State of web scraping report 2026

apify

The Web Scraping Club

In December 2025, we conducted a survey to understand the state of web scraping in 2026, asking both the Apify and The Web Scraping Club communities how people are scraping today, what's changed compared to last year, and what the outlook is for 2026.

In this report, we're sharing insights directly from our survey, extracted from thousands of answers to our questions. Spoiler: Web scraping is a strong, growing, and evolving industry.

Table of contents

How we approached the survey

Hundreds of professionals filled out the survey, including both closed and open questions, which covered four main areas:

1. **Proxy usage:** We asked professionals about the types of proxies they use, their spending, the number of proxy providers they use, and more.
2. **Infrastructure:** For this area, we asked about programming languages, frameworks, and whether they rely mainly on custom and internal solutions, rather than commercial ones.
3. **Bot detection:** This area focuses on anti-bot techniques and the challenges users face when dealing with anti-bots during their scraping activities.
4. **AI scraping:** This last area focuses on questions about whether professionals use AI scraping tools, what they like about them, and the challenges they overcome using them.

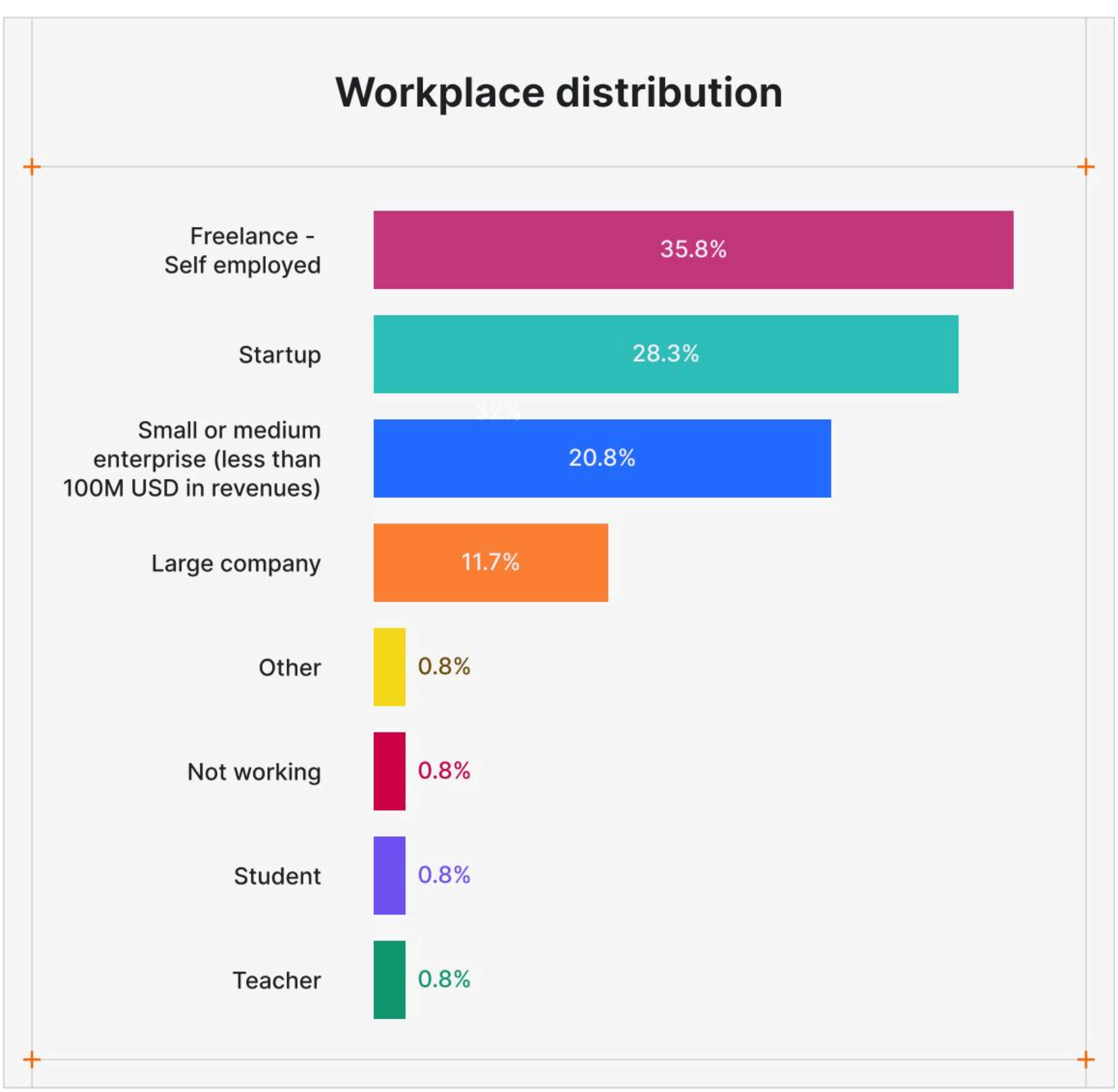
As a result, we've covered a few significant trends, some largely expected and others surprising.

Market overview

Before looking at the results from each of the four sections of the report, let's first analyze the respondents.

The data shows that the majority of web scraping professionals **(35.8%)** work as freelancers. Those working in startups represent **28.3%,** while those in small or medium-sized enterprises account for **20.8%.** If we aggregate this data, professionals working in startups and small enterprises account for **49.1%**, while those working in large companies account for **11.7%**.

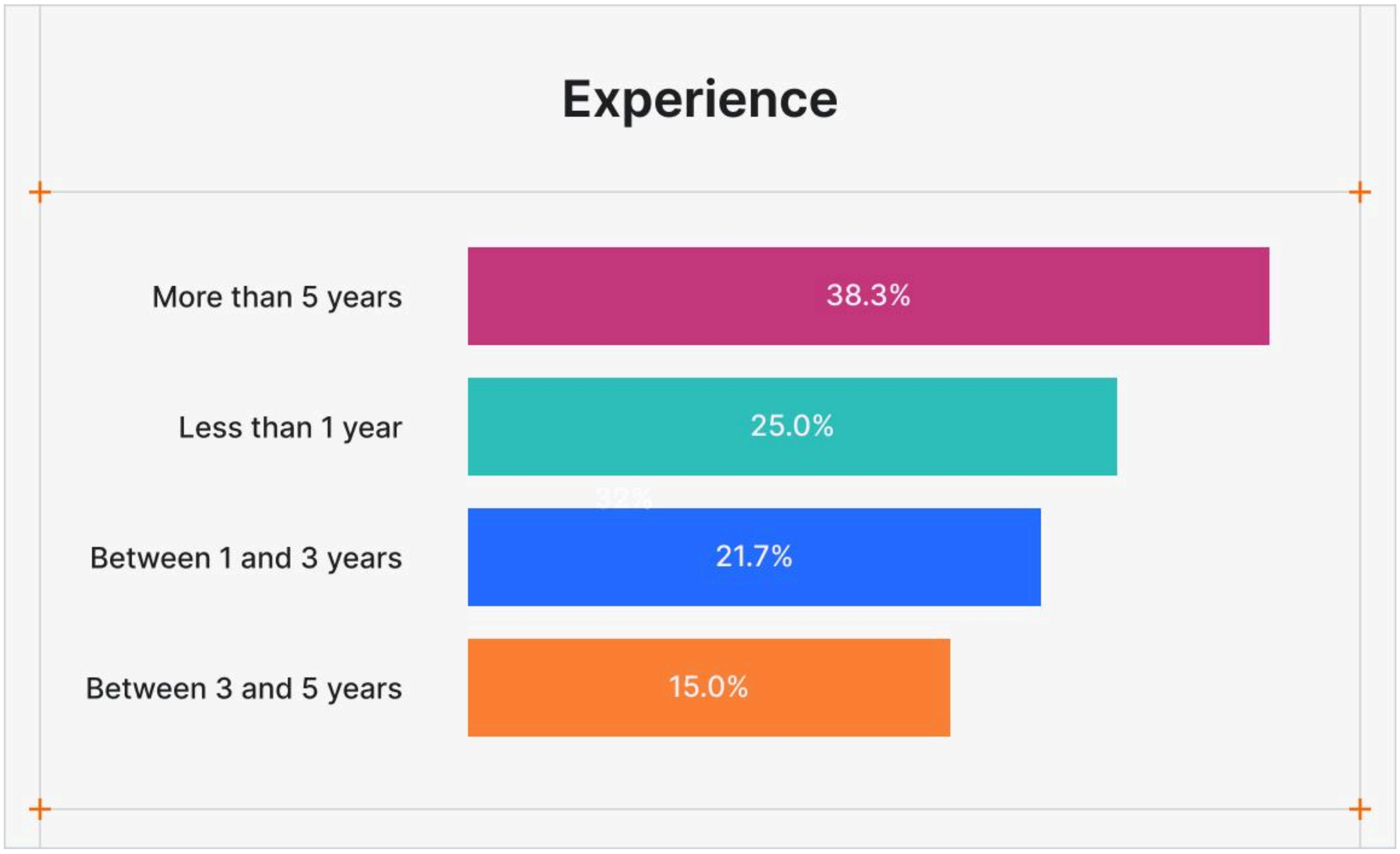
This indicates that the commitment of startups and small companies remains the most relevant in terms of hiring scraping professionals, compared to large enterprises. At a higher level, we can say that web scraping is used across businesses of all sizes.



When it comes to experience, the largest group **(38.3%)** has more than 5 years in the field.

