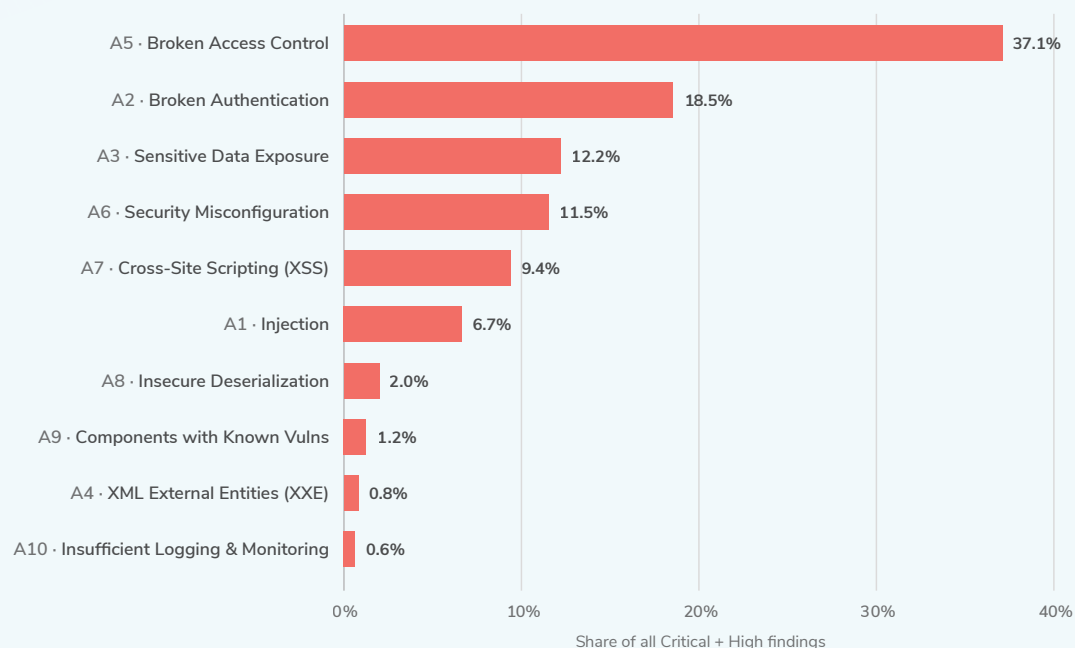
Every finding in the dataset is mapped to an OWASP 2017 category. The following cuts present two complementary views: the **overall distribution** of findings across categories (segmented by severity) and the **concentration of critical findings** by category.

### 2.3.1 Findings by Category and Severity



OWASP 2017 categories are ranked by share of all mapped findings. Bar length represents category volume, and segment color represents severity. The chart intentionally combines these two views. The length of each bar shows how common findings in that category are, and the mix of segments within the bar shows how severe those findings tend to be when they occur.

### 2.3.2 Where Higher-Severity Findings Concentrate



The share of all critical- and high-severity findings, organized by OWASP 2017 category. This band represents findings that most application security programs treat as priority remediation targets.

## Kroll Observations

- **The most common finding class is not the most serious.** The Security Misconfiguration category represents roughly one in three findings, but the great majority are rated with a low or informational severity. Critical- and high-severity findings concentrate elsewhere, in the Broken Access Control and Broken Authentication categories.
- **Broken Access Control and Broken Authentication findings dominate the higher-severity band.** Together, these categories account for roughly 56% of all critical- and high-severity findings, which is well above their combined share (approximately 28%) of overall findings.
- **Cross-Site Scripting (XSS) and Injection findings represent a smaller share than their reputation suggests, but they are disproportionately serious when found.** Together, these categories account for only about one in 10 findings, which is well below the prominence they hold in older OWASP rankings. However, both are overrepresented in the critical- and high-severity bands.

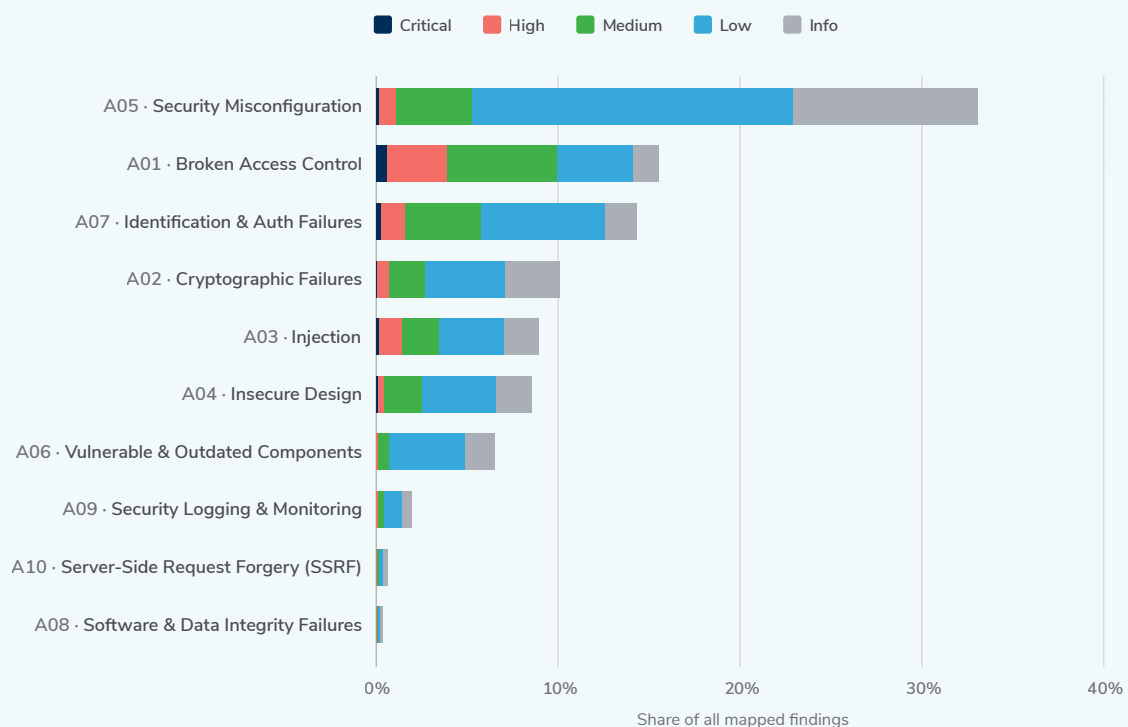
**Nearly 60% of serious vulnerabilities were either Identification and Authentication Failures or Broken Access Control issues.**

**Only 15% of high-severity findings were XSS and Injection vulnerabilities.**

## 2.4 OWASP Top 10 (2021)

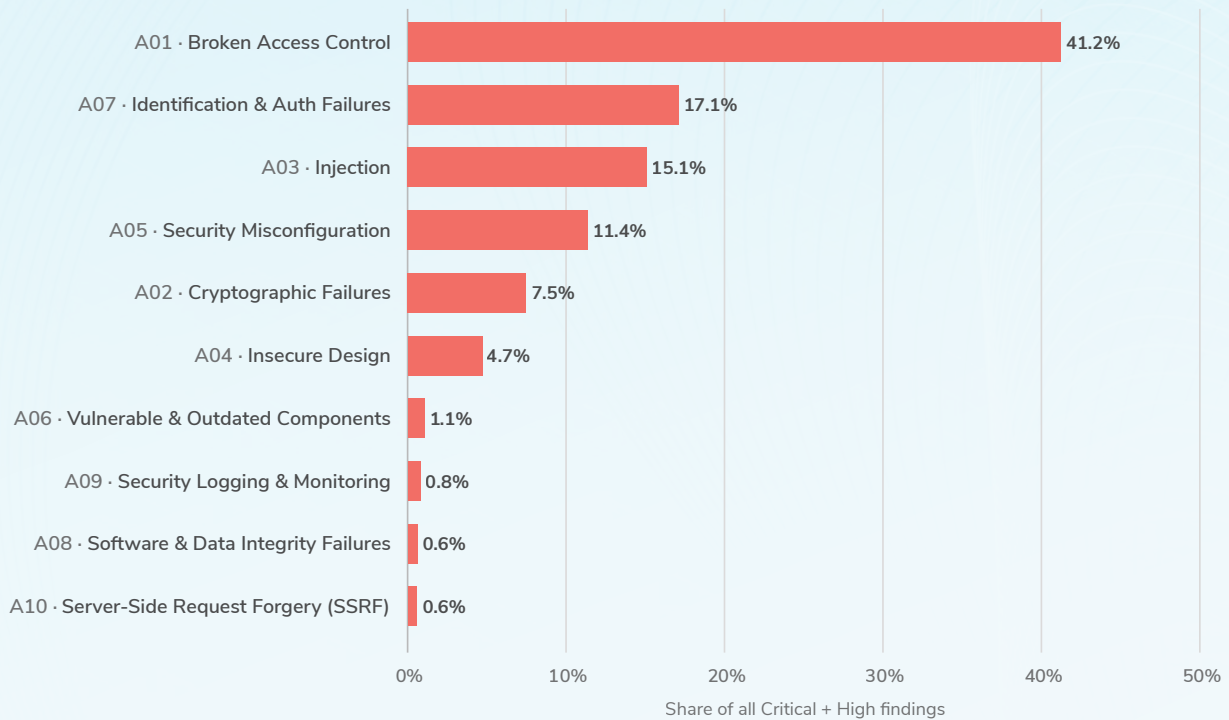
The same dataset of findings is presented in this section but viewed through the newer OWASP 2021 taxonomy. The core purpose of this cut is twofold: to confirm that the picture is stable across taxonomies and to offer an empirical perspective on some of OWASP's own revision choices, particularly the introduction of root cause categories such as Insecure Design.

