PROWESS

Technical Research Study

Which Benchmarks Are Best for Your Workloads?

When evaluating PCs, it's critical to select benchmarks that reflect real-world usage, but many popular benchmarks can be misleading. This study identifies which benchmarks provide the most useful results.

Executive Summary

It is critical to select the right benchmarks to accurately evaluate computer and processor performance across real-world workloads. While many popular benchmarks are easy to install and run, they don't always accurately reflect actual user scenarios, which can lead to misleading or confusing results. In this technical research study, commissioned by Intel, Prowess Consulting investigated which benchmarks best represent real-world performance and which might not be as representative of the applications and workloads your users tend to run.

Our analysis found that synthetic benchmarks like PCMark10® and PassMark® offer limited insight into everyday usage. In contrast, tools such as UL Procyon® office productivity and WebXPRT™ provide more relevant data by running common tasks in Microsoft Office® and modern web applications. Another benchmark, CrossMark®, measures real-world workloads such as frame rendering and video/image colorization, mimicking what users do in applications like Adobe® Lightroom® or Adobe® Premiere® Pro. For AI workloads, the UL Procyon AI benchmarks and MLPerf® Client offer early but promising options for evaluating inference performance across CPUs, GPUs, and neural processing units (NPUs).

We recommend evaluators prioritize benchmarks that align with their users' actual workflows and avoid those that favor convenience over relevance. For the most accurate performance insights, use a combination of real-world benchmarks tailored to your users' specific use cases.

Why Benchmarking Matters

Whether you are an enterprise or a government buyer of PCs, a technical reviewer testing new processors for readers, or even an OEM device manufacturer generating marketing and sales-enablement materials, benchmarks provide critical data for your analysis. In an ideal world, you could install relevant applications and script your own repeatable tests for customized analysis. And in some cases, this might be a viable option. In most cases, however, off-the-shelf benchmarks offer a more convenient and timely way to generate results that can be compared across devices and platforms.

Given the need for benchmarks, how do you determine which ones are most relevant for your workloads? And after you determine that a particular workload is relevant, how do you know if it will be easy to configure and run for your needs?

To answer these questions, Prowess Consulting examined several popular benchmarks designed to analyze different workload categories, from productivity to content creation, web apps, gaming, and battery life. We also examined some emerging benchmarks built to quantify and compare AI performance across devices. Our analysis focuses on which types of benchmarks we do not recommend and why, and we then offer viable alternatives for benchmarks that provide useful, relevant data.

Selecting Benchmarks: What Not to Run

No matter who you are, whether an enterprise IT buyer, a YouTube® reviewer, or a consumer, if you want an accurate portrayal of expected performance, you need to rely on real-world and relevant benchmarks that are built on the apps and workloads your users will actually run.

Highlights

To be useful and accurate, benchmarks need to be:

Relevant:

Run in a way that represents actual scenarios

Real:

Use the actual apps and workloads users will run

Simple:

Easy to install and run repeatedly

This might seem obvious, but for many reviewers and testers, it can be tempting to pick benchmarks based purely on how easy it is to install, configure, and re-run tests without the need to fully reset the PC between runs. However, while ease of use is important, it shouldn't be the primary or only consideration. Several popular benchmarks are easy to run, but they might be irrelevant (those that don't provide a true measure of end-user performance) or synthetic (those that do not run real-world workloads), severely hindering their value in a real-world context.

In this context, *irrelevant* benchmarks are ones that don't provide results aligned to your needs or that measure performance in an unrealistic way. For example, Cinebench focuses on performance rendering a single image repeatedly. While the benchmark performs this specific function well, the simplified workflow does not reflect the robust range of content creation workloads that organizations and users generally perform. As such, the benchmark is not as relevant for evaluating PCs for real-world use cases.

Another example would be with benchmarks that might use real-world applications or workloads, but in ways that are not indicative of typical usage. For example, running an intensive workload repeatedly without concurrently running background applications or workloads.

Synthetic benchmarks are programs designed to evaluate performance by using standardized workloads. These benchmarks are constructed to simulate a wide range of operations and to stress the hardware in a controlled manner, allowing for consistent and repeatable results. However, synthetic benchmarks typically don't reflect real-world application behavior, and they can therefore struggle to reflect actual user experiences. While these benchmarks are designed to have easily repeatable results across test runs on a given PC, those results might not provide a true picture of how the PC will perform for everyday tasks.