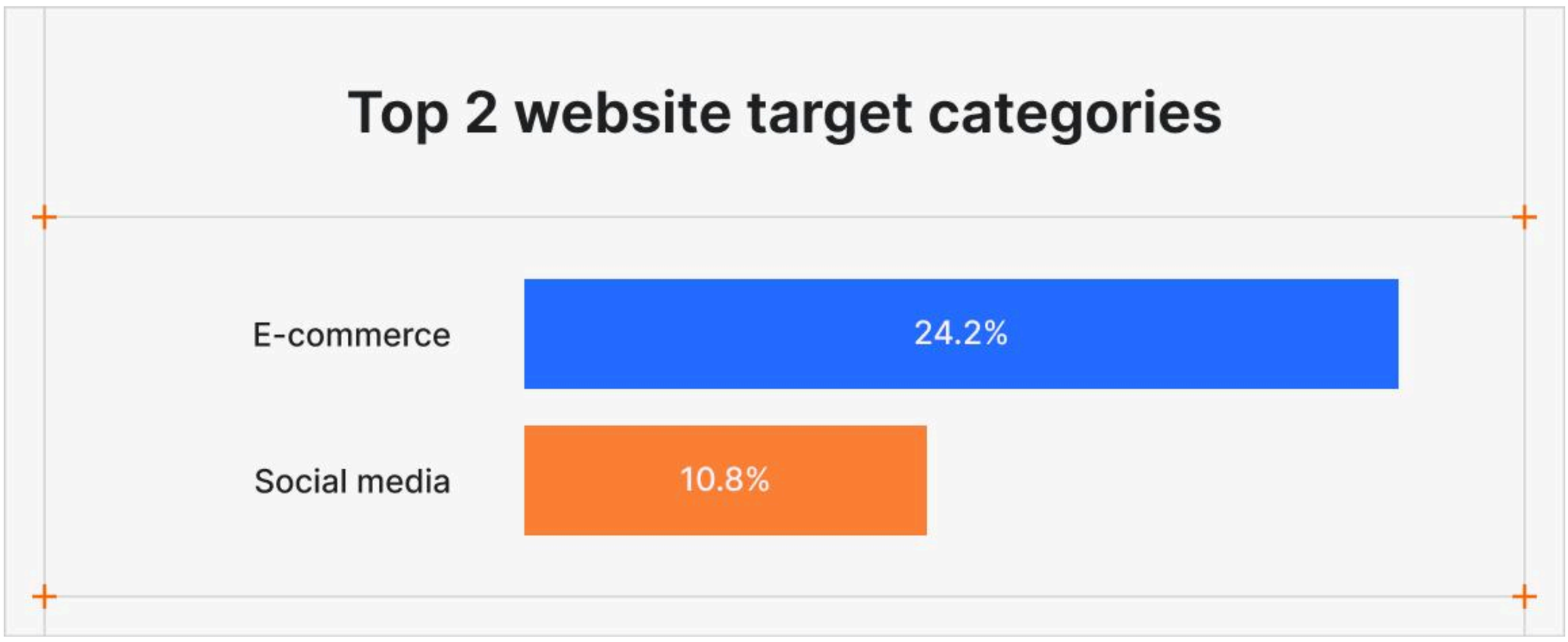
Professionals with less than a year of experience account for **25%**. That shows there's room even for entry-level roles in the web scraping industry.

21.7% have between 1 and 3 years of experience, while **15%** have between 3 and 5 years. Overall, **53.3%** have 3+ years of experience, making web scraping an industry that can create long-lasting careers.



As for websites that get scraped, the survey shows that professionals target websites in different industries. However, the majority works with e-commerce and social media platforms.

In particular, **24.2%** work exclusively with e-commerce websites, and **10.8%** only with social media. The remaining part works with both and other industries.



Now that we've understood the composition of the sample that answered the survey, we have more context for a deep dive into the four sections of our report.

1. Proxy usage

As you know, there's no web scraping without proxies. But we were curious to find out if professionals are using more proxies in their projects than they did last year

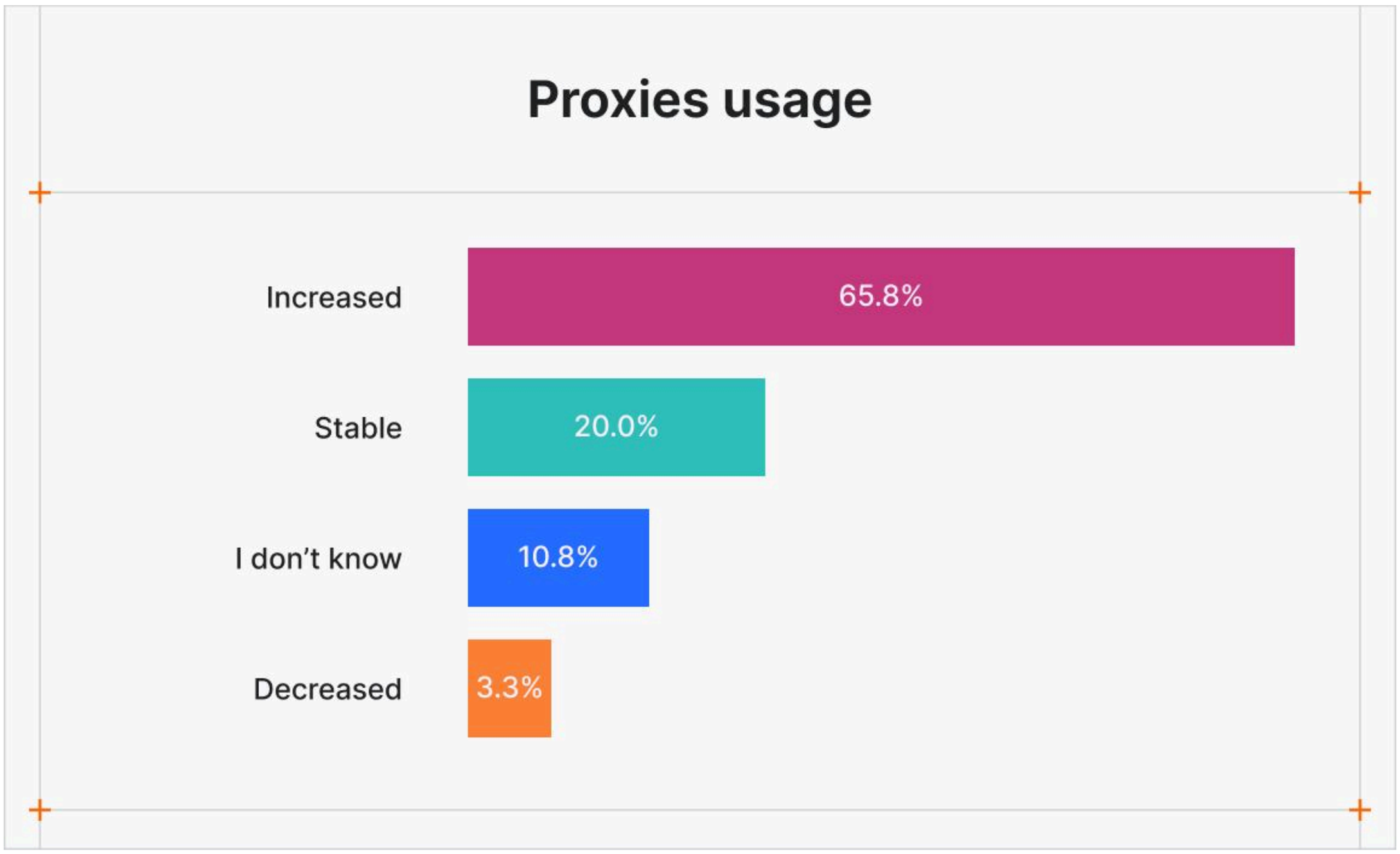
We haven’t specified whether consumption is in the number of requests or GBs. However, the results are the following:

65.8% of users reported using more proxies than they did last year.

20% declared that their proxy usage remains stable.

10.8% can't say what their proxy usage has been in the last year.

3.3% say their proxy usage has decreased.



Still, we weren’t satisfied with that information alone. That’s why we also wanted to know if scraping experts are spending more or less on proxies than they did last year:

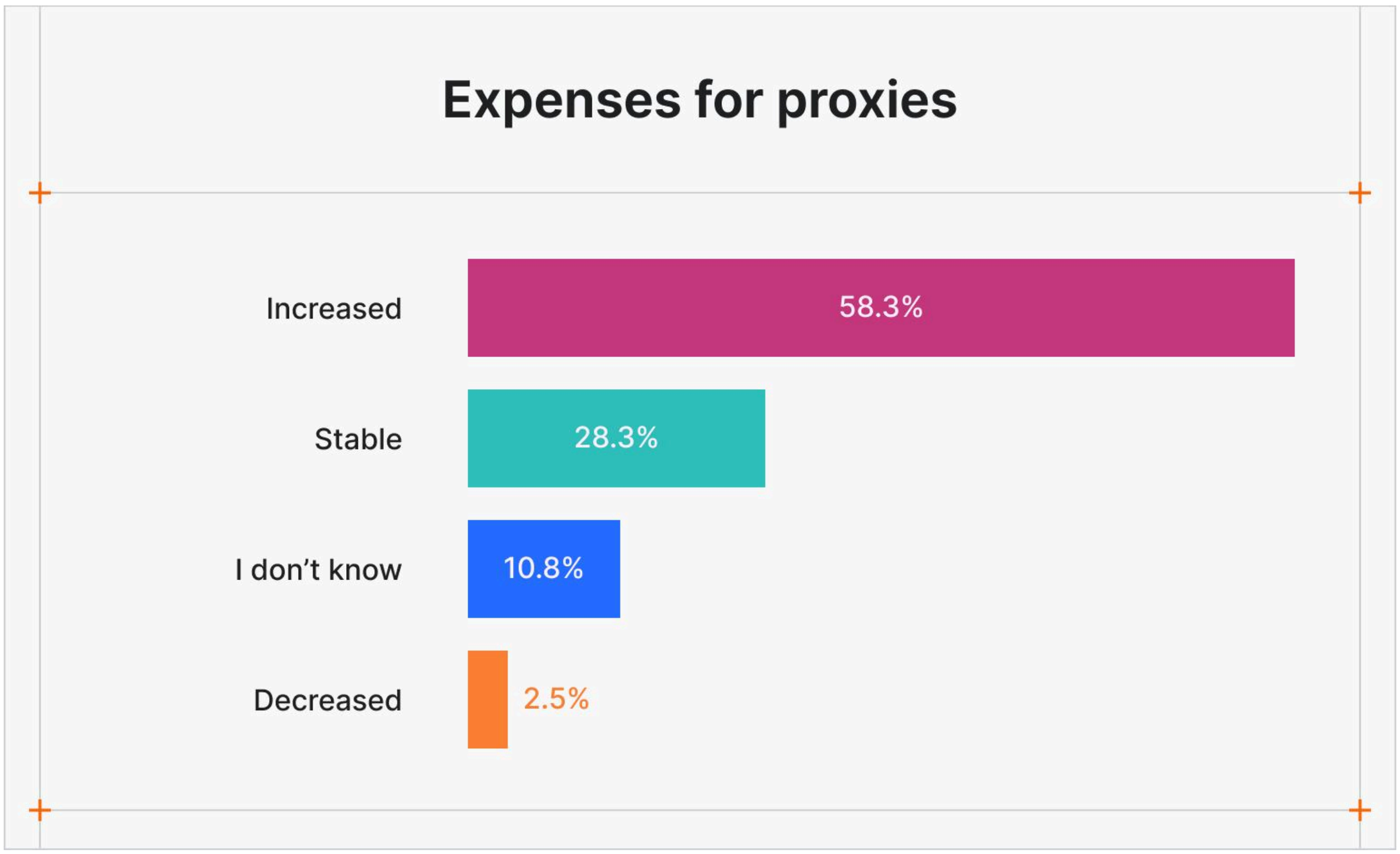
58.3% The vast majority says they've increased their expenses on proxies for their scraping projects in the last 12 months.