### 2.4.1 Findings by Category and Severity



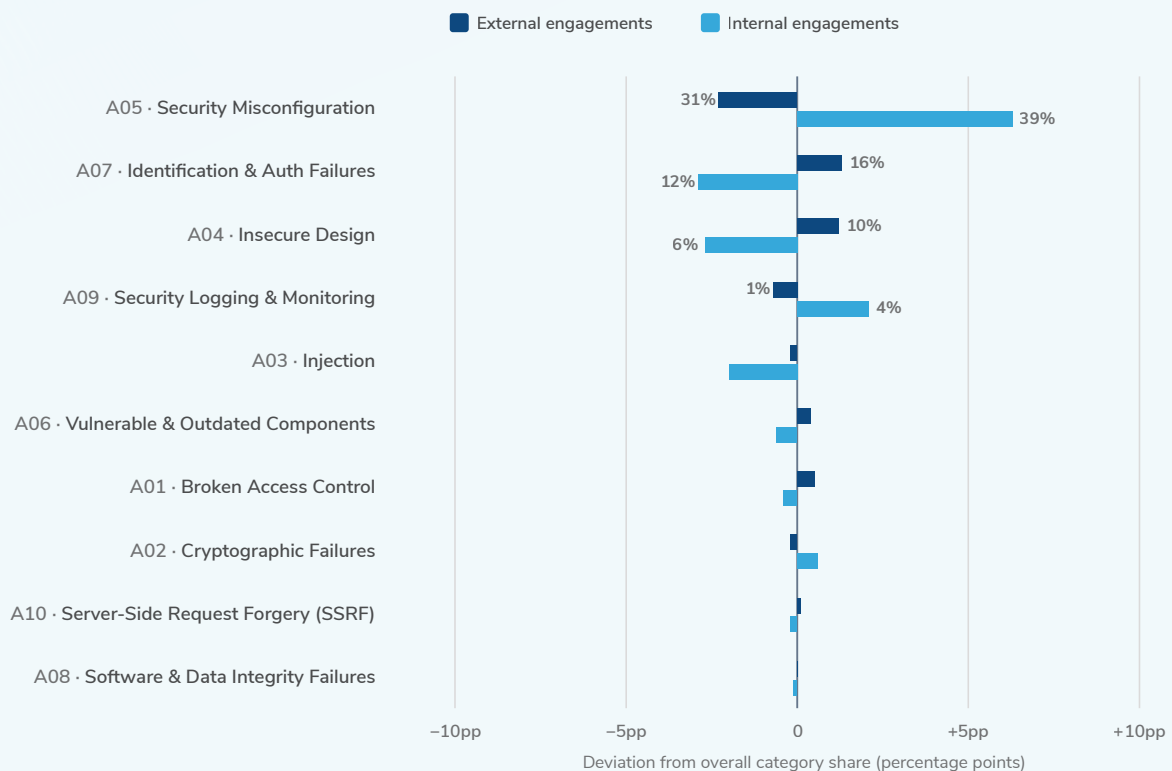
OWASP 2021 categories are ranked by share of all mapped findings. The bar length represents category volume, and segment color represents severity.

### 2.4.2 Where Higher-Severity Findings Concentrate



The shares of all critical- and high-severity findings are organized by OWASP 2021 category. Broken Access Control findings account for over 40% of higher-severity findings in the 2021 mapping.

### 2.4.3 How the Category Mix Shifts by Testing Scope



Category share deviation from the overall mean is shown for both external and internal testing scope. Bars extending right indicate overrepresentation in that scope, and bars extending left indicate underrepresentation. Categories are organized by the size of the external versus internal gap. Engagements that include both scopes are omitted for chart clarity.

## Kroll Observations

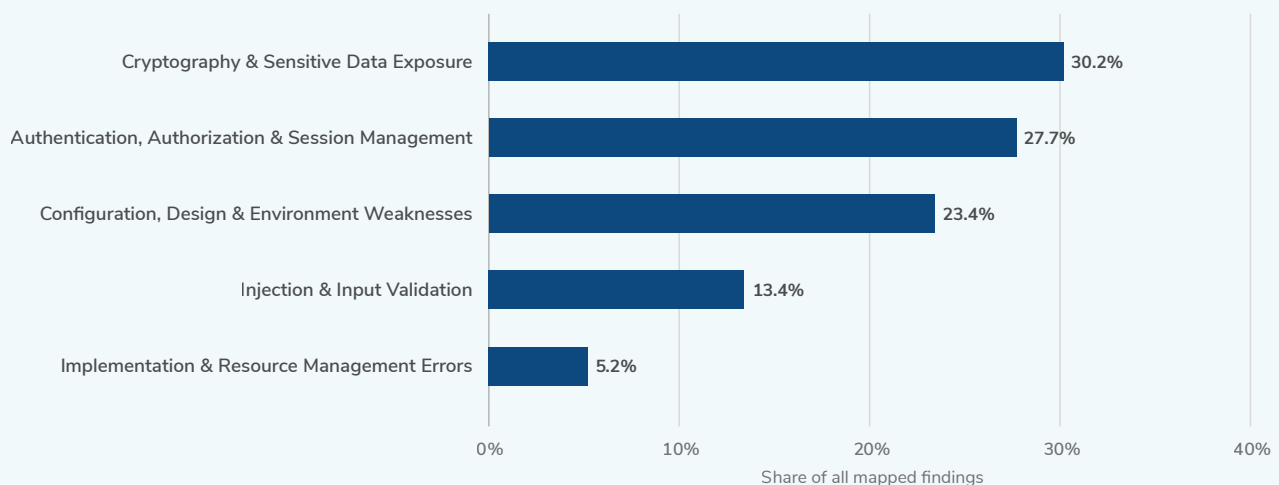
- **Broken Access Control findings alone account for over 40% of higher-severity findings.** The broader picture is stable across OWASP category revisions. Security Misconfiguration leads by volume, and access control and authentication failures dominate the critical- and high-severity band, regardless of which taxonomy you use.
- **The taxonomy undercounts design-level vulnerabilities.** OWASP 2021 mixes outcome categories (e.g., Broken Access Control, Injection and Cryptographic Failures) with root cause categories (e.g., Insecure Design, Software and Data Integrity Failures). When a design flaw manifests as a concrete outcome, testers tag the outcome because it's more actionable for the client. That leaves the root cause categories systematically undercounted. We return to this pattern from a different angle in our Vulnerability Introduction Point analysis.
- **Server-Side Request Forgery (SSRF) findings are rare, accounting for less than 1% of all findings, and they are not disproportionately severe when they do occur.** This point is consistent with OWASP's own note that SSRF was included in the Top 10 partly due to community concern rather than data prevalence. The severity of exploitable SSRF findings likely materializes through consequences they enable downstream in areas that aren't attributed to the SSRF category.
- **Internal and external testing engagements result in different kinds of findings.** Internal engagements reveal more configuration issues and more logging or monitoring gaps. External engagements reveal more authentication and design-level findings. The Insecure Design gap is the most interesting, and it likely arises because external testing more often exposes business logic and workflow design flaws, which are hard to classify under a category that names a concrete outcome.

**Security misconfigurations are the most common vulnerabilities found overall.**

## 2.5 CWE Taxonomies

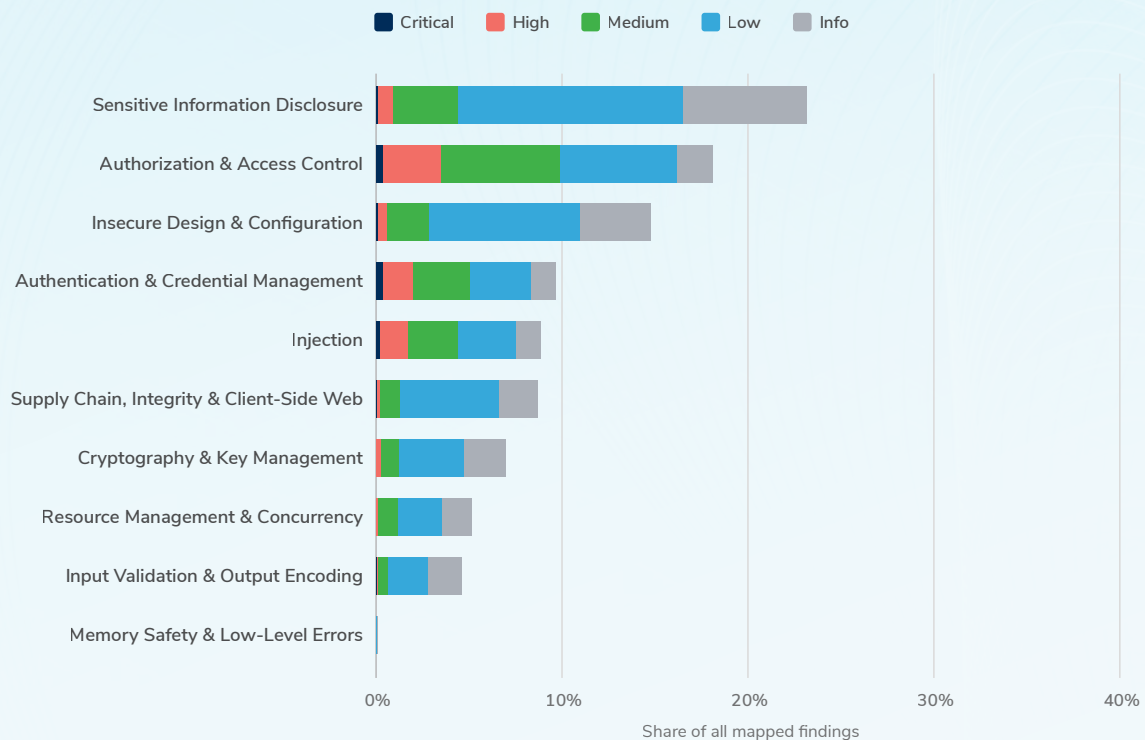
The raw CWE catalog lists roughly 900 distinct weakness categories, which is too granular for a report-level view, and neither OWASP nor CWE standards provide an intermediate grouping that is suited to this style of wide-view reporting. To bridge the gap, our consultants map each finding to two derived taxonomies: a five-category compressed view and a 10-category detailed view. Every finding in the dataset is mapped by both. The same dataset of findings is presented in the two following views.

