

Publication date:

December 2022

Author:

Alexander Harrowell

Enterprise AI Is All About the Developer

Enterprises need platform flexibility to support numerous, diverse, project-driven AI workloads



Sponsored by:



Brought to you by Informa Tech

Contents

Summary	1
Enterprises are adopting AI faster than you think, but the real value is still ahead of us	2
Enterprises have more AI models than you think, and release them more often	5
What does this mean for your production AI architecture?	7
Hardware to beat the implementation gap	10
Conclusions	13
Appendix	14

Summary



Enterprise AI has tipped from the early adopter phase to the early majority – where relatively simple chatbot projects are giving way to more complex analytics. This inflection is reflected in a substantial gap between how many AI projects are implemented versus those only announced. This change means new challenges and nuance comes to the hardware requirements of developers, where early consideration of the right path can bring better outcomes. To help the community understand this, and enterprises make better technology choices, an Omdia survey of enterprise users for Intel investigated how enterprises use AI. This whitepaper explores the key themes and results from the survey, with a conclusion for the developer and supporting community.

Key Highlights

- **Enterprises adopting AI have more distinct and diverse models than previously assumed**, and models are usually trained in the enterprise in support of specific, time-bound projects. Enterprises also re-train or fine-tune their models relatively frequently, sometimes aligning with their software release cycle.
- **Model training and development requirements are therefore high**, whilst audiences using the model for inference is often small. The industry-assumed ‘factor of 10’ difference of AI inference computer workload being larger than training is incorrect – in fact, enterprises use more of their computing resources for model training and development than they do for inference.
- **Enterprise respondents want to avoid porting AI applications from the training environment to a different chip architecture for inference.** Respondents with more AI models in production were more likely to practice a faster software release cycle, typical of modern development methods such as Agile, Scrum, and Continuous Integration/Continuous Delivery.
- All of these factors suggest that training will be a bigger workload relative to inference and that our respondents will need hardware solutions for AI that support an efficient software and model development process. **The problem is not so much how to serve AI inference faster, as how to deliver AI projects and changes to existing AI projects faster.**

“When building infrastructure for your AI development pipeline, the widest possible selection of CPUs with standard programming tools is therefore crucial.”

Omdia believes the intense focus on inference scalability and accelerated computing has been driven by the needs of Silicon Valley hyperscale cloud providers and HPC projects rather than by enterprises, developers, or data scientists. Borrowing a concept from manufacturing, enterprise AI is much more about “mix” than “throughput”, and the survey found that real enterprises value software support and compatibility as much as they do performance.

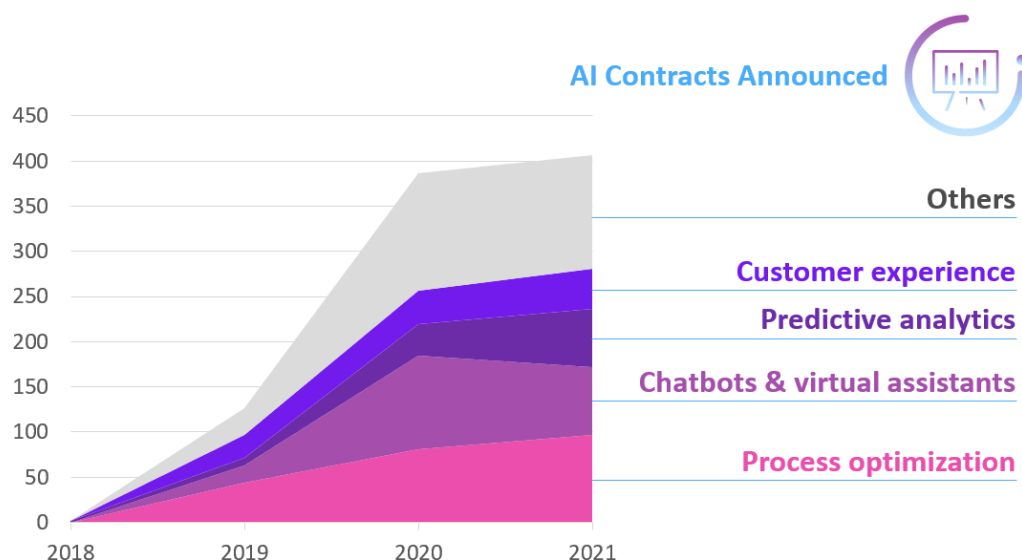
When building infrastructure for your AI development pipeline, the widest possible selection of CPUs with standard programming tools is therefore crucial, as are tools such as AI Blueprints and MLOps services from cnvrg.io, an Intel company. It comes as no surprise that our respondents value CPU compute highly and are “CPU-first” across on-device, edge, and the cloud.