Our research found popular benchmarks that fall into these categories of less relevant or overly synthetic benchmarking tools, including:

- **PCMark 10** is a benchmarking suite that evaluates overall system performance using workloads modeled on real-world tasks such as web browsing, videoconferencing, spreadsheet editing, and digital content creation. PCMark 10 simulates realistic activities, but it does not run common commercial applications such as Microsoft Office apps, which means that its results might not align with real application performance. Perhaps because of this issue, UL Solutions is now replacing PCMark 10 with the UL Procyon suite of benchmarking tests, which are based on relevant workloads using popular applications, including Microsoft Office apps.
- The **PassMark** benchmark is a synthetic performance testing suite that evaluates a computer's hardware—such as CPU, GPU, memory, and disk—using standardized workloads to generate comparative scores. While it is widely used for quick comparisons, PassMark does not simulate real-world application behavior, and it might not accurately reflect actual user experiences.

Identifying Relevant, Real-World Usage Benchmarks

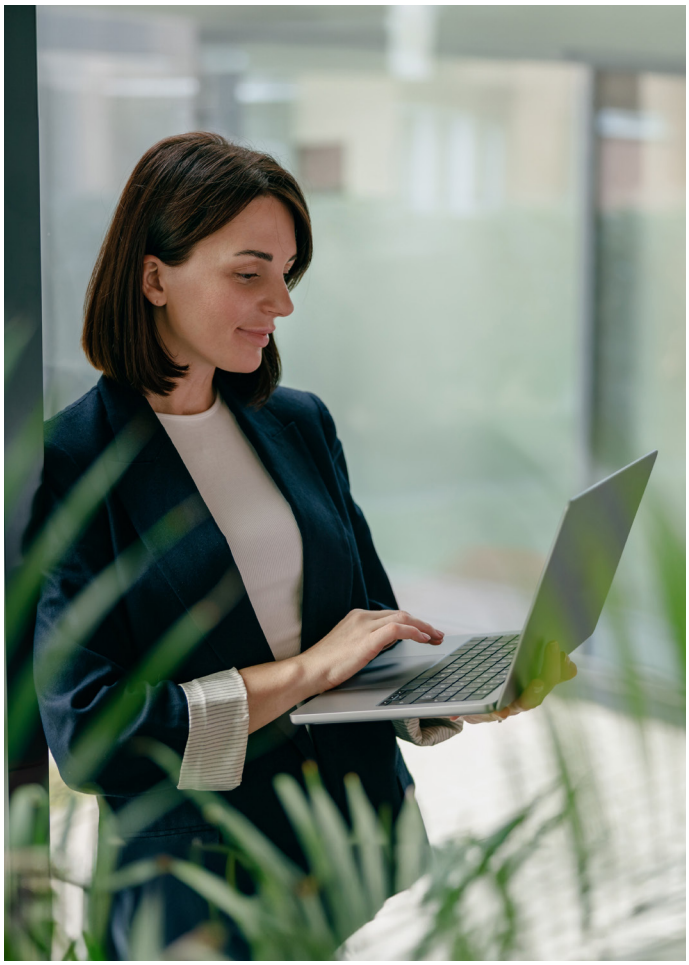
Fortunately, we found several benchmarks that offer relevant, real-world usage tests. These are all widely available to users and reviewers and can help determine the best devices for your needs. To choose the most relevant benchmark for your users, begin by identifying the category or categories that best fit your use case, from productivity to content creation, web apps, gaming, battery life, or AI workloads.

Whenever possible, we recommend testing by running your apps and workloads in combination with one of the benchmarks listed below to create a broader and more realistic representation of performance. For example, you could run the UL Procyon office productivity benchmark while the user shares their screen in a Microsoft Teams® call, including applying a blurred background effect to their webcam video.

Overall Performance and Productivity

We found that the following easy-to-run benchmarks are well-suited for measuring real-world productivity performance using common applications, including the Microsoft Office suite.

- The **UL Procyon office productivity benchmark** is a multiplatform performance test that uses real Microsoft Office applications (including Microsoft® Word, Excel®, PowerPoint®, and Outlook®) to evaluate how well Windows® PCs and Apple® Mac® devices can handle typical office productivity tasks. This benchmark simulates real-world workflows such as editing documents, managing emails, and working across multiple Office apps simultaneously. It is widely recognized as a real-world benchmark because it reflects actual user workloads rather than synthetic or isolated component tests.
- **CrossMark** is a cross-platform benchmark that evaluates overall system performance and responsiveness using models of real-world applications across productivity, creativity, and responsiveness scenarios. It is designed to be easy to run, and it supports a wide range of platforms including Windows, macOS®, iOS®, and Android™. CrossMark is considered a real-world benchmark because—in contrast to PassMark, for example—it can model realistic usage scenarios, such as document editing, web browsing, and photo/video editing, making it a reliable tool for assessing everyday PC performance.