### 2.5.1 The Compressed View (Five Categories)



The CWE compressed taxonomy is expressed in five categories. The first four categories together account for 95% of the finding population.

### 2.5.2 The Detailed View (10 Categories, by Severity)



The CWE detailed taxonomy is expressed across 10 categories. The bar length represents category volume, and the segment color represents severity. The Sensitive Information Disclosure category has the most volume, and the Authorization and Access Control category leads in higher-severity concentration.

#### Kroll Observations

- **Authorization and Access Control failures dominate serious findings, whichever taxonomy you use.** Across all three lenses—OWASP 2017, OWASP 2021 and CWE—the pattern is the same: Descriptive-outcome categories dominate volume, and access control and authentication failures dominate the higher-severity band. The consistency across independently constructed taxonomies confirms that the pattern is a true property of the finding population and not of any one classification scheme.
- **Sensitive Information Disclosure is one of the largest sources of findings, but these issues are rarely serious.** CWE groups findings by what leaks rather than by which mechanism failed. Under this lens, Sensitive Information Disclosure (through verbose error messages, over-permissive logging or side channels) is the single biggest category by volume, though most of these issues have a lower severity.
- **Memory safety issues are rare in the modern web and mobile stack.** These issues account for a negligible share of findings, which is a strong signal that decades of industry investment in memory-safe languages and safer runtimes have materially reduced the attack surface. When these issues do appear, they skew toward a higher severity. The category is best understood as a legacy risk that industry practice has largely engineered away, but it has not been fully retired. Practitioners working with memory-unsafe components should still assume the risk is present.

**Information disclosure is the most common type of leak found across penetration tests.**

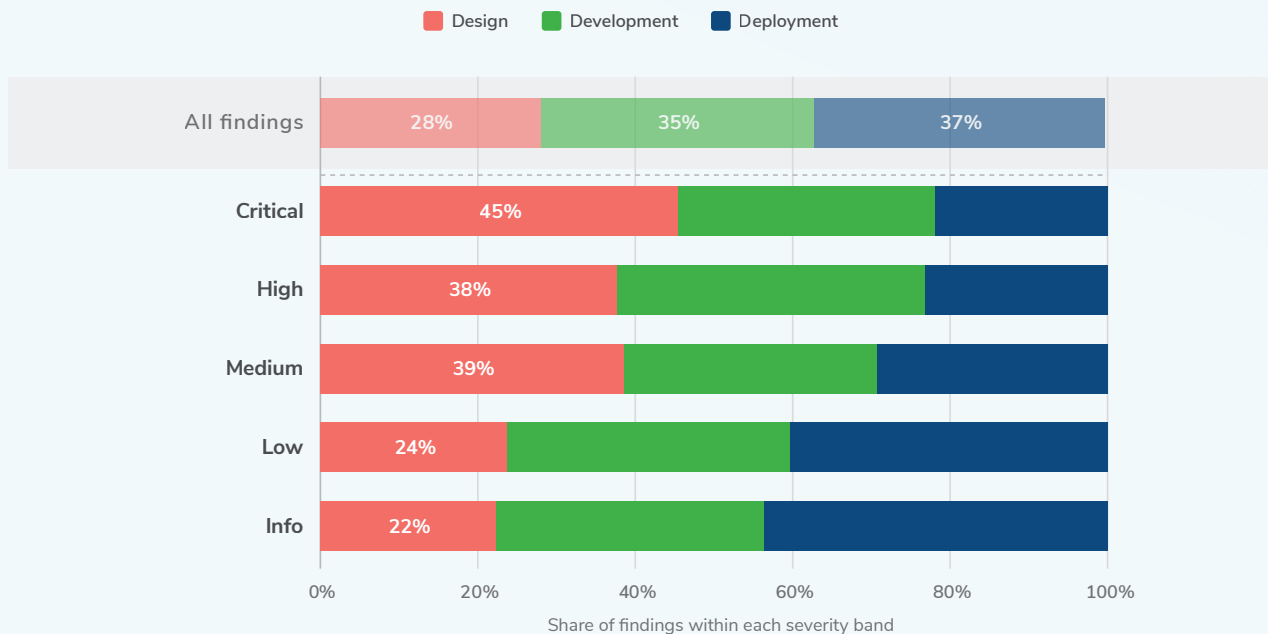
## 2.6 Vulnerability Introduction Point

Each finding in our dataset is categorized by the stage in the target's life cycle during which the vulnerability was most likely introduced. Three categories are used, along with a fourth reserved for cases where categorization cannot be reliably determined:

- **Design:** The vulnerability was introduced during the planning or architectural stage. The system was built as intended, but the design itself was flawed (e.g., a single-role authorization model that permits privilege abuse and a shared-kernel container model that permits escape).
- **Development:** The vulnerability was introduced during implementation. The design was sound, and defensive mechanisms were intentionally put in place, but a coding error allowed those mechanisms to be circumvented (e.g., validation bypasses, filter edge cases and check-order mistakes).
- **Deployment:** The vulnerability was introduced during the target's exposure to its operating environment. The code and design were sound, but the deployment context introduced the weakness (e.g., weak TLS configuration on a reverse proxy, exposed administrative endpoints, and permissive access control rules on cloud infrastructure).
- **Inconclusive:** This category is reserved for findings with write-ups that do not contain enough detail to attribute the introduction point. Consultants are instructed to categorize honestly rather than guess. In practice, less than 0.2% of findings fall into this category, and it is omitted from the charts in this section for readability.

The categorization is inherently a qualitative judgment. The boundary between the design and development categories in particular can involve interpretation. For example, it can be difficult to decide whether missing input sanitization reflects a design that failed to specify sanitization as a requirement or whether a development error was made in a system in which sanitization actions were intentionally implemented but done so incorrectly. This is a call that the reviewing consultant makes based on the context of the finding.

### 2.6.1 Introduction Point Mix Across Severity Bands

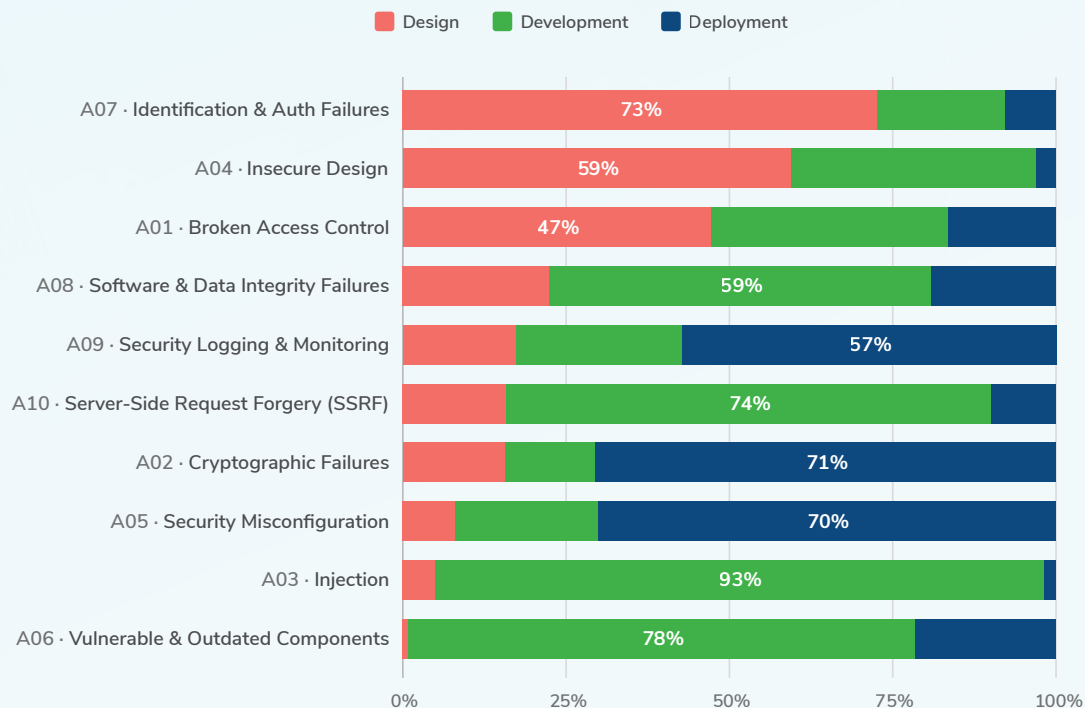


The All findings row at the top shows the introduction point mix across the full population as a reference and at a reduced opacity to distinguish it from the severity bands below it. The five severity rows show the mix within each band. The design category's share is labeled on each bar. Reading down the chart, the aggregate of findings looks roughly balanced, but the severity bands do not.

## Kroll Observations

- The aggregate of findings looks roughly balanced among design (about 28%), development (about 35%) and deployment (about 37%), but when they are split by severity, the balance disappears:
  - Design share ranges from 45% critical to 22% informational.
  - Deployment share ranges from 22% critical to 43% informational.
  - Development share is roughly steady, hovering between 32% and 39% across all five bands.
- The stage at which a vulnerability is introduced tracks closely with how serious it turns out to be. A design issue is more likely to be severe than a deployment issue.