Enterprises are adopting AI faster than you think, but the real value is still ahead of us

Omdia's Enterprise AI Contracts Tracker, a database of publicly announced enterprise AI contracts, was monitoring 1,017 contracts at the end of 1Q22. During 2019-2020, the number of contract announcements surged dramatically, driven most of all by the chatbot and virtual assistant category. Since 2020, the excitement around this category has cooled and growth has shifted into the categories of process optimization, predictive analytics, and customer experience.

"In most cases, this shift will be one towards greater computational complexity as well as to greater integration with the core business, making demands of the technology as well as the people."

Figure 1: Enterprise AI contracts by category, 2018-2021



Measure: Number of contracts announced
Source: Omdia Enterprise AI Contracts Tracker

Omdia understands this as a shift from low-hanging fruit – projects that are above all easy to demonstrate, show quick wins, and very often come shrink-wrapped from a vendor – to a deeper digital transformation engaging AI in fundamental business decisions and processes. As such, we expect this second wave of adoption to generate more value, although over a longer payoff period and with more investment needed to operationalize the opportunity. Data scientists and software

developers will be key to this process, working on one hand with line-of-business domain experts and on the other with infrastructure engineers. In most cases, this shift will be one towards greater computational complexity as well as to greater integration with the core business, making demands of the technology as well as the people.

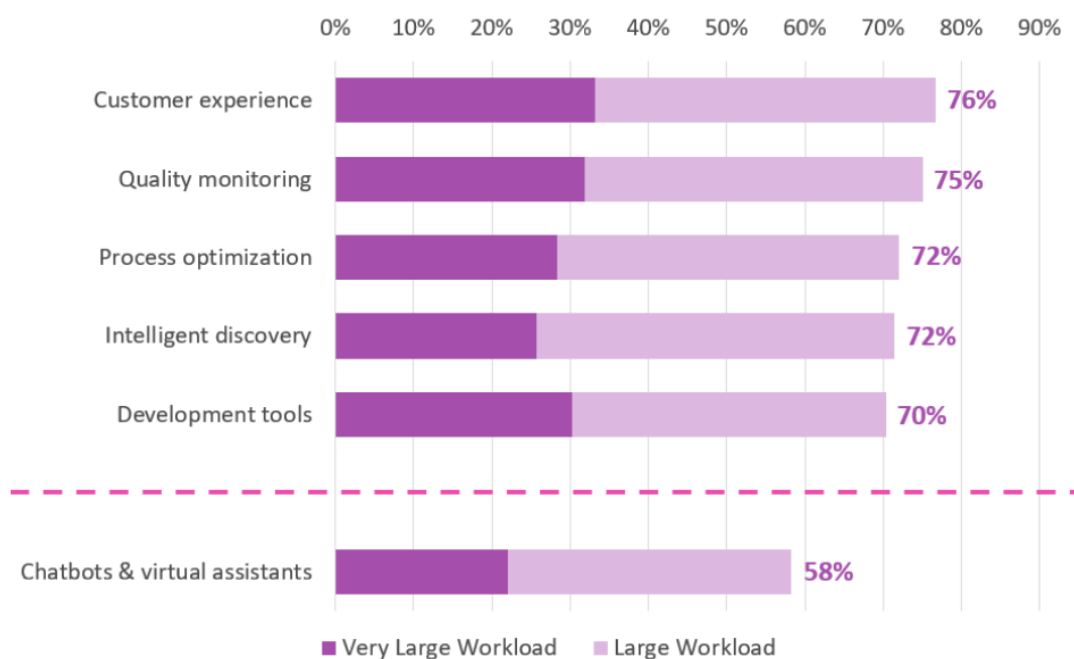
In a new Omdia survey of 304 global enterprises, we found that the rising use cases are already driving greater demand for compute. “Process optimization”, “intelligent discovery”, and even “customer experience” are much more likely to be rated as a “very large” or “large” compute workload than is “chatbots and virtual assistants”.