28.3% declared their expenses on proxies remained stable in the last year.

10.8% aren’t aware of the costs, which isn’t surprising, since not all employees have access to that kind of information

2.5% spent less than the previous year on proxies.

So, the trend is clear. Not only have they increased their proxy usage, but they've also spent more money on them. This is an interesting outcome: over the past few years, the prices of almost every type of proxy have fallen, thinning the margins for proxy providers. Despite that, most survey respondents say they’re both using and spending more on them. This means the proxy demand has grown so much that it has compensated for the diminishing margins of the providers.



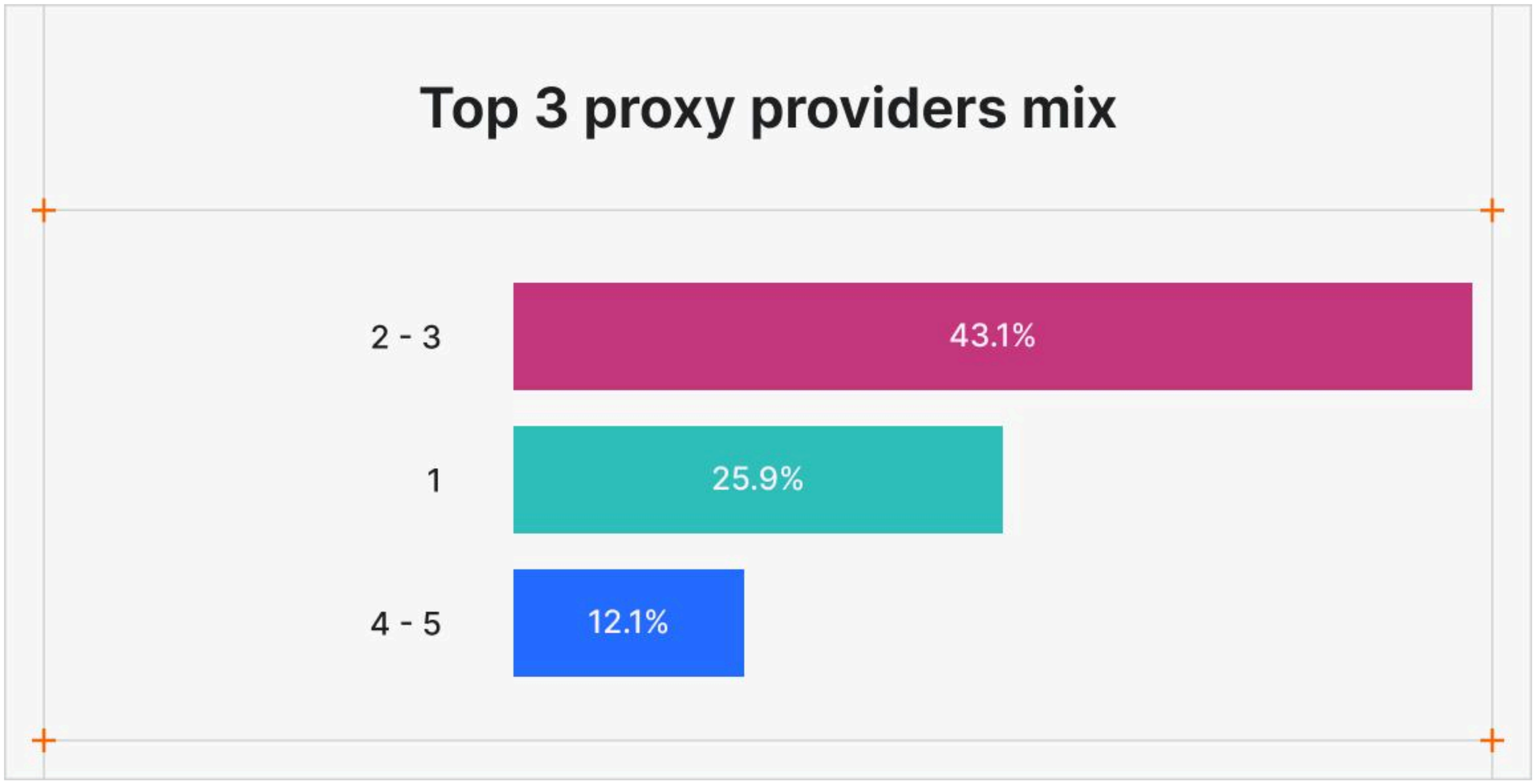
As there are several proxy providers on the market, we also asked if they purchase proxies from a single provider or multiple providers.

We gave several possibilities, and below are the top three responses:

43.1% of scraping professionals say that they're buying proxies from 2-3 proxy providers.

25.9% buy proxies from one provider.

12.1% use proxies that come from 4-5 different providers.



2. Infrastructure

In web scraping, infrastructure is often a significant expense for companies, both in terms of time and money. Still, it’s the basis on which data-powered companies build their future.

First, we asked professionals whether they develop their scraping infrastructure in-house or rely on external services:

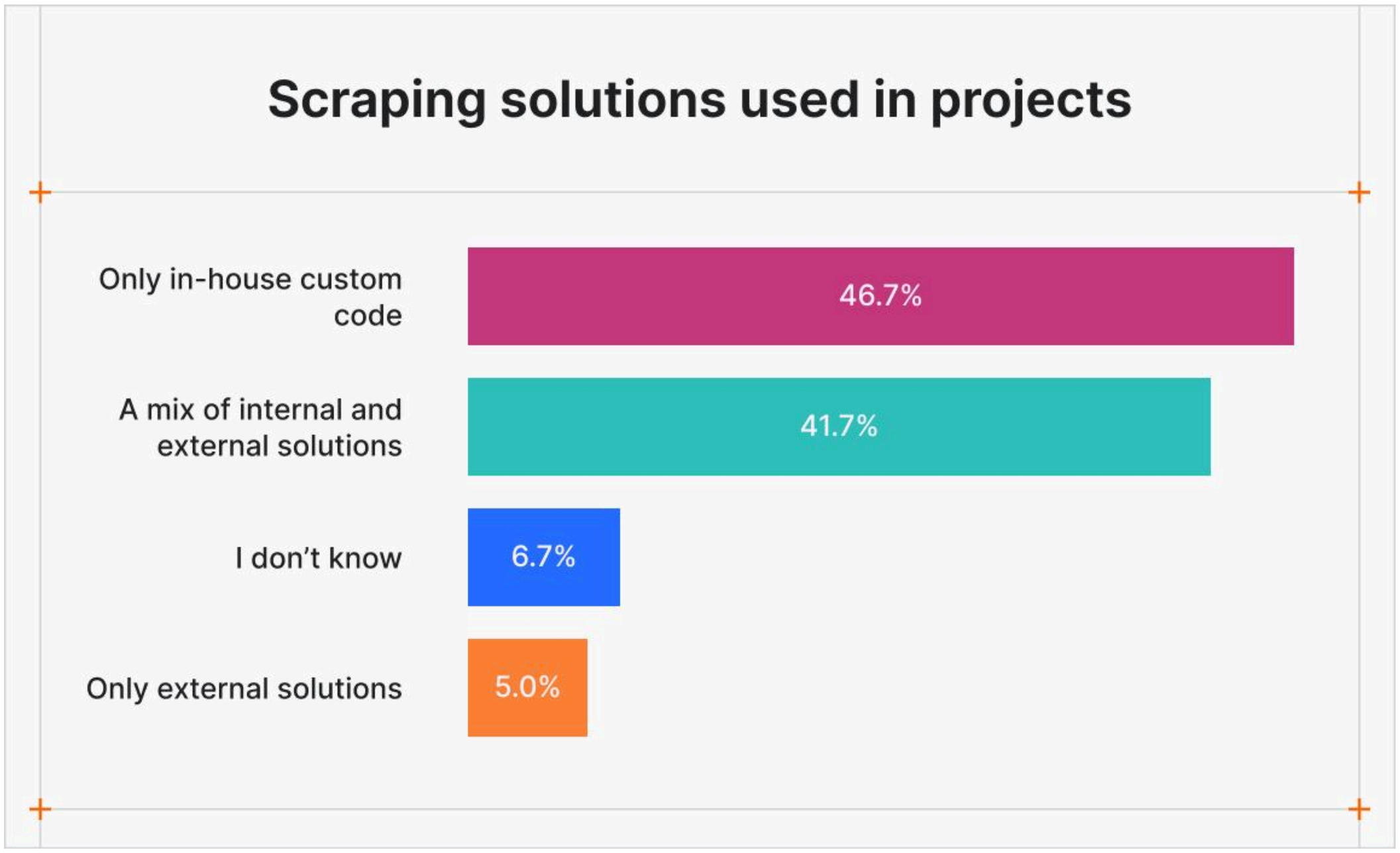
46.7% use only in-house custom code. This stresses once again that scraping is very often an industry where custom solutions are the majority of the cases.

41.7% use a mix of internal and external solutions.

6.7% aren’t aware of the underlying technical solution implemented by their company. Although that may seem surprising, some survey respondents may not be involved with this aspect of the web scraping process.

5% rely only on external solutions.