### 2.6.2 OWASP 2021 Categories, Decomposed by Introduction Point

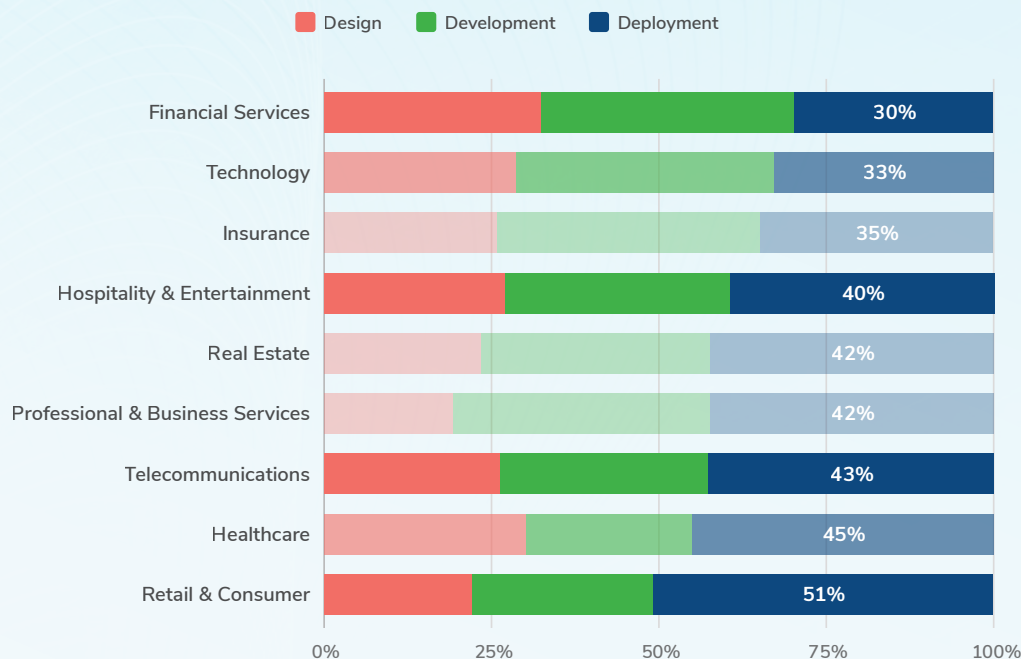


Each OWASP 2021 category is broken down to show the distribution of introduction points for the underlying findings. Issues in the Broken Access Control and Identification and Authentication Failures categories (which are both outcome categories) are primarily caused by design choices.

## Kroll Observations

- Design-stage vulnerabilities are three times more common than OWASP's Insecure Design category suggests.
  - **Categories dominated by design decisions:** Identification and Authentication Failures (73%), Broken Access Control (47%) and Insecure Design itself (59%)
  - **Categories dominated by development mistakes:** Injection (93%), Vulnerable and Outdated Components (78%) and SSRF (74%)
  - **Categories dominated by deployment configuration:** Cryptographic Failures (71%), Security Misconfiguration (70%) and Security Logging and Monitoring (57%)
- To assist clients in focusing on remediation activities, consultants often rate findings with outcome-focused categories. However, doing so can hide the true scale of design mistakes.

### 2.6.3 Introduction Point by Industry Vertical

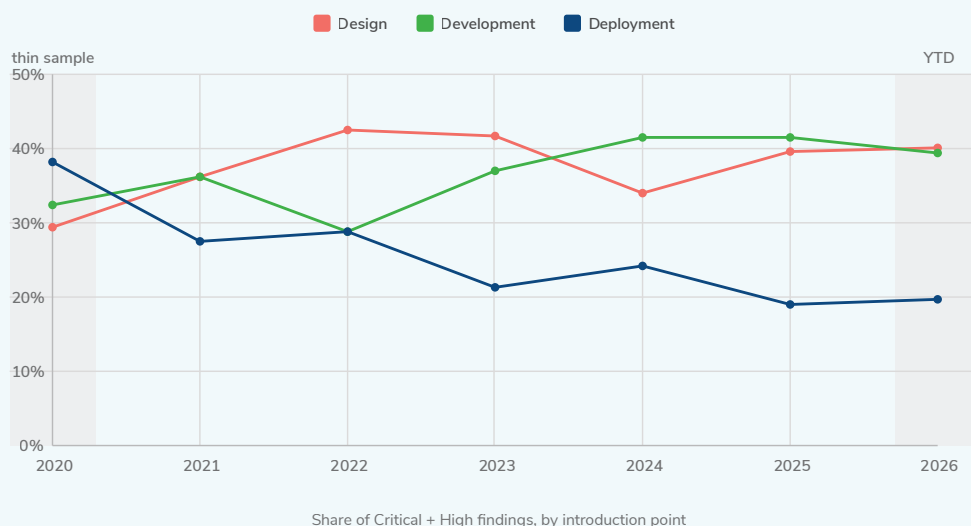


The mix of introduction points is organized by industry. The rows are shown in ascending order by deployment share, which is labeled on each bar, from the most deployment-lean industry (Financial Services) to the most deployment-heavy industry (Retail and Consumer). Bar shading indicates relative sample size within this cut (solid = larger sample, faded = smaller sample), consistent with the convention introduced in Section 2.2. Industries with less than 50 engagements are omitted for reliability.

#### Kroll Observations

- The industries with the lowest vulnerability density (see Section 2.2) also show the lowest share of findings introduced in the deployment stage. Two independent frameworks focused on the same industries produce broadly the same rankings.
- Financial services sits at one end of both rankings, with the lowest density (about 3.9 findings per report) and the lowest deployment share (about 30%). At the other end, retail and consumer (about 51% deployment) and healthcare (about 45%) exhibit the opposite profile, with a heavier weighting toward the introduction points that are, in principle, cheaper to remediate.

### 2.6.4 How the Mix Has Shifted over Time (Critical and High Only)



The share of critical- and high-severity findings attributed to each introduction point, by year. The years 2020 and 2026 are shaded to indicate a thinner sample size because 2020 preceded consistent vulnerability introduction points tagging at scale, and 2026 is incomplete at the time of writing. The most robust signal in the middle years is the declining share of higher-severity findings introduced during the deployment stage.

### Kroll Observations

- The residual serious risk is concentrating upstream. Findings introduced during the deployment stage make up a shrinking share of the higher-severity band. Conversely, findings introduced during the design and development stages represent an increasing share.
- Findings introduced during the deployment stage accounted for approximately 28% of critical- and high-severity findings in 2021 and 2022, and approximately 19%–24% from 2024 to 2026. These results are consistent with genuine industry-level improvements in deployment discipline. Automation and hardened defaults have reduced the surface through which environment-level mistakes can be introduced. Wherever automation is more difficult and human judgment matters more, the risk persists.

**Critical- and high-severity vulnerabilities in the deployment stage decreased from 2021 and 2022 (28%), compared with 2024 to 2026 (19%–24%), which is consistent with industry improvements to automate and harden defaults.**

**Wherever automation is more difficult and human judgment matters more, risk persists.**

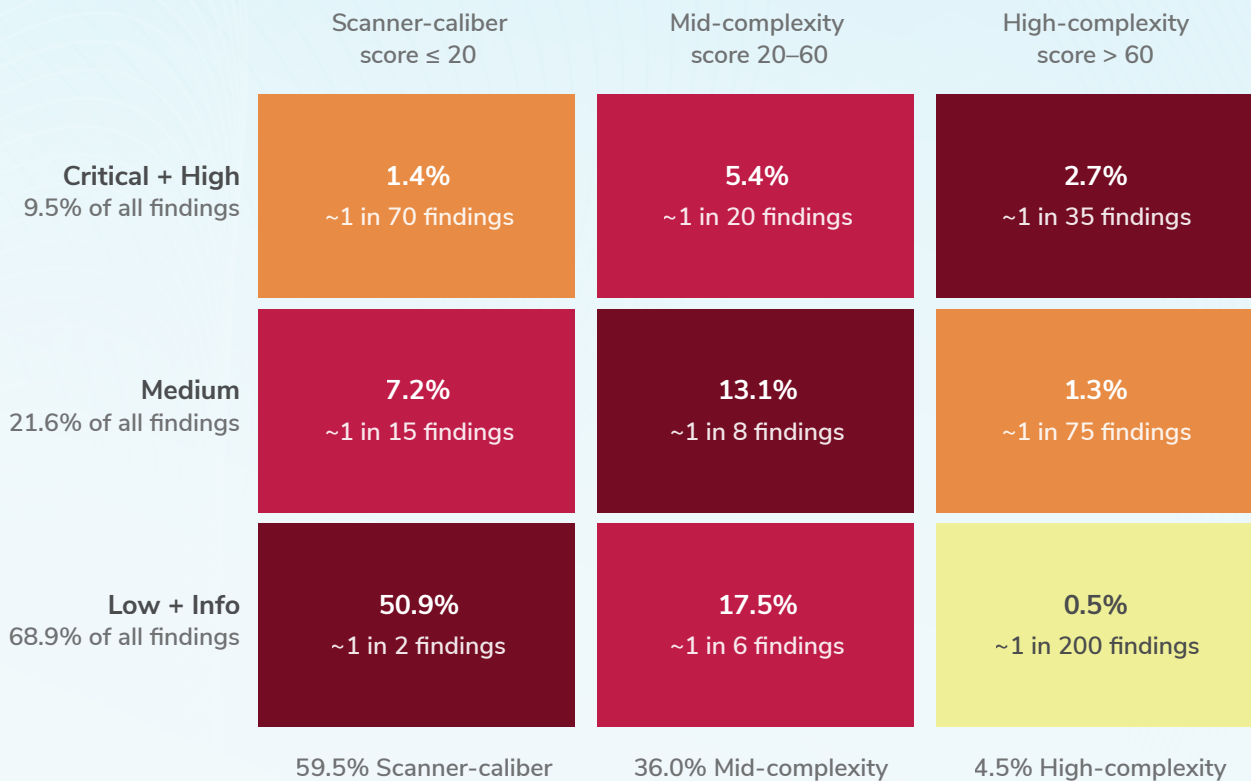
## 2.7 Finding Sophistication

Each finding in our dataset is scored on a **0–100 sophistication scale** that measures the degree of penetration tester ingenuity that was required to identify the finding. A finding that a vulnerability scanner could reasonably catch scores at the low end, and a finding that required multiple chained observations, deep reverse engineering, or substantial reasoning about the target's behavior scores at the high end.