Figure 2: Rising AI use cases are using more computing power



“What is the scale of the workload created by the following AI use cases?”

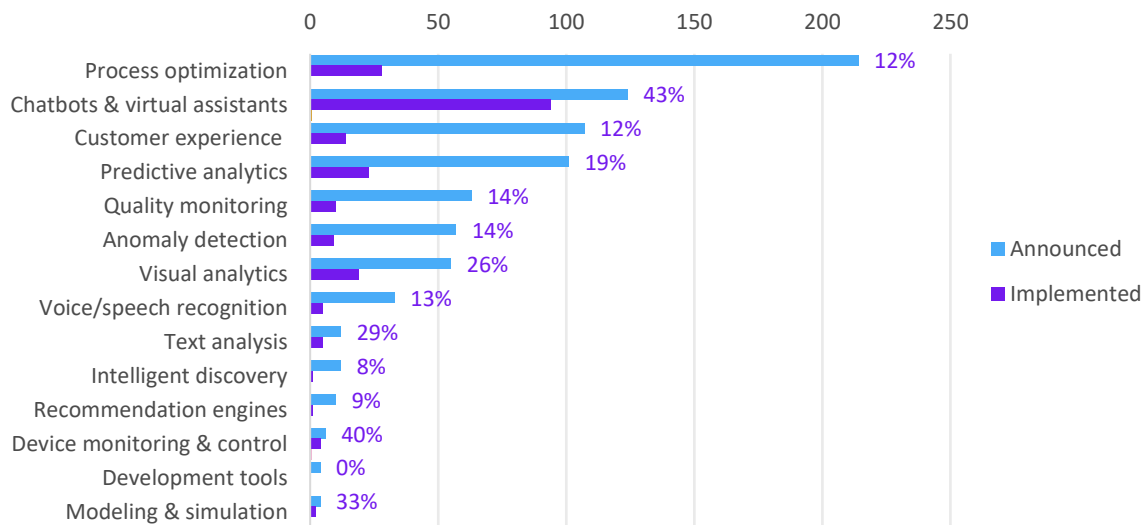
Showing selected answers only



Source: Omdia

Meanwhile, these categories are also proving more difficult to operationalize than simple chatbots. The Tracker makes a distinction between projects that have been implemented and those that have been merely announced, making it possible to measure the implementation gap. Implementation takes time, of course, so newer projects are less likely to be implemented yet. However, there is a clear pattern. Chatbots are by far the easiest projects to implement, followed by visual analytics, while the projects of the future are running at an implementation rate of around 10-20%.