Content Creation

We found that the following benchmarks are well-suited for content creators because they rely on realistic workloads and are run using the same commercial software these users tend to work with.

Photo Editing

This first pair of recommended benchmarking tests focuses on photo editing workloads:

- The **PugetBench for Photoshop®** benchmark runs directly within Adobe® Photoshop to evaluate system responsiveness and throughput across real-world photo editing tasks such as filter application, layer manipulation, and AI-based features like Generative Fill. This benchmark was designed in collaboration with industry professionals to reflect actual creative workflows.
- The **CrossMark Creativity sub-score benchmark** is a component of the CrossMark suite that evaluates a system's performance in content creation tasks such as photo editing, video editing, and media manipulation. It provides a useful, real-world representation of content-creation workflow performance.

Video Editing

Video editors are best represented by the following benchmark, which relies on real usage scenarios:

- **PugetBench for Premiere Pro** runs directly within Adobe Premiere Pro to evaluate system performance across real-world video editing tasks such as encoding, GPU effects, and media processing. It is another benchmark in the PugetBench for Creators suite that was designed in collaboration with industry professionals to reflect actual workflows used by video editors and content creators.

Web Applications

Many organizations rely on web browser–based applications for their workflows. The following benchmark provides reliable, real-usage tests that can be used to evaluate web application performance:

- **WebXPRT 4** is a browser-based benchmark that evaluates the performance of web-enabled devices using real-world scenarios implemented in HTML5, JavaScript®, and WebAssembly. WebXPRT 4 covers common tasks designed to reflect everyday web application usage, including photo enhancement, organizing albums using AI, stock option pricing, encrypting notes, and optical character recognition (OCR) scanning using WASM, generating sales graphs, and performing online homework. WebXPRT simulates realistic browser workloads rather than synthetic component tests, making it a useful tool to run alongside other benchmarks like CrossMark and UL Procyon for a more comprehensive evaluation of system performance in practical, user-relevant contexts.

Gaming

For game testing, we recommend using the built-in benchmarks for games that provide them. These benchmarks offer a repeatable and easy-to-run test for generating frames-per-second (FPS) and other stats that can be used to compare performance between devices. When running these benchmarks, both PC and in-game display and graphics settings need to be configured consistently across test devices. Considerations include PC display resolution and refresh rate, use of a GPU, and in-game settings for graphics APIs, resolution, and other settings.

When game-specific benchmarks are not available, you can try using a manual or scripted sequence to ensure that each PC runs the same in-game scenario, eliminating as much variability as possible due to user input or random events.

Battery Life

It can be tricky to measure battery life because it must be viewed in the context of both time and performance. The reality is that performance levels often drop once a laptop is unplugged from the wall. This is because high performance generates heat inside the chassis, causing the fan to spin up, and increased fan usage impacts battery life. Many manufacturers account for this by strategically reducing performance to extend battery life. As a result, it is important to compare plugged and unplugged test results in order to select a PC that provides the right power/performance balance for your specific needs. After all, what good is peak performance if your battery dies before you complete your work? Or conversely, how useful is long battery life if your performance is completely throttled?

When examining power/performance, be aware of the Windows power management strategy and how it is configured. Laptops running in “Balanced Power Mode” will prioritize battery life over performance for some apps when unplugged. However, the performance drop might be small enough to go unnoticed by users, while their battery life might go up significantly.

Options for Testing

An ideal way to conduct battery testing would be to run your primary applications and workloads under battery power by scripting actions that can be run repeatedly. This could include running Office applications, content creation applications, browser workloads, video playback, and videoconferencing workloads. Obviously, this method requires a high degree of upfront effort to configure. Another, simpler option is to run the **UL Procyon Battery Life benchmark**. This benchmark measures the real-world battery life of Windows laptops, notebooks, and tablets across multiple usage scenarios, including video playback, office productivity, and idle mode. It also provides a detailed battery life profile that can help you compare devices under realistic conditions.

You can also run other benchmarks listed earlier in this document, such as CrossMark, WebXPRT 4, or other UL Procyon benchmarks, while both plugged in and on battery power to compare performance. By combining those results with those of the UL Procyon Battery Life benchmark, you can compare power/performance numbers between devices to determine which device offers the best combination according to your use cases and your users’ workloads.