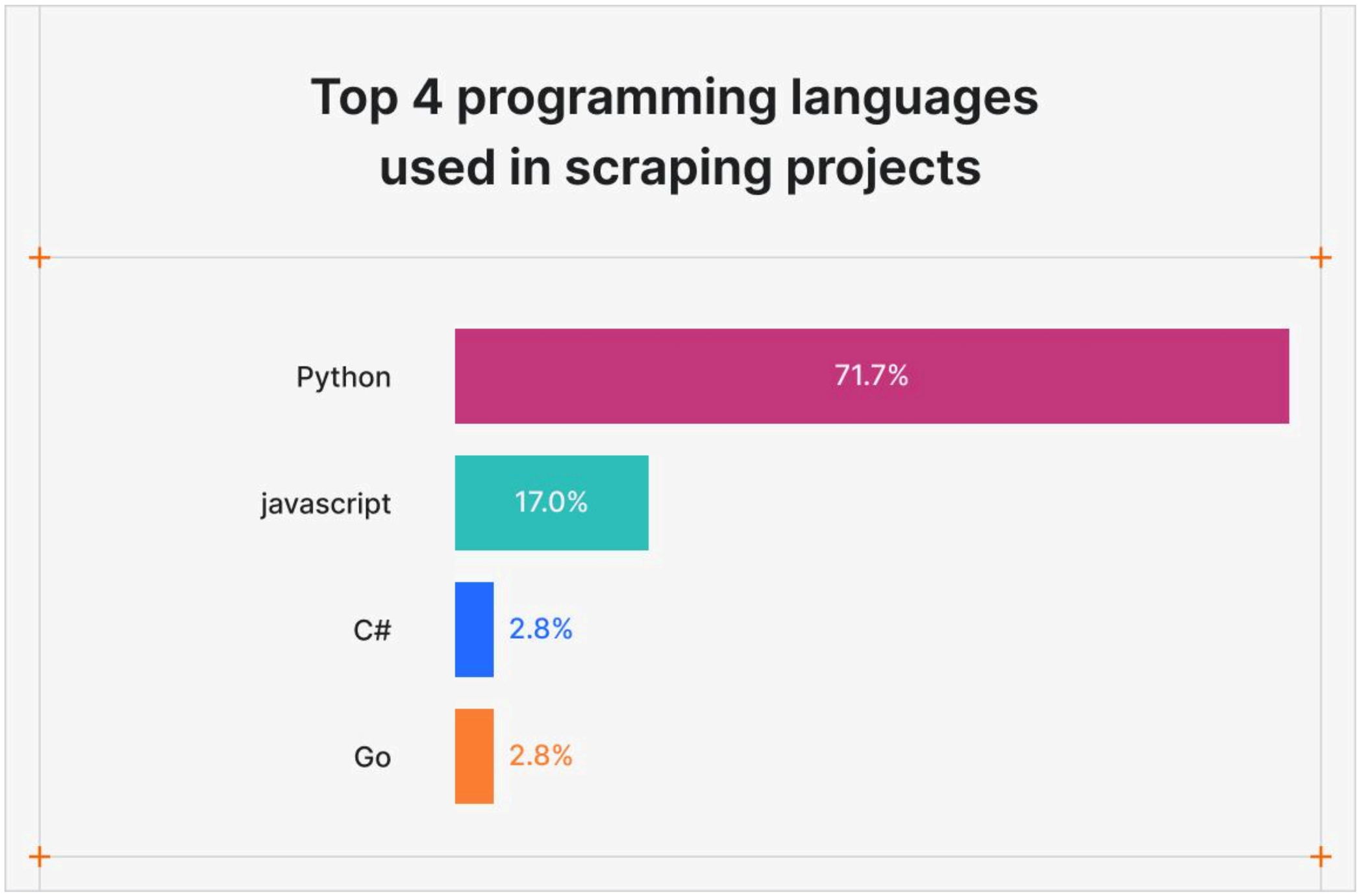
Given that this survey has been shared in the Apify and The Web Scraping Club communities, which are mainly composed of experts in web scraping, it was expected that most of the audience is building scraping tools internally rather than buying them externally.



As for programming languages used for scraping, the results are just as you'd expect:

- **71.7%** use Python, while **17%** prefer JavaScript.
- There are a few who develop in C# and Go, both at **2.8%**.
- Less than **6%** of users write code for their scraping projects in PHP, Ruby, or other programming languages.

The most used frameworks are Selenium, Puppeteer, Playwright, and Scrapy.



As we did for proxies, we asked the audience whether they had spent more or less on infrastructure over the past 12 months. By infrastructure costs, we mean the computing power for running scrapers (virtual machines, hosting) and all the connected running costs for scrapers, excluding proxies:

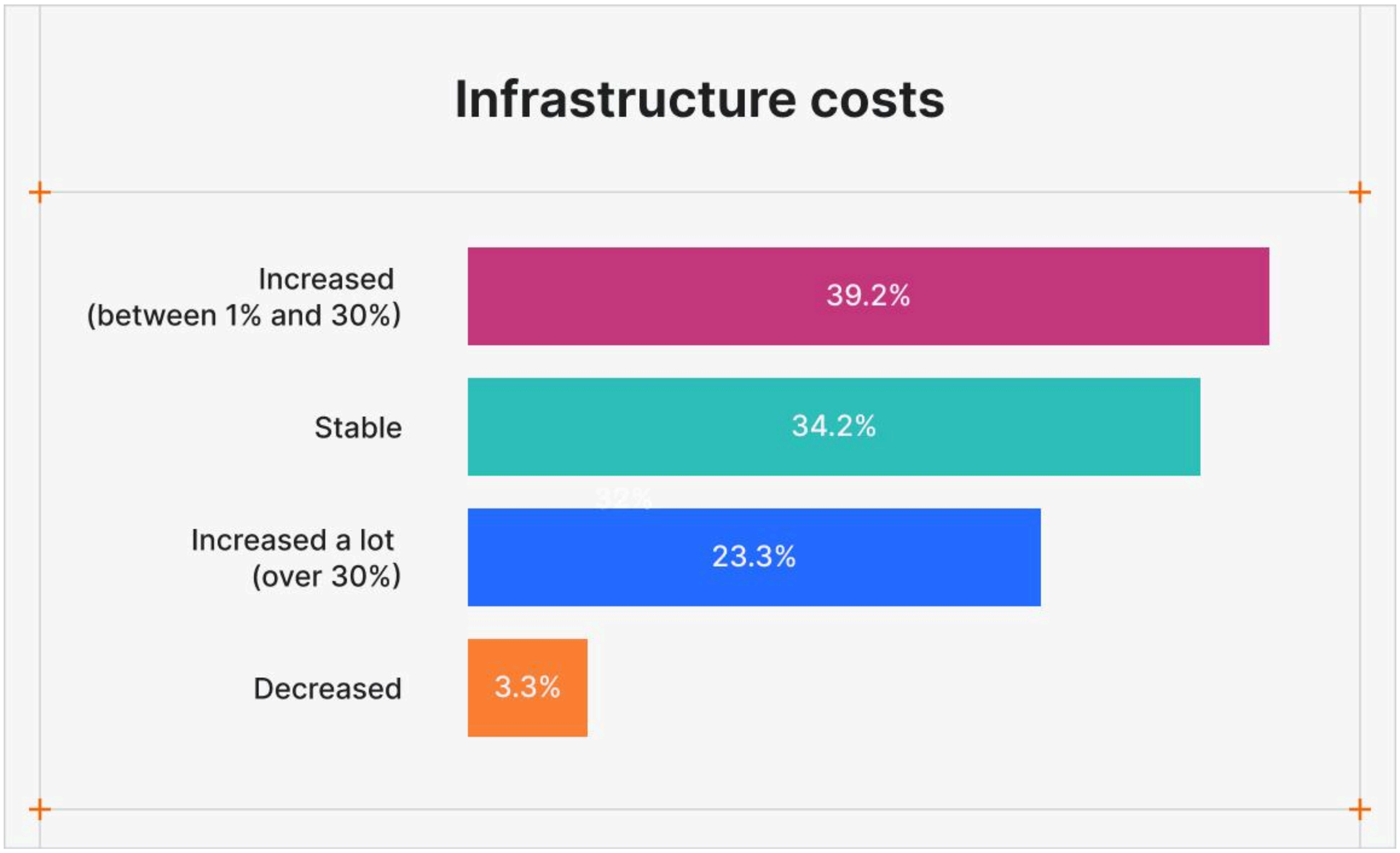
39.2% reported that their infrastructure costs increased by **1-30%** in the last 12 months.

34.2% declared their costs remain stable.

23.3% said that their costs increased more than **30%.**

For the 3.3% infrastructure costs decreased in the last year.

More than two-thirds of the audience noted that web scraping is becoming increasingly expensive over time. This is not surprising: antibot protections are becoming increasingly accurate, and bypassing them requires more effort. And thanks also to AI companies scraping the web in bulk, more websites are deciding to protect themselves from bots. We can see these trends in the next section dedicated to bot detection.