Figure 3: The implementation gap on AI contracts



Source: Omdia Enterprise AI Contracts Tracker

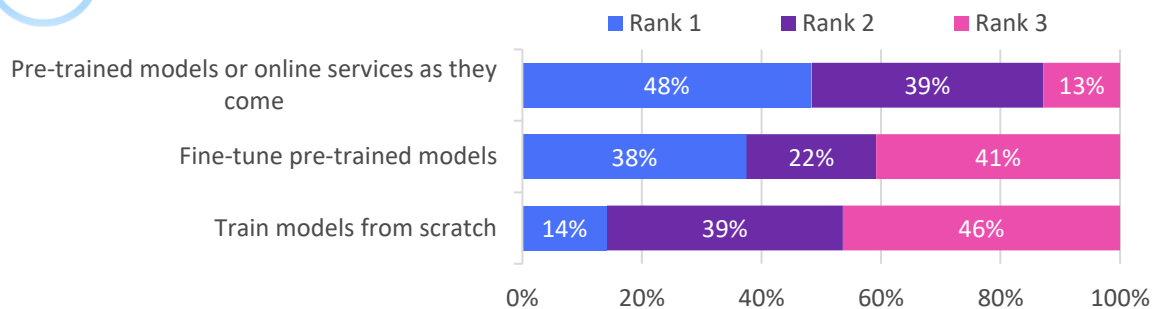
It's worth remembering that if you want a simple chatbot, image classifier, or text miner, these models are increasingly available as online services in products such as Microsoft Azure Cognitive Services APIs or Amazon Web Services' Rekognition. This is an obvious way to reduce the implementation gap. Where it gets tougher is where projects have to become more task-specific and more customized. In Figure 4, our survey results show that although enterprises are keen to deploy ready-made cloud services, this definitely doesn't exclude training from scratch. Interestingly, industrial users are more likely to fine-tune their models, and financial services are more likely to use pre-trained models as they come. Applications developers – who probably know best – are less likely than the other occupational groups to train from scratch.

"...although enterprises are keen to deploy ready-made cloud services, this definitely doesn't exclude training from scratch."

Figure 4: Enterprises are keen on using productized cloud services, but even those that do are doing substantial custom training and development



"Rank the following model training options in order of your organization's most common practice"



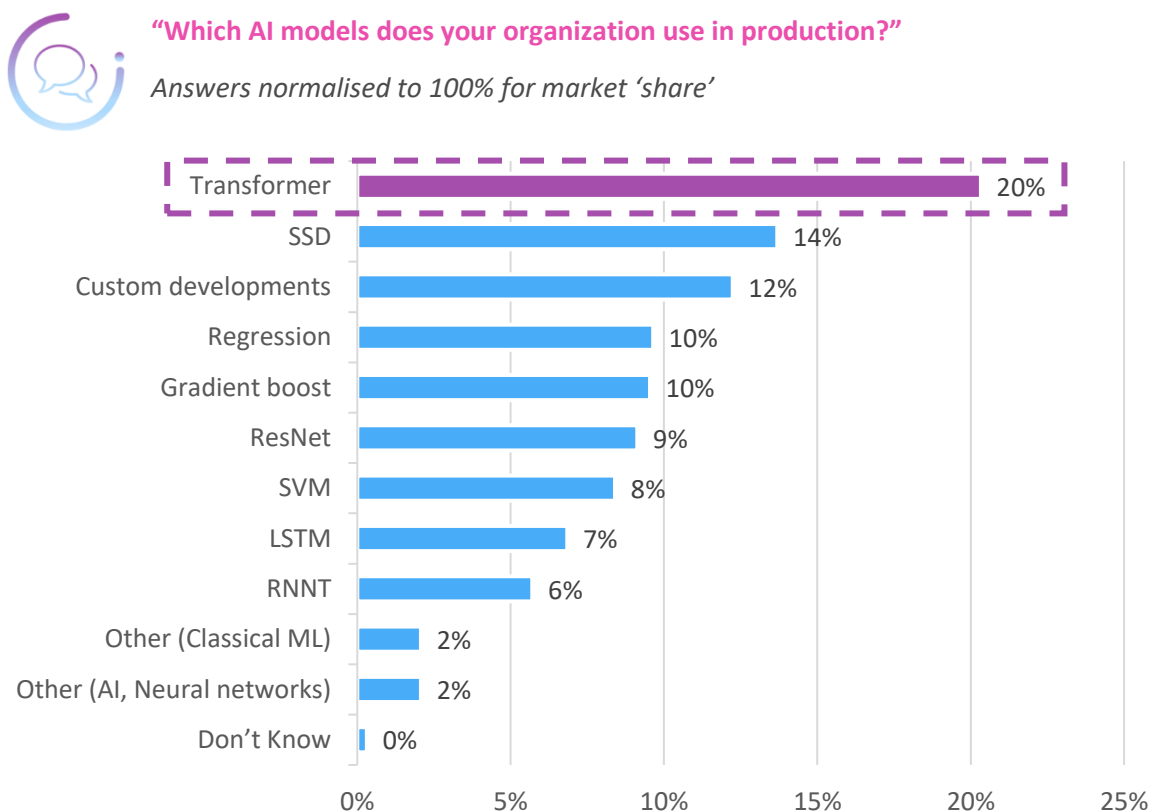
Source: Omdia

© 2022 Omdia. All rights reserved. Unauthorized reproduction prohibited.