The score exists because taxonomy and severity together don't fully describe a finding. Two critical findings can be very different regarding what an attacker needs to exploit them. One may be a straightforward misconfiguration that casual attackers could weaponize, while another may require the chained analysis of a well-resourced attacker. The sophistication score adds this missing dimension of how much ingenuity exploitation requires. It gives clients a second axis for prioritizing remediation and grounds each finding in a plausible attacker profile. Sophistication is a qualitative property, but not a subjective one. Internal reviewers scored the dataset, working to apply a consistent judgment across findings other than their own.

One helpful way to use the score is as a gauge for what could have been caught by a scanner. We treat scores of **20 or below** as scanner caliber and scores **above 20** as manual caliber. The threshold is a policy choice rather than a data-driven boundary and sits at the shoulder of the overall score distribution.

### 2.7.1 Severity and Sophistication Together

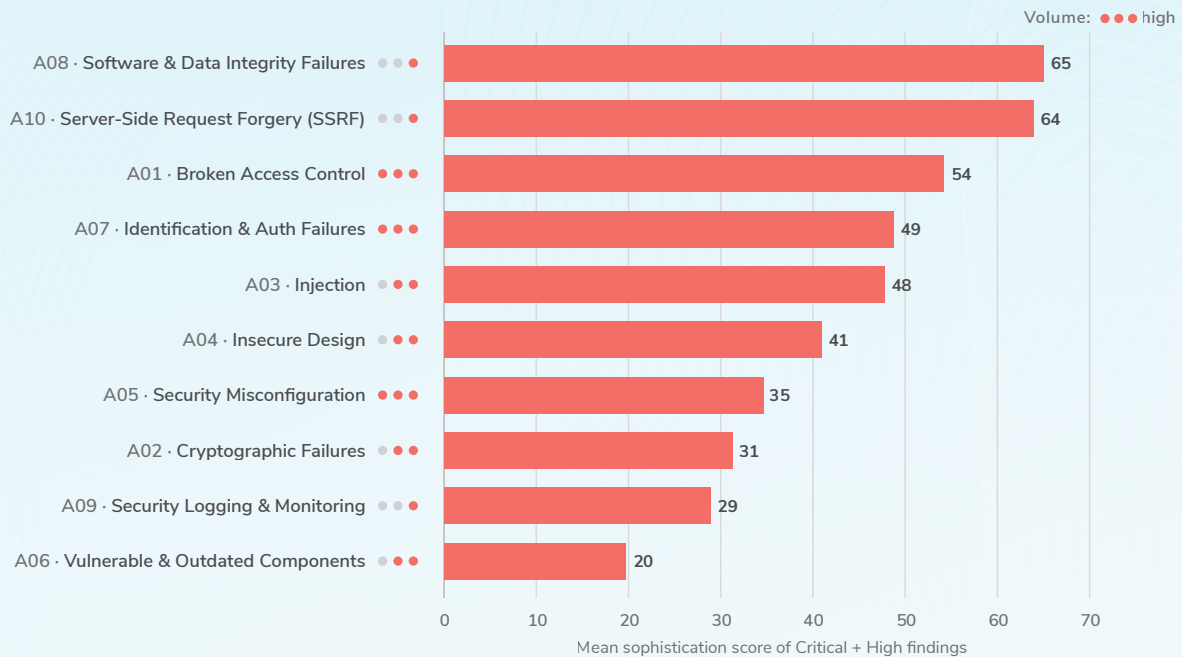


Findings are partitioned by severity band (rows) and sophistication band (columns). Cell color intensity is just a visual aid; the numbers themselves carry the story. The top-right cell (Critical and High, Manual caliber) shows the population for which human ingenuity is required to identify findings that meaningfully matter for risk decisions. The bottom-left cell (Low + Info, Scanner caliber) is the ambient noise that competent automation should identify without a penetration test engagement being performed.

#### Kroll Observations

- About half of all findings carry weight (i.e., they are serious enough to matter for risk, beyond what scanners can reach, or both):
  - **Half the findings are noise.** These are low- or informational-severity findings that a scanner should catch. These findings represent what any competent application security program should already be handling before commissioning a penetration test.
  - **A finding can carry weight for two reasons:** severity or sophistication. About 31% of findings had nontrivial severity (medium or above), and about 41% required manual work to identify. About 22% do both, and roughly half of all findings satisfy at least one. This is why penetration tests exist.
  - **The sharpest edge is a smaller population.** About 3% of findings are critical or high severity and require high-complexity manual work.
- The manual columns concentrate at the serious end of the severity scale, and the scanner column concentrates at the low-severity end. This data demonstrates a real difference in what automated tooling and human testing each find at scale.

### 2.7.2 Mean Sophistication of Serious Findings, by OWASP 2021 Category

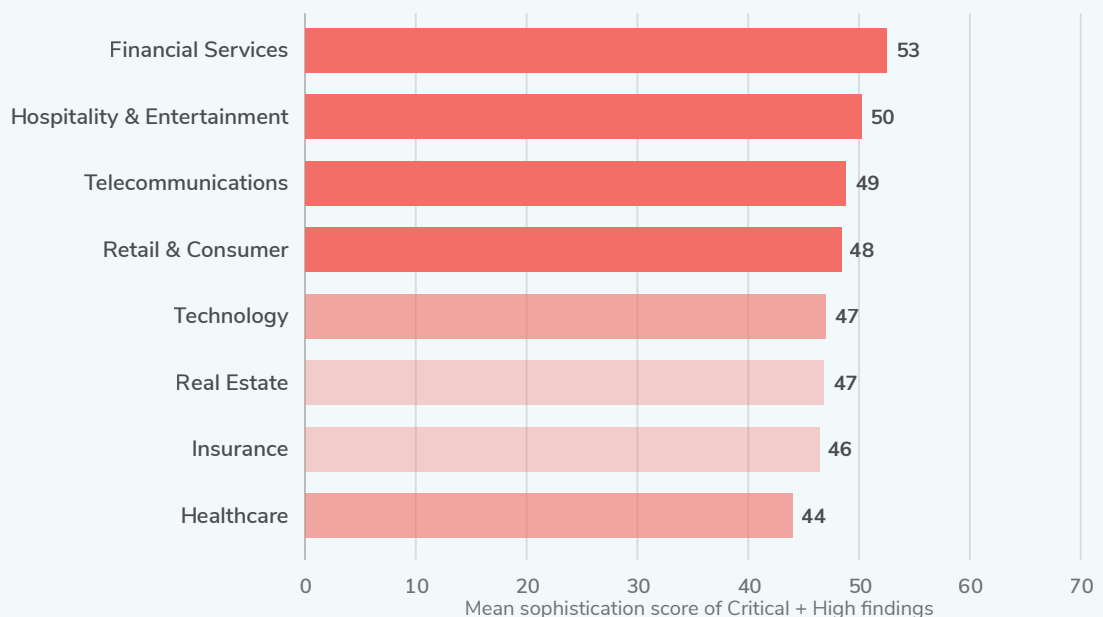


The mean sophistication score of critical- and high-severity findings is presented by OWASP 2021 category. Each bar is the mean of the findings we observed in that category and severity band. Dot markers to the left of each label indicate category volume across the whole dataset (●●● high, ●● medium, ● low), which is a signal for how much of the ranking rests on a substantial base of findings.

#### Kroll Observations

- Not all serious findings are equally difficult to identify. Access control failures require deep ingenuity to uncover. Cryptographic failures, even when serious, are much closer to what scanners can catch.
- Broken Access Control findings top the sophistication ranking among the higher-volume categories (with a mean score of about 54), while Cryptographic Failures findings sit near the bottom (with a mean score of about 31), showing a 23-point spread within the same severity band. The highest scores in the chart sit in low-volume categories in which a single unusual finding can move the mean substantially, so we treat the higher-volume comparison as the more representative one. The gap reflects a real difference in the kind of testing that tends to discover each type. A serious access control failure typically involves architectural reasoning and chaining observations, whereas a serious cryptographic failure is more often something visible enough for a well-tuned scanner to spot.

### 2.7.3 Sophistication of Serious Findings by Industry



The mean sophistication score of critical- and high-severity findings is organized for each industry. Bar shading indicates relative sample size within this cut (e.g., solid = larger sample, faded = smaller sample). Industries with fewer than 50 engagements are omitted. The Financial Services category shows the highest sophistication of serious findings of any industry with a substantial engagement history.

## Kroll Observations

- Industries with the lowest vulnerability density (\$2.2) also produce the most sophisticated, serious findings in this slice.
- Financial services leads both rankings, with the lowest density of findings per report and the highest mean sophistication (about 53) among the serious findings. The telecommunications and hospitality and entertainment industries are close behind for both.