3. Bot detection

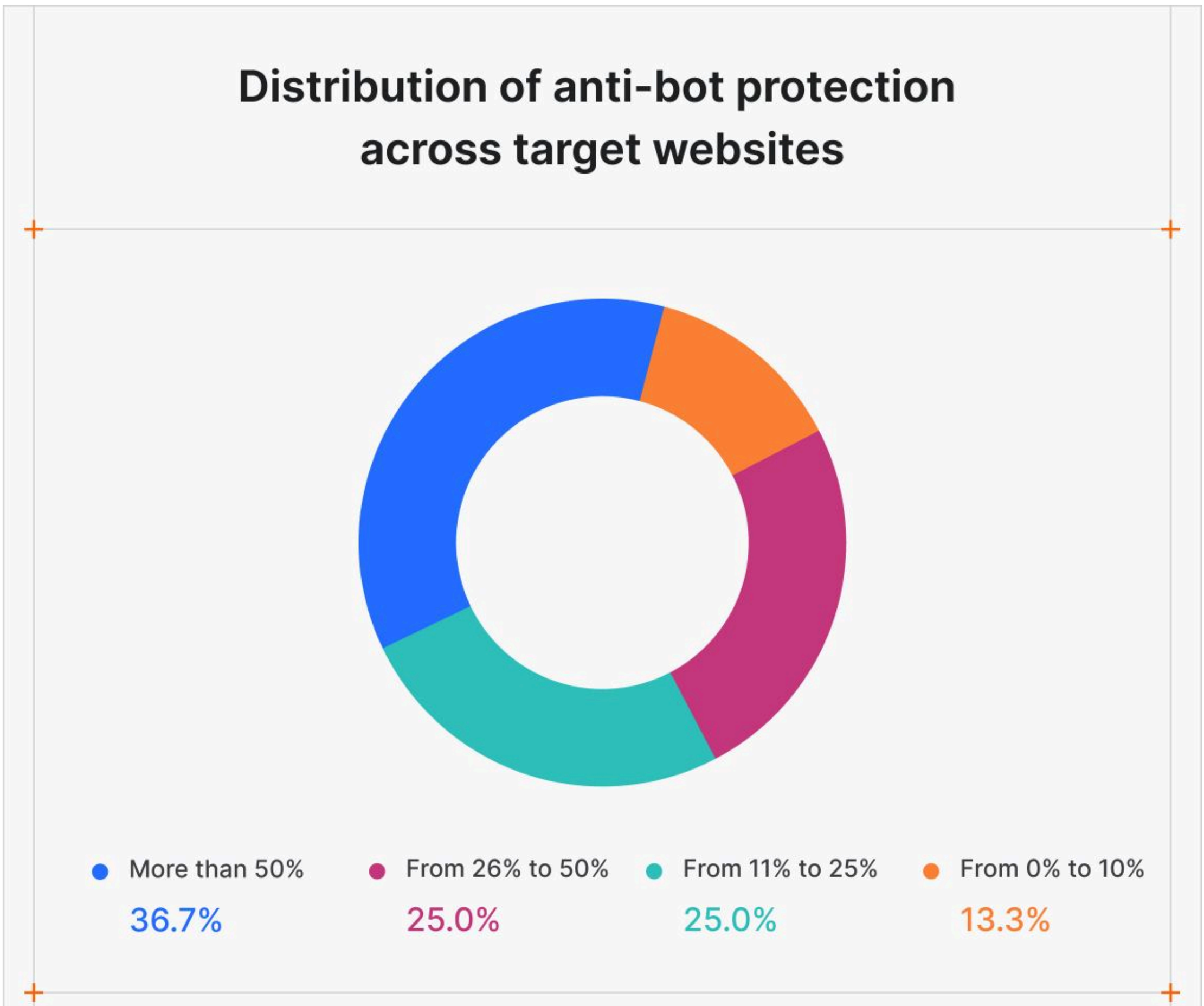
Bot detection is as old as web scraping, and anyone familiar with building automated web scrapers knows how challenging it can be due to anti-bot protections. In this section, we wanted to measure the impact of anti-scraping solutions in 2025 and how this is affecting web scraping operations.

To better understand the scope of those challenges, we asked about the percentage of protected websites targeted in their scraping activities:

36.7% indicated that more than **50%** of their target sites employ some form of anti-bot measures.

50% reported dealing with protected sites in **26–50%** and 11–25% of cases, evenly split at **25%** each.

13.3% reported that only **0–10%** of the sites they target are protected.



These findings indicate that a substantial portion of web scraping activity involves interacting with websites shielded by some anti-bot protections. This seems to contradict the DataDome report, which claims that less than 3% of top domains are protected by an anti-bot solution. We believe that our findings indicate that most scraping activity is concentrated on well-known sites, which, in response, implement anti-bot measures.

In fact, we asked respondents to specify the ones they currently find most challenging to scrape.

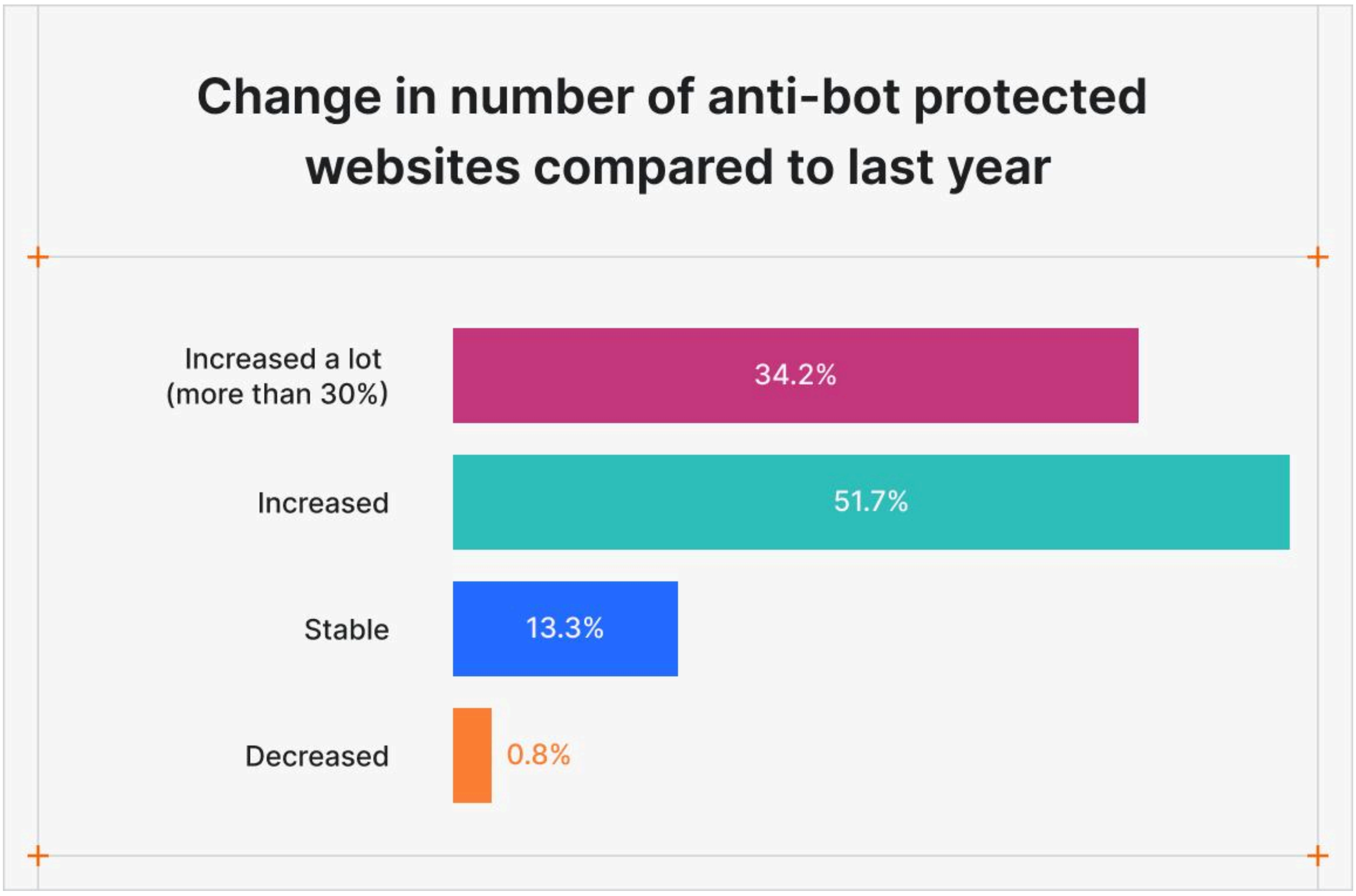
The most frequently mentioned websites were Amazon, Shopee, Walmart, LinkedIn, TikTok, Google, SHEIN, Hermes, Indeed, Expedia, Naver, Facebook, Instagram, YouTube, Zillow, and Home Depot. Less commonly cited but still noteworthy sites included Bol.com, Leboncoin, and Radisson Hotels.

Where our audience seems to disagree with the DataDome report is the trend in anti-bot adoption. The DataDome report says that there are fewer websites protected than in the past year. But compared to 2024:

34.2% of respondents reported that websites protected by anti-bot mechanisms have increased by more than **30%.**

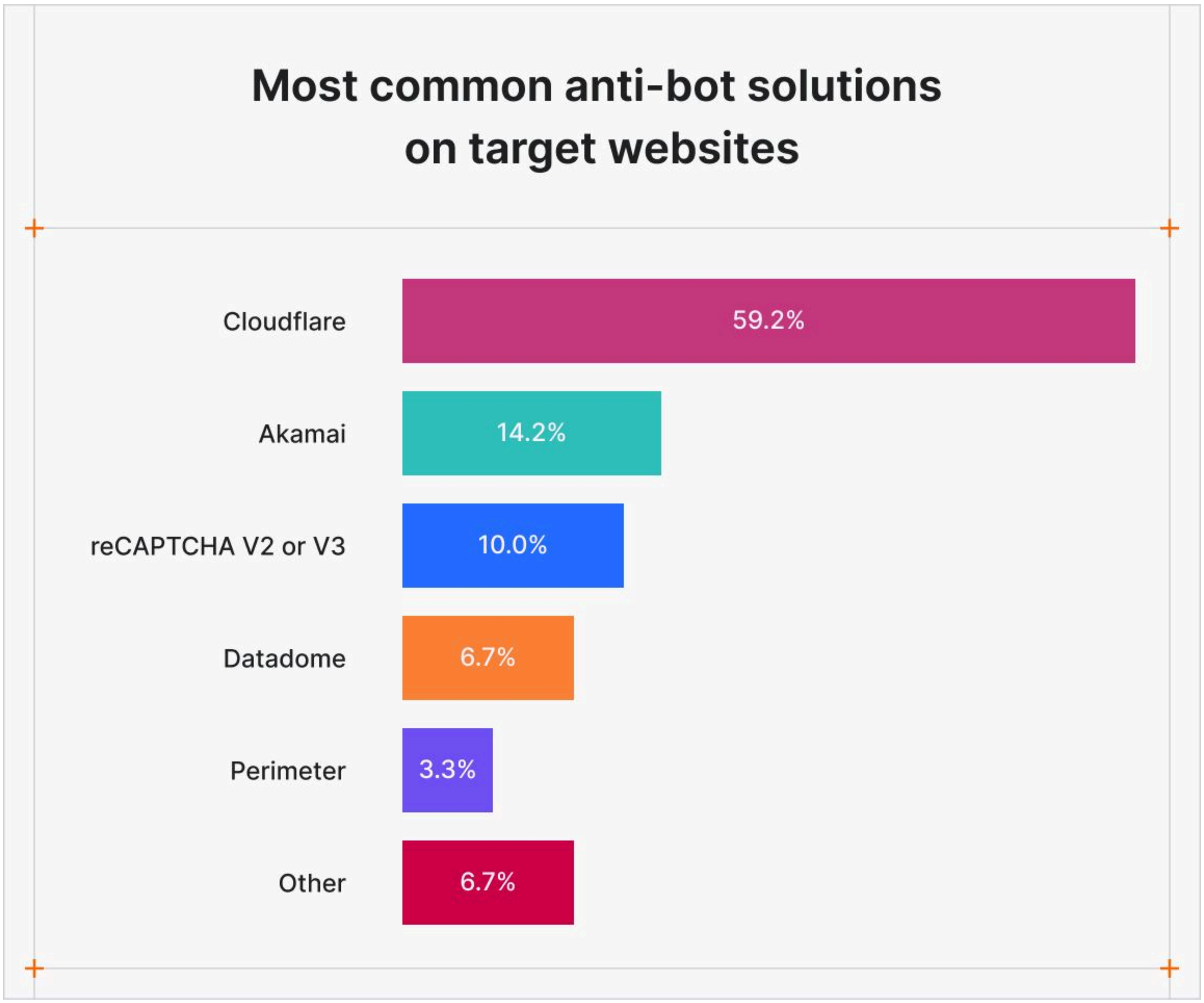
51.7% observed a general increase in anti-bot protections.