Enterprises have more AI models than you think, and release them more often

In Figure 5, we can see which model architectures our respondents report using in production. The most common AI models encountered in the survey were the various Transformer multi-functional large language models, followed by SSD-Mobilenet, a high efficiency image classifier, and then by custom developments. This last detail was a surprise, and points to the increasing ambition and confidence data scientists and developers have in tackling more demanding business problems with AI. It also points to the importance of data preparation, development, training, and evaluation relative to inference, something we'll see more of later.

Figure 5: Enterprises' choice of AI model architectures – Transformers come top, SSD second, and custom developments third



Source: Omdia

Enterprises that are training Transformers from scratch, or creating custom models, are taking on the biggest implementation challenges. Fortunately, there are some well-known methods to overcome the implementation gap. AI projects are software projects, and for some years now best practice in software development has called for frequent, incremental releases driven only by specific user requirements or bug fixes, usually on a regular cycle and linked to automated unit tests. Table 1 compares respondents' AI model fleets and their software development practices.

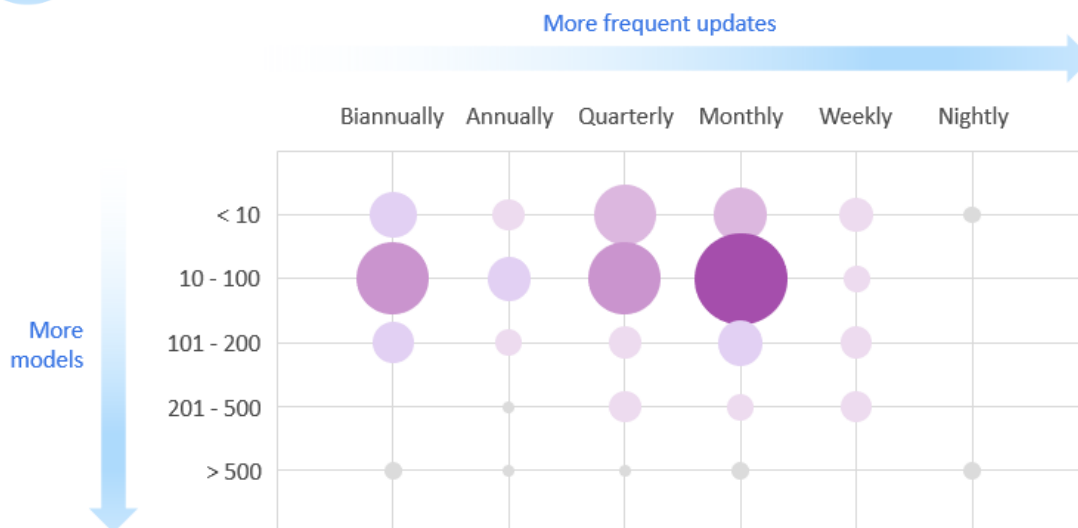
"...enterprises implementing AI need first of all to set up a strong software development cycle, and their infrastructure and hardware choices have to support this if they're going to bridge the implementation gap and deliver value beyond chatbots"

We were surprised, to say the least, by the depth of AI adoption and the sophistication in software development that Figure 6 shows. The modal combination of answers – in darker purple on the heatmap in Figure 6 – reports having 10 to 100 models in production and a monthly software release cycle. 2 companies who claim to have 500+ models also claim to do a nightly build, implying a major commitment both to AI and to cutting-edge software development practices.

Figure 6. Model size and update frequency for AI



"How many AI models do you have in production, and how often do you release the software applications that use them?"