**Penetration tests of financial services firms typically find the fewest vulnerabilities, and the vulnerabilities they do find are the most difficult to identify because exploiting them requires more sophistication.**

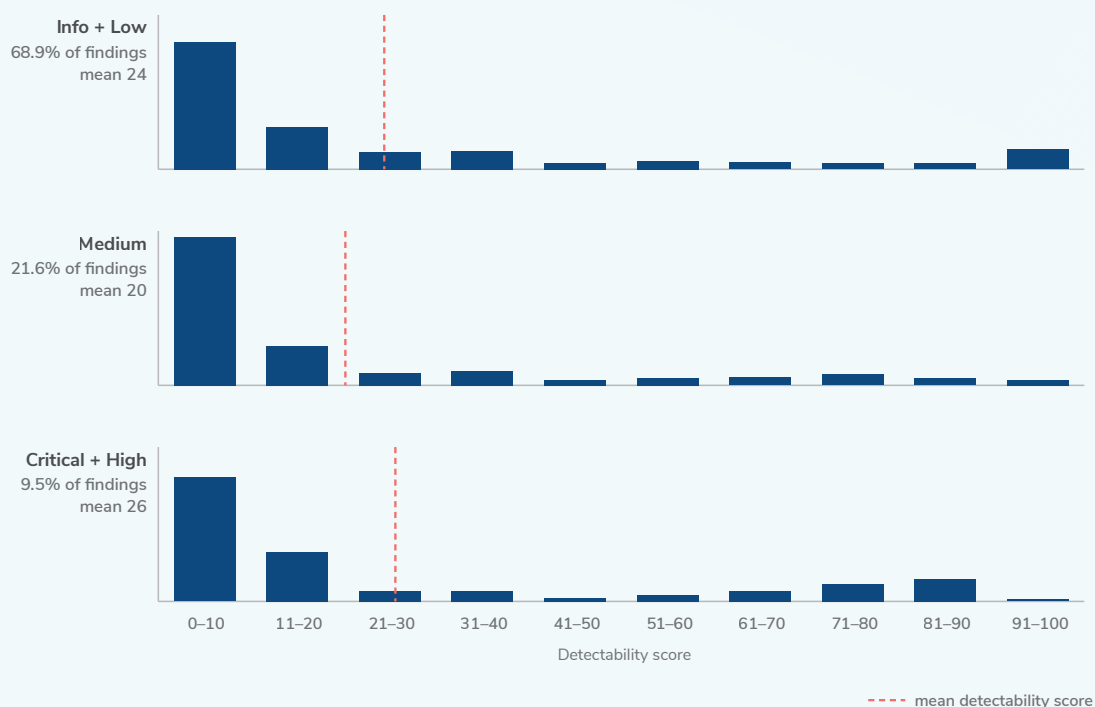
## 2.8 SAST/SCA Detectability

Each finding is scored on a 0–100 detectability scale that measures how likely it is that a SAST/SCA tool would have caught it earlier in the SDLC. Findings with a code-visible signature (e.g., dependency CVE, injection patterns or unsafe library calls) score at the high end of the scale. Findings that depend on runtime state, business logic or architectural context score at the low end.

The score complements finding sophistication. Sophistication ratings are used to assess who could have discovered a finding after deployment (e.g., a scanner running against the live target or a human). Detectability ratings are used to assess whether the finding could have been prevented earlier, before the target was deployed at all, through static analysis of code and dependencies during development. Together, the two scores describe the reachability of a finding across the SDLC from opposite ends.

As with our Vulnerability Introduction Point, a small share of findings (about 1.5%) could not be reliably scored from the write-up and were excluded from the following analysis.

### 2.8.1 Detectability Distribution Across Severity Bands

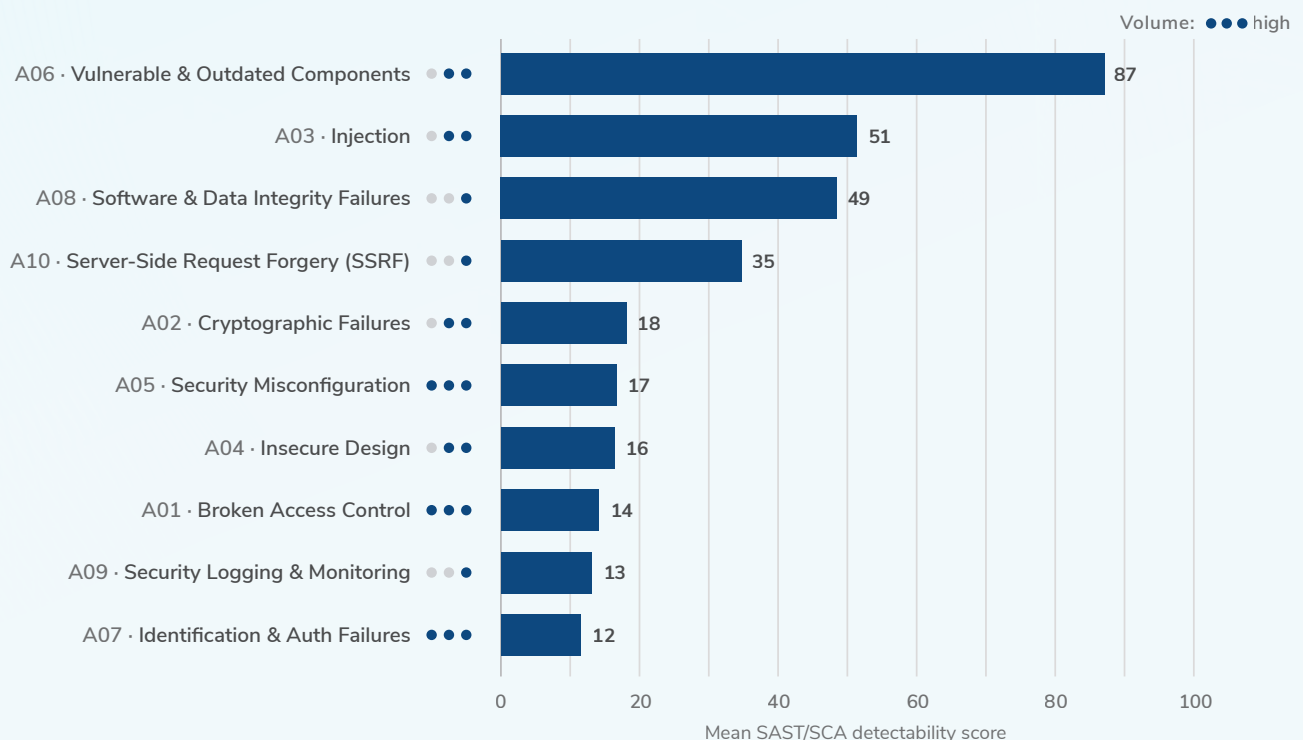


The distribution of SAST/SCA detectability scores is shown separately for three severity groupings. Each bar's height indicates the share of that severity group's findings that falls in the corresponding 10-point detectability bucket. Dashed red vertical lines mark each group's mean score. Findings scoring 80 or above (i.e., the two rightmost buckets in each panel) represent about 9% of the population and mark the range where SAST/SCA tooling would reliably have caught the finding earlier.

## Kroll Observations

- **Most of what penetration tests find is inherently dynamic. Only about 9% of findings have a detectability score high enough to indicate that a SAST/SCA tool would reliably have caught them earlier. The rest, to varying degrees, are invisible to static analysis.** This is not a comment on SAST/SCA quality; modern static tools do what they do well. It is a comment on what penetration testing discovers, which are findings that depend on runtime state, business logic, deployment context or architectural reasoning that a static tool cannot reach by design. Penetration tests and static analysis are largely non-overlapping instruments, and their outputs reflect that.
- **Severity is largely independent of detectability.** The three severity groups show broadly the same distribution shape and mean. A common intuition is that a well-invested SAST/SCA program would preferentially catch the more serious findings. However, our data does not support that. Severity is a property of consequence (how much damage a finding can cause in context), and detectability is a property of pattern (whether the finding's signature is code visible). A dependency CVE is easily detectable whether it turns out to be critical or low, but an authorization flaw is undetectable regardless of what data it exposes.

### 2.8.2 Mean Detectability by OWASP 2021 Category



The mean SAST/SCA detectability score of critical- through informational-severity findings is separated by OWASP 2021 category. Dot markers to the left of each label indicate category volume across the whole dataset (●●● high, ●● medium, ● low).

## Kroll Observations

- **SAST/SCA works well if the finding's signature is code visible but provides essentially no coverage if the finding is a matter of authorization, authentication or design.** Vulnerable and Outdated Components, which is a category of known CVE dependencies that SCA is built to detect, tops the ranking with a mean score of about 87. Injection, which is the category that SAST is built to detect, ranks second, with a mean score of about 51. At the other end, Broken Access Control and Identification and Authentication Failures score at the bottom (with means of 14 and 12, respectively) because they are runtime properties of the application, not code shapes that a static tool can inspect. Investment in SAST/SCA buys strong coverage in a few specific categories and essentially none in others.