AI Capabilities and Performance

When evaluating PCs for AI performance, you might be tempted to simply compare the published specifications for trillions of operations per second (TOPS) between PCs. TOPS represents the number of computing operations an AI chip (such as an NPU) can handle in one second, which makes it useful for providing a single number to encapsulate an AI chip's computational capability. When it comes to expected performance, however, TOPS can be misleading for two primary reasons:

- TOPS does not differentiate between the types or quality of operations that a chip can process.
- TOPS is not relevant if it refers to a specific compute engine in your PC, but your workload primarily relies on a different processor; for example, if TOPS refers to the NPU, but your workload relies on the CPU.

In addition, other factors such as memory bandwidth or software optimizations are not factored into TOPS specifications. We recommend that you don't rely solely on these benchmarks or TOPS for AI workloads. Wherever possible, test your specific applications with a focus on the expected AI use cases for your workers.

Ideally, AI-specific benchmark tests would be used to replace or accompany TOPS specifications for evaluating and comparing PCs. However, AI benchmarking tools are still in early development, and the landscape is changing rapidly. The following list covers useful benchmark tools that are available as of the publication of this study:

- The **UL Procyon AI Computer Vision benchmark** evaluates the performance of AI inference engines on tasks such as object detection, classification, and segmentation using state-of-the-art neural networks like MobileNet V3 and YOLOv3®. This benchmark supports multiple inference engines (for example, NVIDIA® TensorRT™, Intel® OpenVINO™, Qualcomm® SNPE, and Microsoft® Windows ML), and it runs on CPUs, GPUs, and NPUs. It's designed to reflect real-world machine vision workloads, and it is easy to run via graphical user interface (GUI) or command-line interface (CLI).
- The **UL Procyon AI Image Generation benchmark** measures the inference performance of on-device AI accelerators (like high-end discrete GPUs) in generative image tasks. It provides a consistent and transparent workload for evaluating how well a system can generate images from prompts, simulating real-world creative use cases.
- The **UL Procyon AI Text Generation benchmark** assesses natural language generation capabilities, such as those used in tools like ChatGPT®. It evaluates how efficiently a system can run large language models (LLMs) to generate coherent and contextually appropriate text, offering insights into performance across CPUs, GPUs, and NPUs.
- The **MLPerf Client benchmark** assesses how well PCs handle generative AI workloads such as LLMs, including Meta's Llama® 2 and Llama 3. This allows users and vendors to compare AI capabilities across different PC hardware configurations in realistic, inference-focused scenarios.

You will want to select the benchmark that most closely aligns with your workload requirements. Note that you might also need to explicitly target the relevant processors—CPU, NPU, or GPU—for testing so the benchmark correlates with whatever processors your specific apps and workloads will rely on.

For a deeper dive into AI PCs and how to benchmark and compare them, see the Prowess Consulting study, "**How to Understand and Evaluate AI PCs for Your Apps and Workloads**."

Benefits for Government Agencies

Government agencies evaluating PC performance can benefit from more than just free or discounted versions. Benchmark providers often offer tailored support for public-sector users, including streamlined licensing, compliance with government standards, and priority technical assistance. These benefits help agencies ensure their testing aligns with IT policies and regulatory mandates.

The discounts offered to government organizations can provide access to low-cost, high-quality benchmarking tools and expertise without the financial barriers of commercial licensing. Qualified government users might be eligible for free or discounted access to the CrossMark suite, the UL Procyon suite, and WebXPRT, while MLPerf Client is free for all users (see **Appendix** for full details).

Recommendations for Real, Relevant Benchmarking

Benchmarks provide a critical tool for users, reviewers, and government or business buyers of PCs. They provide useful metrics for evaluating performance and for comparing PC and processor performance between vendors. For those metrics to be useful, they need to closely align with the tasks end users will be performing.

Based on our research, we recommend running multiple benchmarks to test a variety of usage scenarios. Avoid synthetic benchmarks whenever possible and instead rely on those based on real and relevant apps and workloads. If you are comparing different platforms, you will obviously want to run benchmarks that can are supported on both. For example, UL Procyon office productivity benchmarks offer a useful and realistic way to assess the performance of Microsoft Office apps on both Windows and Mac computers.

Some relevant and easy-to-run benchmark suites cover multiple user scenarios, like CrossMark, which provides tests for generating both productivity and content creation scores. For web applications, WebXPRT is a well-known and relevant benchmarking tool that is easy to run. UL Procyon offers the best options currently for both AI benchmarking and battery life testing. For gaming performance comparisons, in-game benchmarks are the simplest and most relevant option, when they're provided.