Source: Omdia

A key conclusion is therefore that enterprises implementing AI need first of all to set up a strong software development cycle, and that their infrastructure and hardware choices have to support this if they're going to bridge the implementation gap and deliver value beyond chatbots.

What does this mean for your production AI architecture?

In a follow-up interview, one of the respondents (a data scientist working in the media industry) told us that “most of the projects are small, and all the projects are different”, and that most projects in their experience were commercially driven, about sales, and would be initiated tactically, to investigate specific problems or inform decisions. As such, project turn-around time was much more important than, for example, inference performance.

An engineer/executive with a global investment bank, meanwhile, told us that the bank had numerous and very diverse AI models, and although they worked with several silicon vendors, a substantial sub-set of their fleet of models were unsuited to most accelerator architectures and consequently had to be trained on CPUs. In general, our respondents reported taking an eclectic approach to

AI, with a wide variety of model architectures, data sets, and applications. The variety of business problems they were asked to address is both large, and constantly changing, while the typical data set was at least semi-structured (e.g. text or time-series data) rather than media (e.g. images or video).

“Most of the projects are small, and all the projects are different.”

*Data Scientist
Media Industry*

The combination of lots of relatively small models with diverse characteristics, frequent software releases, and a project-driven approach to AI implies that our respondents will be doing a relatively large amount of model training and development. It is commonly assumed that AI inference will always be a much bigger workload than training or development, but this makes certain key assumptions that do not necessarily hold in the enterprise context:

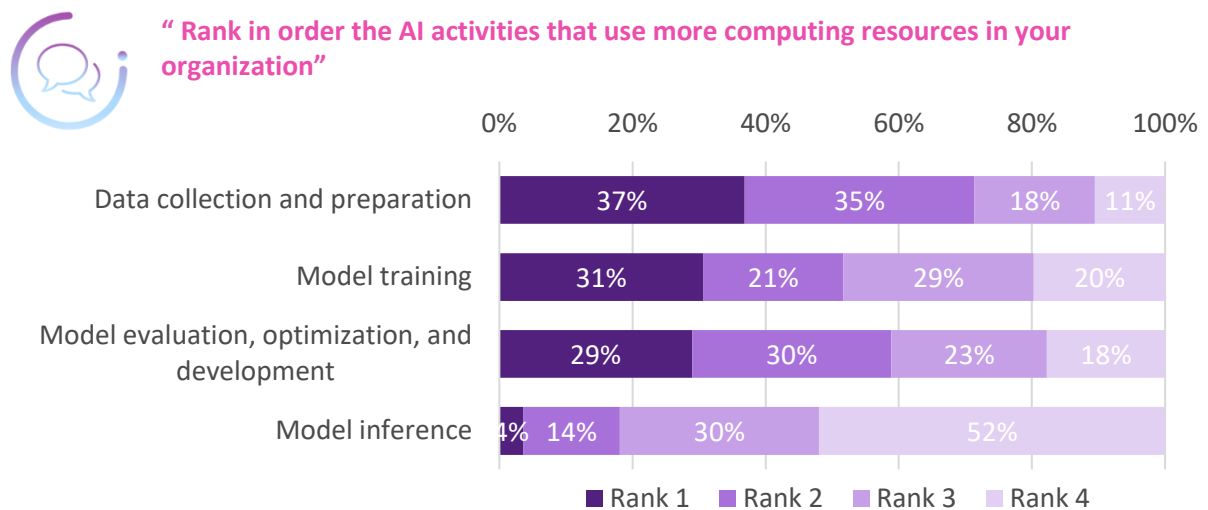


Incorrect Assumptions

- Models are relatively stable once trained
- Releasing new AI applications is a relatively rare event
- Inference is served to a web-scale audience

Instead, in the context our respondents work in, new models and applications are released frequently, as are new data sets, and inference is not necessarily served at scale. To borrow a term from manufacturing industry, “mix” is predominant rather than “throughput”, and consequently flexibility is crucial. In fact, the survey results in Figure 7 showed that respondents were using substantially more computing resources for data preparation, model training, and model evaluation than they were for inference serving. Infrastructure engineers, interestingly, thought training was a bigger workload than data preparation, while other groups thought the reverse. **This is a reminder that it’s important to have visibility over your compute utilization across your workloads** – with monitoring tools such as the infrastructure dashboard in cnvrg.io.