13.3% didn’t notice any change, while **0.8%** even reported a decrease.



Those results clearly indicate that anti-bot adoption isn't only widespread on major sites but is also growing steadily overall, highlighting the increasing complexity of web scraping in current environments.

The next logical question was which anti-bot solution is most commonly encountered. The responses were quite predictable, with Cloudflare cited in **59.2%** of cases, followed by reCAPTCHA V2/V3 at **10.0%**, Datadome at 6.7%, Akamai at **4.2%**, PerimeterX at **3.3%**, and other technologies making up the remaining **6.7%**.



Considering that Cloudflare is used by **81.9%** of websites with known reverse proxy services — representing **20.9%** of all websites (at least, according to W3Techs) — it's not surprising that it emerged as the most popular anti-bot solution, affecting **~60%** of respondents.

Beyond Cloudflare, scraping experts reported challenges with other WAFs such as Datadome and Akamai. For CAPTCHAs, the most frequent obstacles were reCAPTCHA and hCaptcha. Other highlighted measures were rate limiting and TLS fingerprinting.

Keeping in mind that AI is making bots become more and more sophisticated, it's no surprise that WAFs, CAPTCHAs, and other anti-bot systems are evolving in response. This helps explain why **46.7%** of surveyed users reported increased difficulty when scraping data from protected sites.

When we inquired about what users employ to overcome those obstacles, open-source web scraping libraries cited were Camoufoux, curl_cffi / curl-impersonate, Scrapling, and Botasaurus.

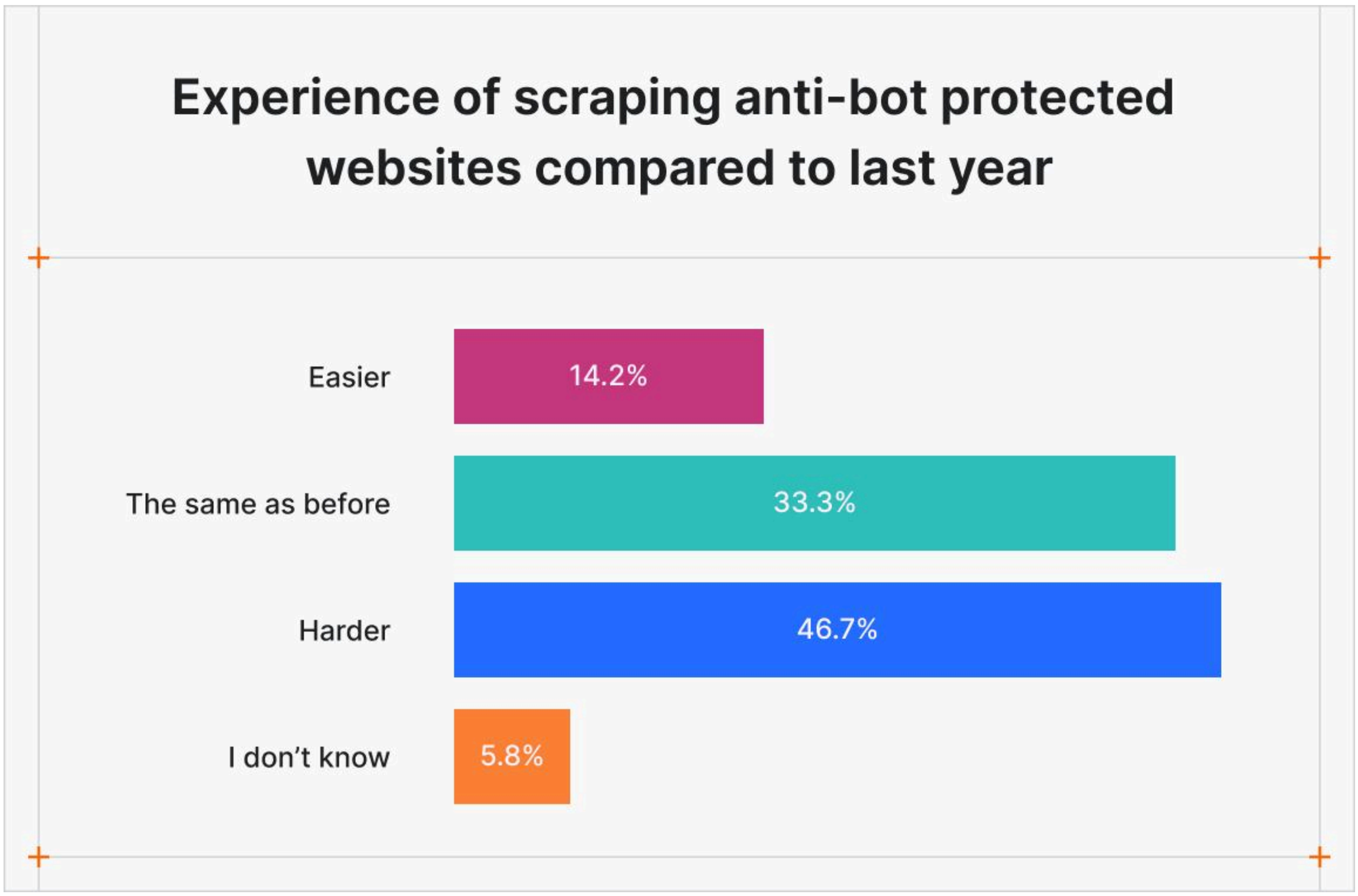
In addition to these, respondents specifically highlighted techniques such as TLS fingerprint emulation, cookie manipulation, and browser automation.

Yet, open-source libraries alone don't appear to be sufficient. That's because the most common approaches reported were:

- Residential proxies / rotating proxy services
- CAPTCHA solvers
- Web unlockers

Other methods mentioned included anti-detection browsers.

These are all primarily paid solutions, which aligns with the finding that **88.7%** of respondents declared increased costs when targeting websites protected by anti-bot measures.

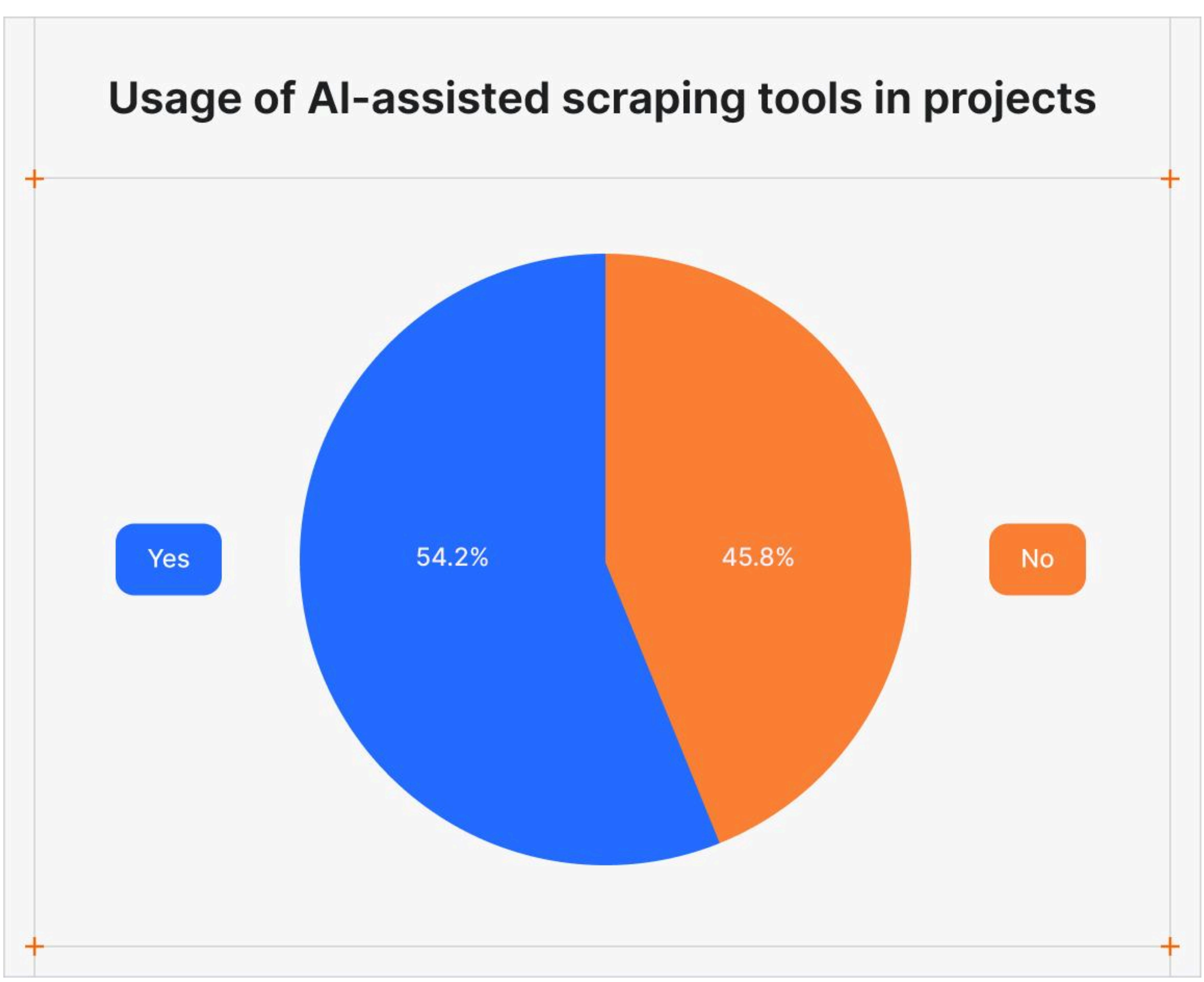


Considering that proxies were the most common solution mentioned for bypassing anti-bot mechanisms, and given the observed increase in proxy expenses, the rise in overall scraping costs when targeting protected sites is a natural consequence.