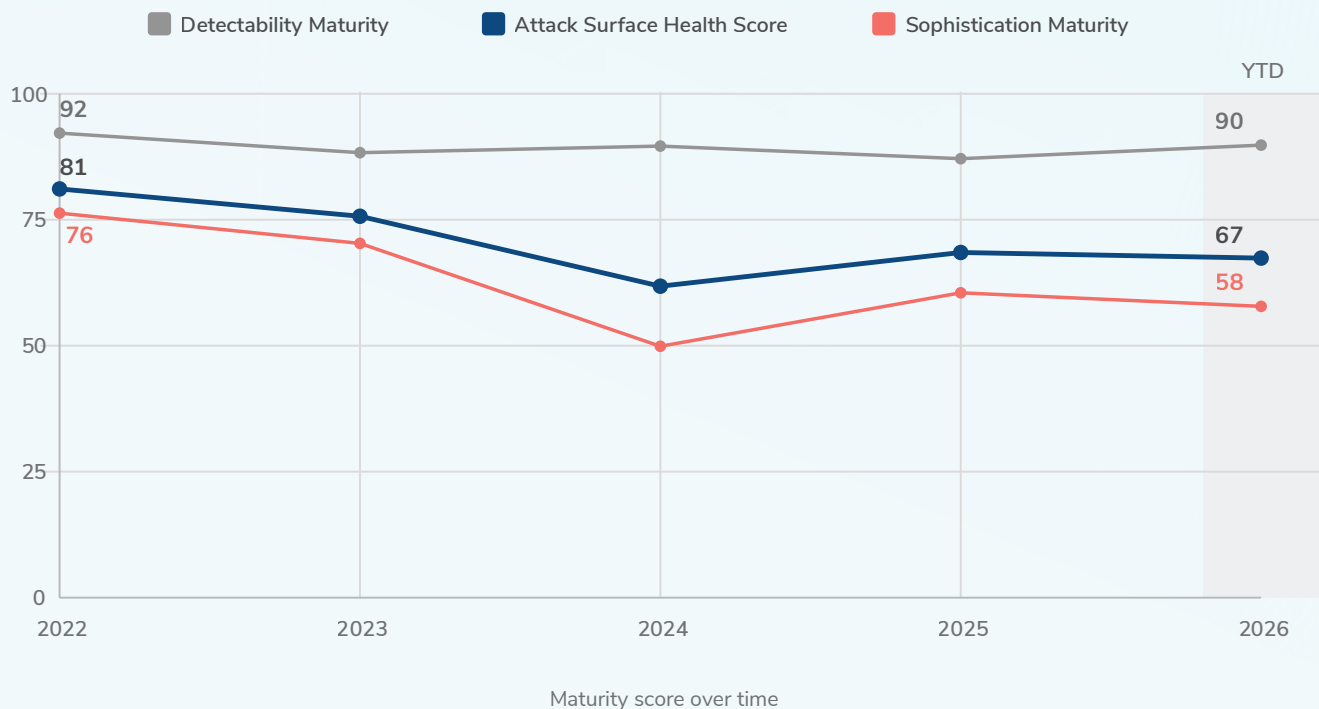
## 2.9 Attack Surface Health Score

The Attack Surface Health Score brings together the other metrics in this report (i.e., severity, sophistication, static detectability and finding density) into a single measure of what the penetration test surface looks like. Averaged across the entire dataset, the score sits at 69 out of 100, which is midrange, with room to improve, and no signs of catastrophic failure. The individual score number is less useful than the trajectory. The score has drifted from around 81 in 2022 to around 67 in 2026, and that shift is where the story lives.

The score combines two component maturities, detectability and sophistication, weighted 30% and 70%, respectively. Sophistication is weighted more heavily because although detectability captures what a specific class of tooling can reach, sophistication captures a broader signal about how much residual attack surface remains visible to any competent tester. Both components are normalized to control for engagement volume.

The score is derived only from penetration test outcomes. It does not incorporate security governance, policy maturity, staff awareness or blue-team tooling. It is best understood as an observed maturity measure (i.e., what the penetration test surface looks like from the attacker's side of the engagement) rather than a complete evaluation of a security program.

### 2.9.1 The Two Components Over Time



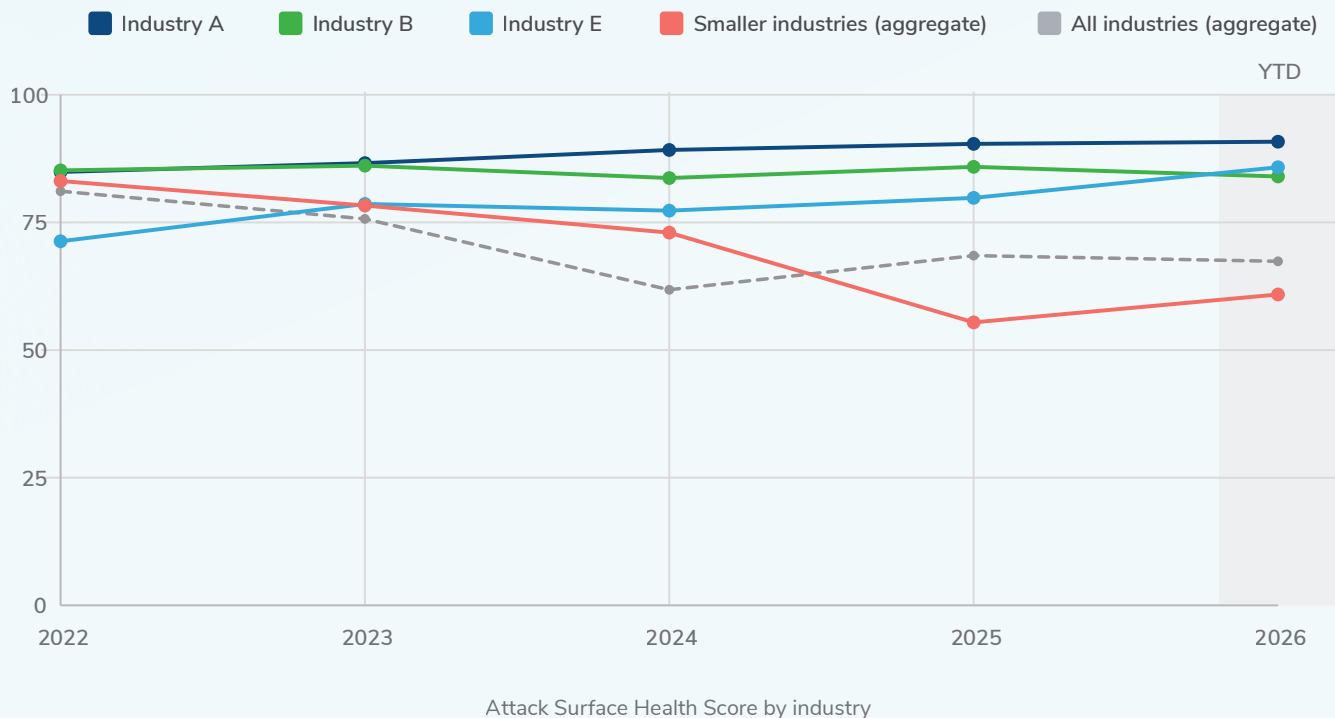
The Attack Surface Health Score (black) shown alongside its two components across the reliable observation window (2022–2026). Detectability maturity sits high and moves within a narrow band of about five points. Sophistication maturity has declined from 76 to 58 over the same period. The composite score tracks primarily with sophistication, reflecting its 70% weighting in the score's formula, falling from 81 in 2022 to 67 in 2026.

## Kroll Observations

- The industry is holding in some ways and weakening in others:
  - **Detectability maturity has remained broadly stable.** The score has moved within a narrow band (87–92) across the reliable observation window. The specific slice of surface that SAST/SCA tooling can reach has neither gained ground nor lost it.
  - **Sophistication maturity has moved meaningfully lower**, from 76 in 2022 to 58 in 2026. The wider surface that requires human testing to identify has become harder to secure.
  - **The composite score has therefore tracked with sophistication.** Because sophistication carries 70% of the composite's weight, the Attack Surface Health Score has moved almost entirely because of the sophistication decline. Detectability has been essentially inert relative to the overall trajectory.
- **Why this shift has happened is an open question that penetration test data alone cannot definitively answer.**

**Detectability of vulnerabilities remains fairly consistent, and sophistication of vulnerabilities to exploit is declining.**

### 2.9.2 Attack Surface Health Score Trajectory by Industry



The Attack Surface Health Score trajectory for 2022–2026 for three heavily tested industries is shown (individually with an additional aggregate line for the six smaller industries with data in each of the five years). Individual industries are anonymized, with the whole-industry aggregate shown as a dashed gray reference.

## Kroll Observations

- **The aggregate decline masks a diversity of trajectories.** To keep the focus on trajectory shapes rather than industry comparison, individual industries are anonymized, with sample size rank preserved so the reader can see how representative each example is:
  - **Industry A** (largest sample in the dataset) shows a steady climb (from 85 to 91), with improvements in each of the five years.
  - **Industry B** (second-largest sample) has held essentially flat, maintaining a score of 85 across the window.
  - **Industry E** (one of the top five by sample size) has moved upward but with meaningful year-to-year variation (moving from a score of 71 to 86).
  - **Smaller industries as a group** (i.e., six industries with consistent multiyear data) have declined together over the same period, moving from a score of 83 to 61.
- The whole-industry aggregate sits between these lines, which is closer to the smaller-industries line because there are more of them in the pool. The aggregate describes the industry mix and not any specific industry.

## 2.10 Persistent Problems

Some vulnerability classes recur every year, with similar frequency, regardless of which clients we test. We group findings by their specific CWE category (using the CWE catalog directly rather than the derived taxonomies in Section 2.5) and calculate each category's share of that year's findings. Averaged across years, with each year weighted equally, we generate an Annual Recurrence score. A higher score means that the category shows up consistently each year at a meaningful volume.

### 2.10.1 Top Persistent CWEs, at Each End of the Sophistication Scale

| Quick wins (Very low sophistication tier)                               | Needs attention (Critical and high-severity in all sophistication tiers) |
|-------------------------------------------------------------------------|--------------------------------------------------------------------------|
| CWE-200 · Exposure of Sensitive Information— <b>12.5%</b>               | CWE-639 · Authorization Bypass Through User-Controlled Key— <b>10.7%</b> |
| CWE-693 · Protection Mechanism Failure— <b>9.5%</b>                     | CWE-79 · Cross-site Scripting— <b>8.4%</b>                               |
| CWE-327 · Use of a Broken or Risky Cryptographic Algorithm— <b>8.0%</b> | CWE-306 · Missing Authentication for Critical Function— <b>6.5%</b>      |
| CWE-1104 · Use of Unmaintained Third Party Components— <b>7.9%</b>      | CWE-269 · Improper Privilege Management— <b>5.8%</b>                     |

The top CWE categories are shown by Annual Recurrence score at each end of the sophistication scale. The left column lists findings that require little penetration tester ingenuity to identify. The right column lists findings that carry critical- and high-severity risk. Percentages are the average share of a year's findings for each CWE account.

## Kroll Observations

- Persistent problems come in two shapes:
  - The low-sophistication findings that recur year after year are almost entirely hygiene issues: information leaks, missing headers, weak cryptographic configurations and outdated dependencies. Fixing them is well-understood work, and the persistence is consistent with incomplete deployment discipline rather than a technical challenge.
  - The findings that recur at the higher-severity end are almost entirely architectural: authorization gaps, missing authentication and privilege management. These are more difficult to fix, and the persistence is consistent with architectural remediation moving slower than configuration remediation.
- The two lists barely overlap, which is the key observation. The industry is fighting two distinct populations of persistent problems, and the tools and disciplines that address one are largely not those that address the other.