See the comparison tables in the **Appendix** for a comparison of benchmarking tools showing real-world relevancy, ease of use, time to run, relative cost, and cross-platform support.

Benchmarks can never fully replace the benefits of installing and running the actual apps and workloads you or your users access regularly. But by running a variety of relevant benchmarks, you can efficiently produce quantifiable results for making a useful, informed purchasing decision.

Appendix: Benchmark Test Comparison Tables

The following tables list the most useful, relevant, and real-world benchmarks by category. Ease of use, time to run, and cost are relative and can depend on device used, configurations, and (for cost) end-user versus enterprise purchases. In general, we categorized each as follows:

- **Ease of use**: Simple means no setup or configuration required, moderate means some setup or configuration required, and complex means significant effort required to install and/or configure for use.
- **Time to run**: Fast is up to 30 minutes, medium is between 30 minutes and 2 hours, and long is more than 2 hours.
- **Cost**: Low is under $100.00 (USD), medium is $100.00 to $500.00, and high is more than $500.00. (Only enterprise usage costs are indicated.)

Table 1 | Recommended benchmarks for performance and productivity

Benchmark	Platform	Category	Relevant Workload	Benchmark Type	Ease of Use	Time to Run	Cost
UL Procyon® office productivity	Windows® and macOS®	Performance and productivity	Yes	Real-world	Moderate	Medium	High*
CrossMark®	Windows, Linux®, ChromeOS™, macOS, iOS®, and Android™	Performance and productivity	Yes	Real-world	Simple	Fast	High**

* Qualifying public-sector organizations are eligible for **complimentary benchmark licenses and procurement advice**. Contact UL Solutions to **confirm eligibility** and request access.

** Qualified government agencies might be eligible for free or discounted access through the **BAPCo Government Network (BGN)**. Email **government_support@bapco.com**, to confirm eligibility and request access.

Table 2 | Recommended benchmarks for content creation

Benchmark	Platform	Category	Relevant Workload	Benchmark Type	Ease of Use	Time to Run	Cost
CrossMark® Creativity sub-score	Windows®, Linux®, ChromeOS™, macOS®, iOS®, and Android™	Content creation	Yes	Real-world	Simple	Fast	High*
PugetBench for Photoshop®	Windows x86/64 and macOS	Content creation	Yes	Real-world	Moderate	Medium	High
PugetBench for Premiere® Pro	Windows x86/64 and macOS	Content creation	Yes	Real-world	Moderate	Medium	High

* Qualified government agencies might be eligible for free or discounted access through the **BAPCo Government Network (BGN)**. Email **government_support@bapco.com**, to confirm eligibility and request access.

Table 3 | Recommended benchmark for web applications

Benchmark	Platform	Category	Relevant Workload	Benchmark Type	Ease of Use	Time to Run	Cost
WebXPRT™	Any web-enabled device	Web apps	Yes	Real-world	Simple	Fast	Low*

* **Free to use** for personal, educational, government, and non-commercial purposes. **$20 one-time membership fee** required to redistribute the benchmark, automate benchmark runs, access the source code, or publish results in official viewers.

Table 4 | Recommended benchmark for battery life

Benchmark	Platform	Category	Relevant Workload	Benchmark Type	Ease of Use	Time to Run	Cost
UL Procyon® Battery Life	Windows® and macOS®	Battery life	Yes	Real-world	Moderate	Long	High*

* Qualifying public-sector organizations are eligible for **complimentary benchmark licenses and procurement advice**. Contact UL Solutions to **confirm eligibility and request access**.

Table 5 | Recommended benchmarks for AI

Benchmark	Platform	Category	Relevant Workload	Benchmark Type	Ease of Use	Time to Run	Cost
UL Procyon® AI Computer Vision	Windows® and macOS®	AI	Yes	Real-world	Simple (Windows) Complex (Apple® Mac®)	Fast	High*
UL Procyon AI Image Generation	Windows and macOS	AI	Yes	Real-world	Simple (Windows) Complex (Mac)	Fast	High*
UL Procyon AI Text Generation	Windows and macOS	AI	Yes	Real-world	Simple (Windows) Complex (Mac)	Fast	High*
MLPerf® Client	Windows x86/64	AI	Yes	Synthetic/ hybrid	Complex	Fast	Low**

* Qualifying public-sector organizations are eligible for **complimentary benchmark licenses and procurement advice**. Contact UL Solutions to **confirm eligibility and request access**.

** **Open-source and free** for all users. Membership with the benchmark working group is required to **submit results for publication**.

0925/240157