4. AI scraping

As you might expect, the new entry in this year’s report is AI. What you may not have guessed are some of the interesting insights that emerged from our survey.

First, we asked our experts whether they use AI-assisted tools in their web scraping projects. **54.2%** stated that they don't use AI in their scraping workflows, and only **45.8%** of surveyed web scraping experts are leveraging AI in their projects.



Given this clear dichotomy, it makes sense to split the analysis into two sections:

1. Not using AI
2. Using AI

Not using AI

When we inquired why respondents haven't yet embraced AI in web scraping, they cited several barriers.

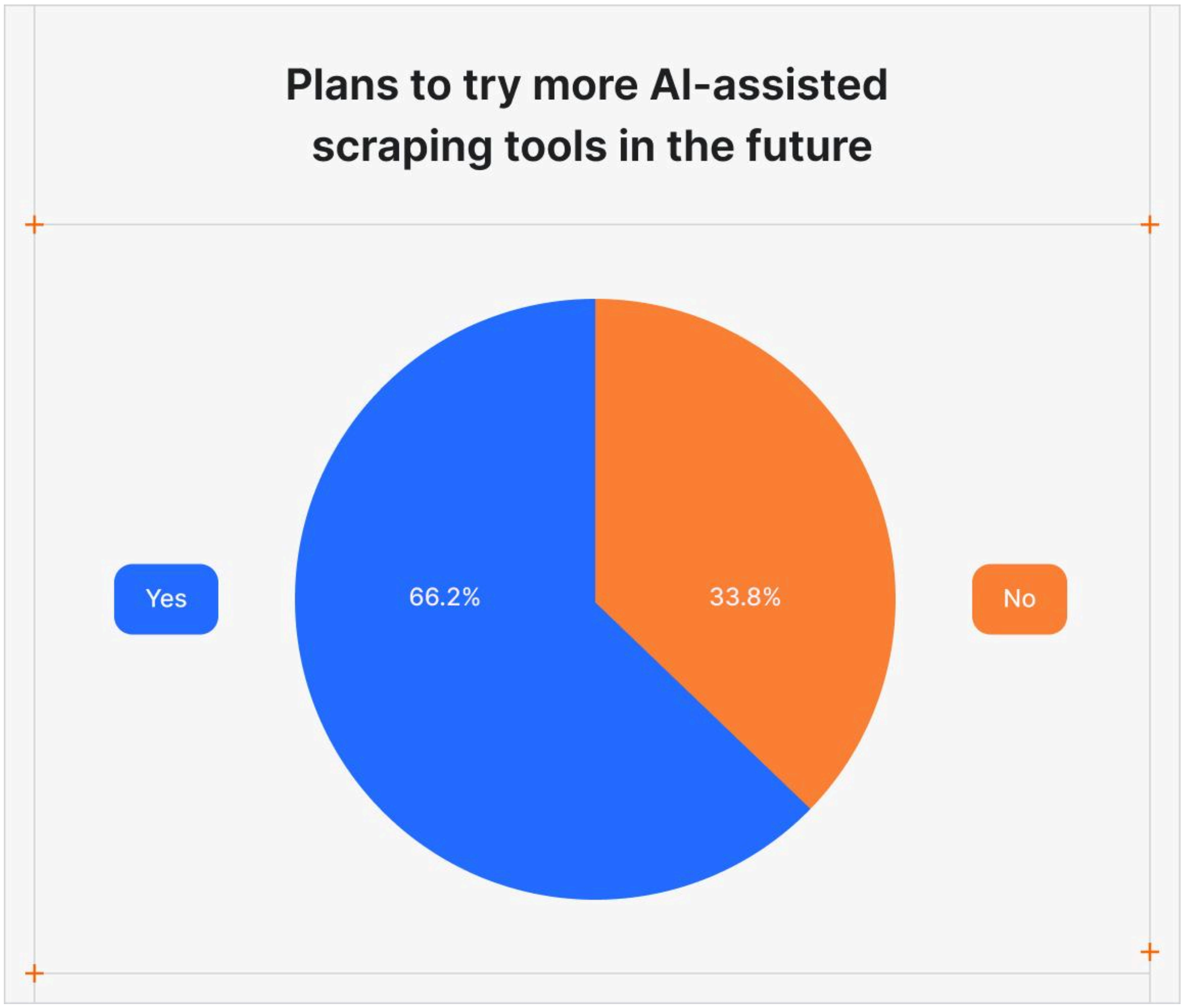
The most common reasons included:

- Lack of trust in the results.
- High costs.
- Technical challenges when integrating or using these tools.

Other mentioned factors were:

- Unreliable performance on certain websites.
- Uncertainty about practical benefits or use cases.

Despite this skepticism, when asked whether they plan to try AI-assisted web scraping tools, **66.2%** of respondents said yes, while only **33.8%** indicated they don't plan to use them even in the future.



This demonstrates that AI in web scraping, while not yet fully adopted, is surely gaining traction, with most users who haven’t integrated AI already planning to do so.

On the other hand, it's notable that ~1/3 of the users who haven't yet adopted AI show no interest in using it at all, meaning that more than 1/6 of the total surveyed population remains highly skeptical of AI.

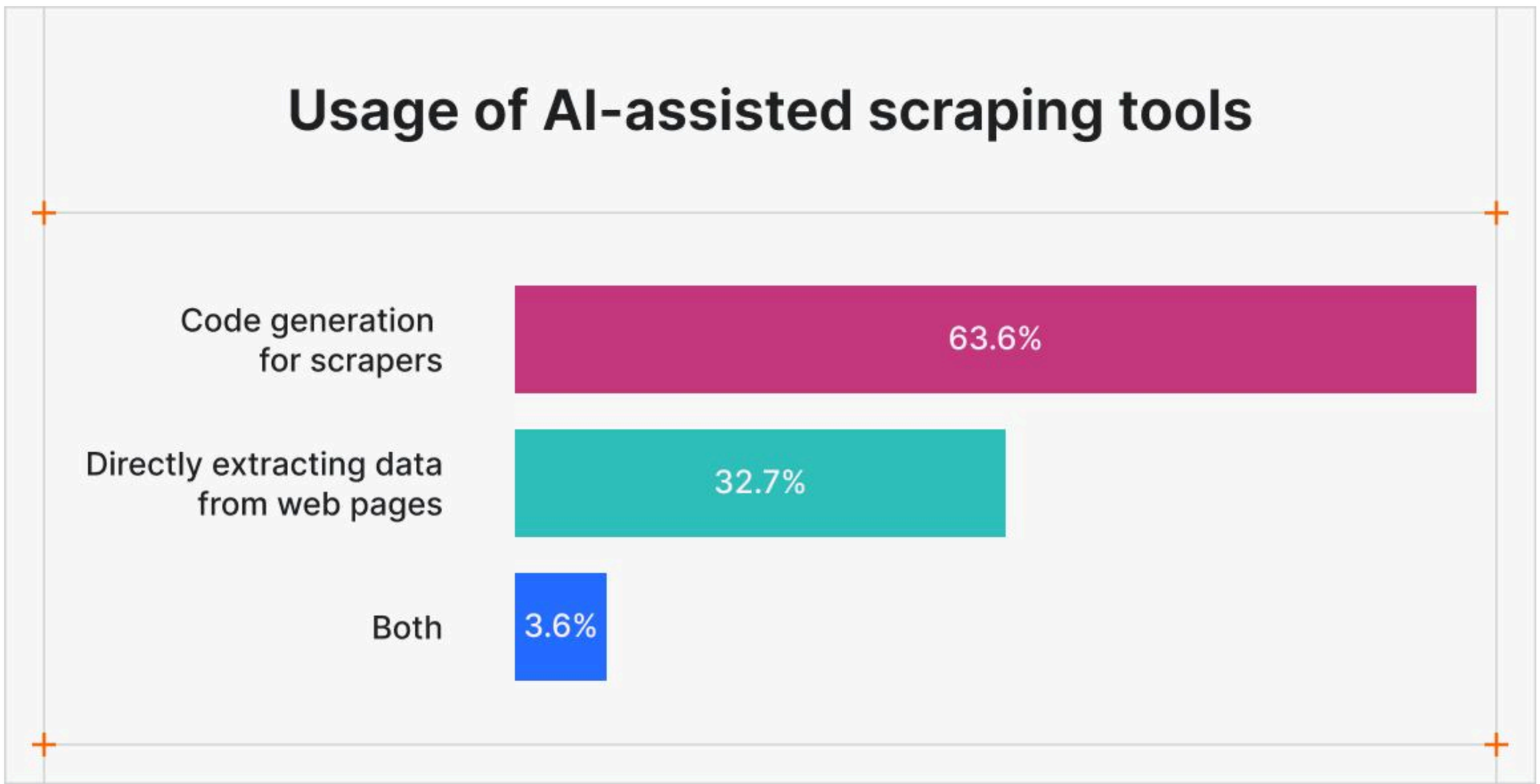
Using AI

When asked to outline their applications, the minority of respondents who use AI revealed the following:

63.6% use it to generate scraping code (e.g. through ChatGPT, Gemini, and similar).