**Persistent critical- and high-severity vulnerabilities are almost entirely architectural: authorization gaps, missing authentication and privilege management. These are more difficult to fix.**

## SECTION 3

# Looking Ahead

### 3.1 What the Next 12–24 Months Look Like

The next 12–24 months will bring meaningful changes to the world of offensive security, particularly within the sphere of application security assessments. AI is likely the most consequential vector regarding both how software is written and how it is analyzed. Other emerging areas warrant attention as well.

#### AI and Large Language Models (LLMs) in Application Security

- **AI-enhanced SAST/SCA tooling:** Current static tooling has a plateau. It does what it is designed to do but does not extend into identifying design-stage, authorization or authentication findings. Whether AI/LLM-enriched static analysis can reach previously unreachable finding classes is one of the more interesting open questions. If it can, the SAST/SCA plateau argument may look different in future editions of this report.
- **AI-assisted coding upstream:** ChatGPT debuted in late 2022. AI-assisted coding tools have reached wide adoption since then. As AI-generated code becomes more prevalent in production applications, the vulnerability profile that penetration tests observe will likely shift in severity, detectability, sophistication or some combination of all these factors. Our current data does not yet show a distinct signal attributable to this trend, but we may be able to reflect these shifts in the future.

#### Other Emerging Areas

- **Evolving mobile threat landscape:** Our study pools web and mobile findings together, but we could look to split these application types in the future.
- **Supply chain and third-party component risk:** This risk is connected with the observations made in §2.5 regarding the diminished but not eliminated presence of memory safety issues and in §2.6 regarding the Vulnerable and Outdated Components category. The economic pressure on transitive dependency risk is likely to grow.

### 3.2 What Our Data Will Track Going Forward

This is intended as a first report on the state of the application security market through the lens of our penetration testing data. In future analysis, several aspects of the methodology and coverage will evolve:

- **Sample composition maturity:** As Kroll's client base grows, the finding population will better represent the broader industry. For example, some industry contrasts were anonymized by necessity due to sample size constraints in smaller industries.
- **Attack Surface Health Score refinement:** The current composite (30% detectability, 70% sophistication, normalized by density) is a first version. Future editions will consider adding factors. Candidates include finding persistence, remediation cadence and category diversity across a client's finding profile. The metric's goal is to describe the penetration test surface honestly, and it should evolve as our understanding improves.
- **Additional metrics:** Cuts of the data we chose not to publish this edition include repeat-offender findings across retests, remediation trajectories, engagement-level observations (e.g., duration, scope and first-time testing versus mature programs) and mobile-versus-web splits.

# About the Firm

## Kroll is a Leader in Cyber and Data Resilience

Global leader in IR with **thousands of incident response engagements annually**

Preferred vendor for **85+** insurance carriers

**MDR exclusively** delivered on CrowdStrike Falcon® Complete

Frontline experience drawn from **Government, Law Enforcement, Intelligence and Fortune 500 environments.**



**700+** experts across 19 countries

**100+** certifications



Expertise in **AI, IAM, Cloud, Data Analytics, OT security and data risk**

**500+** MDR clients

**700k+** actively monitored endpoints

Rated as **industry leaders**



## CONNECT TO LEARN MORE



**Krish Raja**  
Managing Director  
Global Head of  
Offensive Security



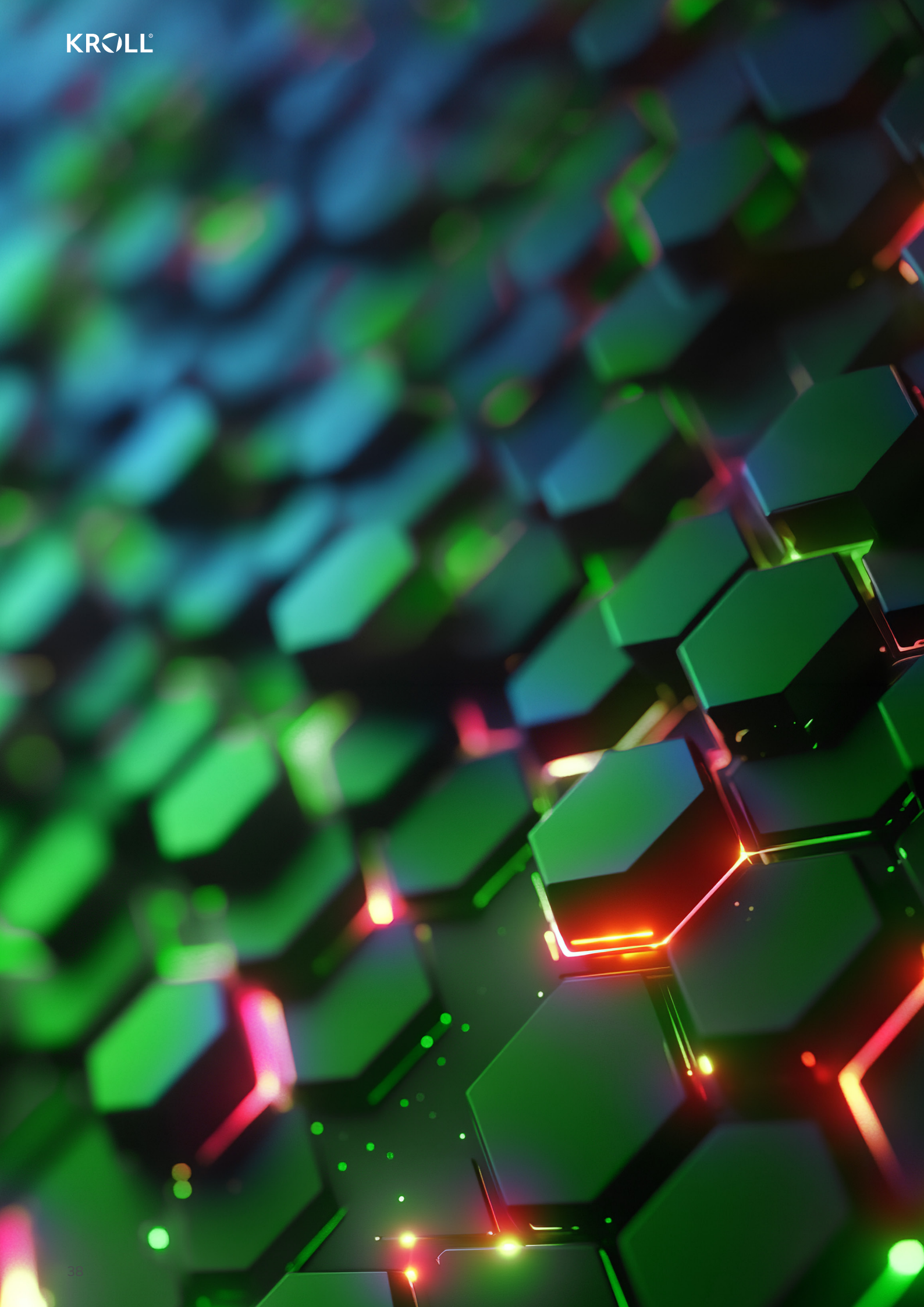
**Rahul Raghavan**  
Director  
Offensive Security



**Alex Cowperthwaite**  
Director  
Offensive Security



**Felix Ang**  
Vice President  
Offensive Security



# Across 30+ Countries and Territories



## The Americas

|              |                |
|--------------|----------------|
| Atlanta      | Nashville      |
| Austin       | New York       |
| Bogotá       | Philadelphia   |
| Boston       | Richardson     |
| Buenos Aires | San Francisco  |
| Chicago      | São Paulo      |
| Dallas       | Seattle        |
| Hamilton     | Secaucus       |
| Houston      | Sunnyvale      |
| Los Angeles  | Toronto        |
| Mexico City  | Washington, DC |
| Miami        |                |
| Morristown   |                |

## Caribbean

British Virgin Islands  
Cayman Islands

## Europe, Middle East and Africa

|                |           |
|----------------|-----------|
| Abu Dhabi      | Mauritius |
| Agrate Brianza | Milan     |
| Amsterdam      | Munich    |
| Berlin         | Padua     |
| Birmingham     | Paris     |
| Cape Town      | Riyadh    |
| Dubai          | Rome      |
| Dublin         | Tel Aviv  |
| Frankfurt      | Turin     |
| Gibraltar      | Zurich    |
| Jersey (CI)    |           |
| Johannesburg   |           |
| Leeds          |           |
| Lisbon         |           |
| London         |           |
| Luxembourg     |           |
| Madrid         |           |
| Manchester     |           |

## Asia Pacific

Bangalore  
Beijing  
Christchurch  
Guangzhou  
Hanoi  
Hong Kong  
Hyderabad  
Jakarta  
Kuala Lumpur  
Manila  
Melbourne  
Mumbai  
New Delhi  
Shanghai  
Shenzhen  
Singapore  
Sydney  
Taipei  
Tokyo



---

#### About Kroll

Kroll is a global financial and risk advisory firm and the definitive authority at the intersection of valuation, risk and transactions, trusted by boards, investors and regulators when decisions matter most. Our team of more than 6,500 professionals worldwide combines financial advisory rigor with investigative and risk intelligence capability, delivering the trusted expertise clients need when the stakes are highest. Learn more at [Kroll.com](https://kroll.com).

M&A advisory, capital raising and secondary market advisory services in the United States are provided by Kroll Securities, LLC (member FINRA/SIPC). M&A advisory, capital raising and secondary market advisory services in the United Kingdom are provided by Kroll Securities Ltd., which is authorized and regulated by the Financial Conduct Authority (FCA). Valuation Advisory Services in India are provided by Kroll Advisory Private Limited (formerly, Duff & Phelps India Private Limited), under a category 1 merchant banker license issued by the Securities and Exchange Board of India.

**Kroll® is a trademark registered in the United States and 50 other jurisdictions, with applications pending in additional countries.**

© 2026 Kroll, LLC. All rights reserved. R26092032