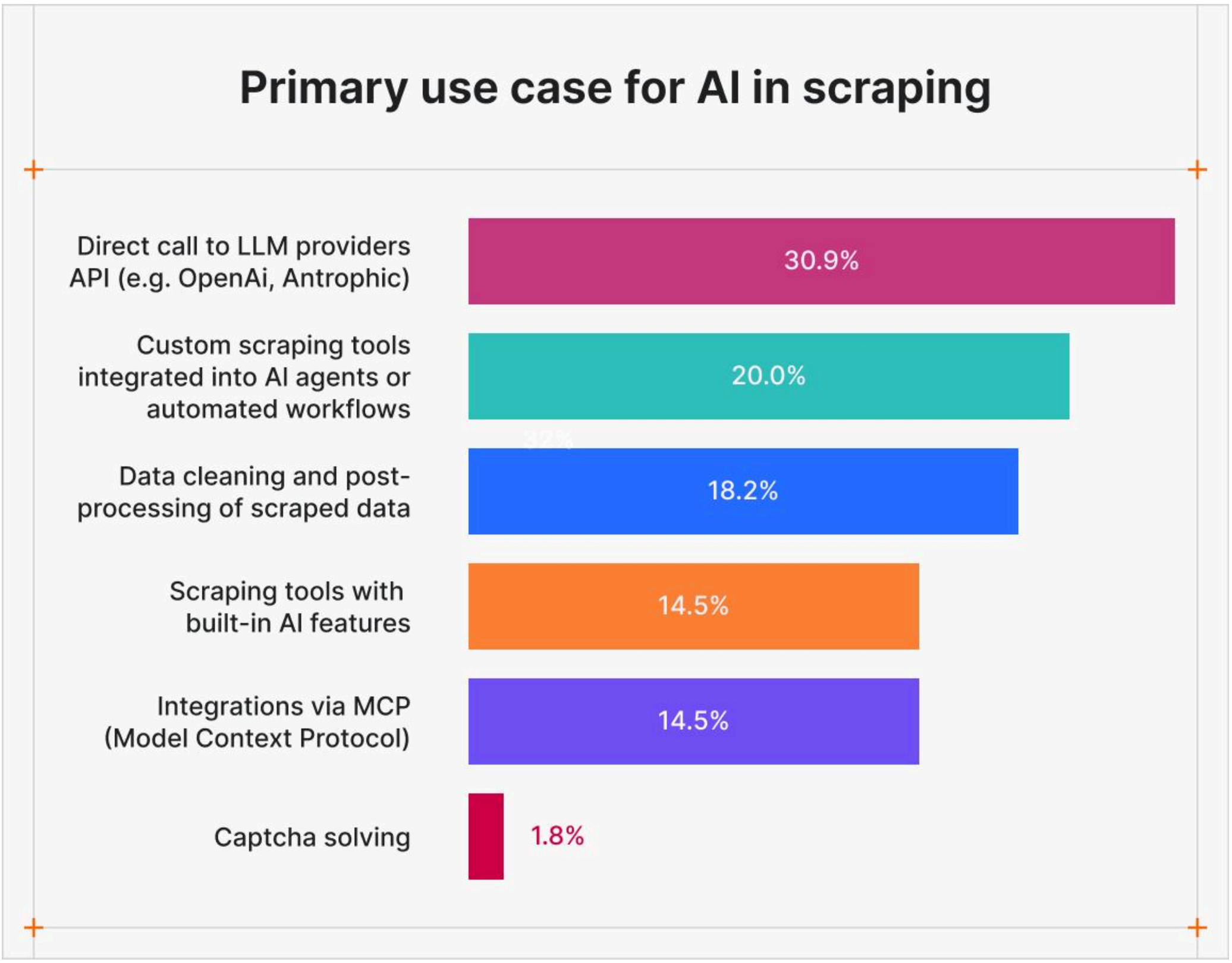
32.7% employ it to extract data from web pages, leveraging its virtually unbreakable data parsing capabilities.

The remaining 3.6% use AI for both tasks.



Next, we asked where exactly AI comes into play. The primary use cases reported were:

- Direct calls to LLM provider APIs **(30.9%).**
- Custom scraping tools integrated into AI agents or automated workflows **(20%).**
- Data cleaning and processing **(18.2%).**
- Scraping tools with built-in AI features **(14.5%).**
- Integrations via MCP **(14.2%).**
- CAPTCHA solving **(1.8%).**



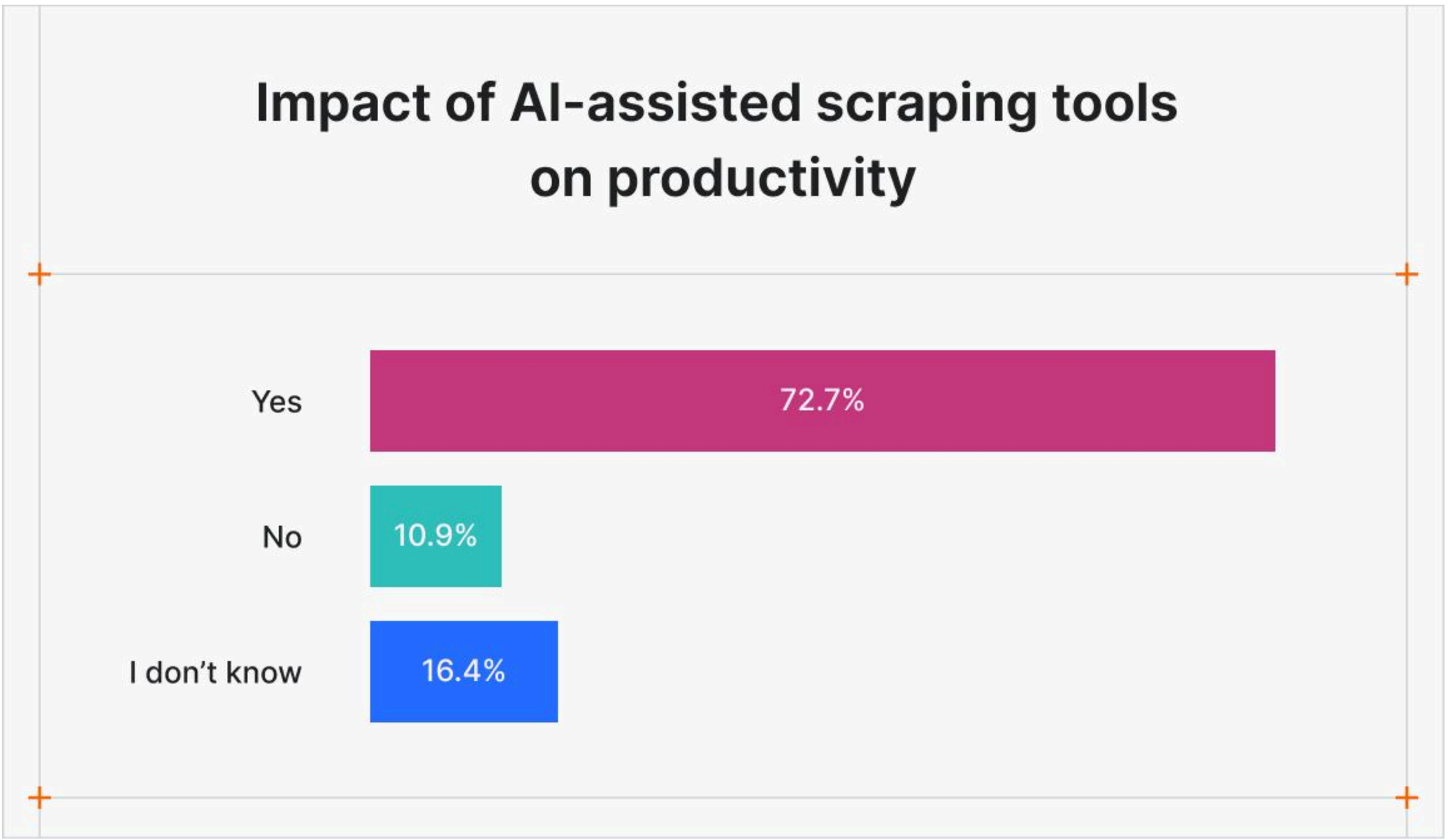
Apart from CAPTCHA solving (which is to be expected, as CAPTCHAs are specifically engineered to stop non-human access), AI is being used fairly evenly across different tasks and in various forms. This suggests that a single dominant usage pattern hasn't yet emerged.

The most cited tools related to AI web scraping included:

- **AI models:** OpenAI (including ChatGPT), Claude, Gemini, DeepSeek, Ollama models.
- **Automation and workflow-building tools:** N8N, Apify, Stagehand.
- **AI-powered web scraping libraries:** ScrapeGraphAI, Crawl4AI, CyberScraper-2077, Firecrawl.
- **AI-assisted IDEs or extensions:** Replit, Cursor, Zyte Copilot.

The mention of IDE solutions reinforces that many use AI for web scraper code generation.

What's certain is that AI speeds up development and prototyping, with **72.7%** of respondents believing that AI in web scraping delivers productivity advantages.



Lastly, when prompted about the pros and cons of AI, respondents highlighted the following:

Pros of AI

- Faster coding, testing, and prototyping.
- Ease of use.
- Simplifying complex tasks and reducing manual effort in parsing.
- Self-healing scrapers.
- Integration in agentic workflows.
- Natural language interfaces.

Cons of AI

- Hallucinations are the most frequently mentioned concern.
- Lack of control and versatility.
- Non-deterministic outputs and inconsistencies.
- Speed and scalability issues.
- Cost.
- Learning curve required to adapt AI tools to real-world scraping tasks.

Despite these issues, **100%** of respondents reported that they plan to increase the use of AI-assisted scraping tools in their future projects.

The obvious conclusion is that users who have embraced AI in web scraping not only find it useful but also expect its adoption to grow steadily, signaling that AI is here to stay.

What to expect for web scraping in 2026

Thanks to the feedback from our communities, we can summarize how web scraping professionals see the industry's present and future.

The rising costs for proxies and the spreading of more sophisticated anti-bot solutions make the game more expensive and challenging. This leads to more complex in-house bot detector bypasses or additional expenses on third-party solutions like web unblockers. With all probability, this trend won't decline in 2026.

While AI can help simplify writing scrapers and generally increase the productivity of scraping teams, AI usage is not yet widespread in the industry. There's still great skepticism around it, especially when it comes to hallucinations and inconsistencies in outputs and behavior. Will this change in 2026? Maybe, with hopefully better models and better data access.