Figure 7: Model development is a bigger deal for enterprise users than inference



Source: Omdia

The results in Figure 7 seem counterintuitive, but this may just be a consequence of projecting the needs of hyperscale companies onto enterprises in general. If your AI application runs inference for every video uploaded to a global social media platform, it's inevitable that inference will be a very large workload even if the model is retrained nightly. However, there are a lot of enterprise AI applications that will never attract those volumes of traffic, even though they may be very valuable indeed. Figure 8 shows a range of different AI applications in a notional banking scenario.

Figure 8. Example AI application scenarios



	Line of business	Typical data	Scope	Inference volume
Bespoke modelling	Investment banking	Valuation, regulatory capital	Individual deals	< 100
Derivatives pricing	Structured finance	Pricing, value at risk	Products	100 – 10,000
Decision support	Trading	Interest rate	Markets	1,000 – 1 Million
Credit scoring	Retail banking	Default probability	Customer base	> 1 Million
Transactional fraud detection	Operations	Binary classification	Whole bank	> 1 Billion

Source: Omdia

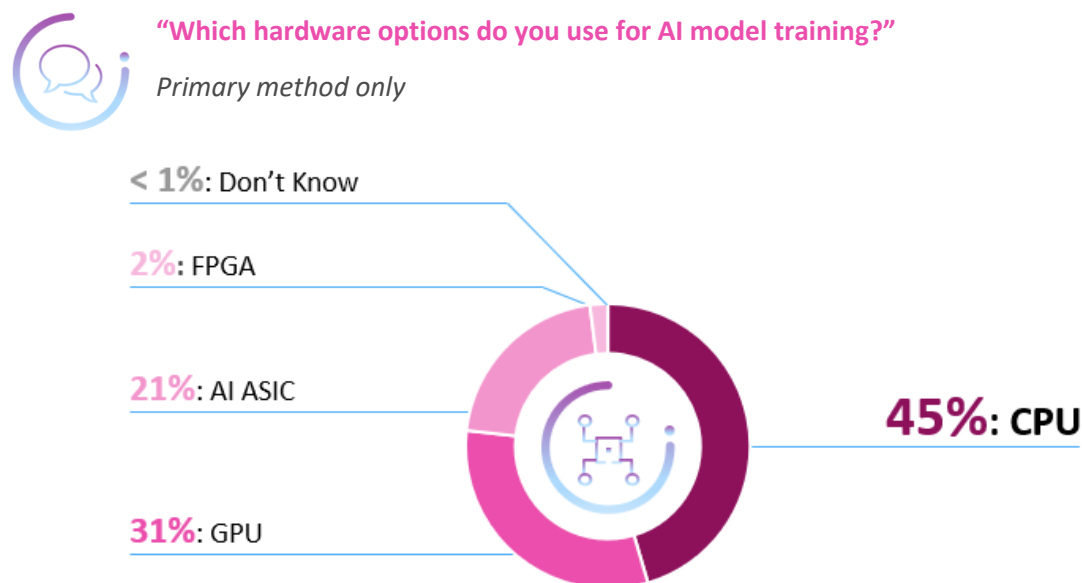
If, as one of our interviewees said, “our training/inference ratio? I would say 1000:1. Inference scalability isn’t a thing for us”, it follows that hardware choices will be rather different from those facing a hyperscale operation serving the same model to an Internet-sized audience. Flexibility and mix will dominate over headline TOPS numbers and throughput. The mix of configurations across the fleet of servers will have to be richer in CPU compute, and indeed, that’s what we found in the survey results. 45% of respondents said that CPUs were their primary AI model training option.

“Our training to inference ratio? I would say 1000:1. Inference scalability isn’t a thing for us”

*Executive/Engineer
Investment Bank*

One reason for this may be the implementation gap discussed in Figure 3 earlier.

Figure 9. 45% consider CPU to be their primary training option



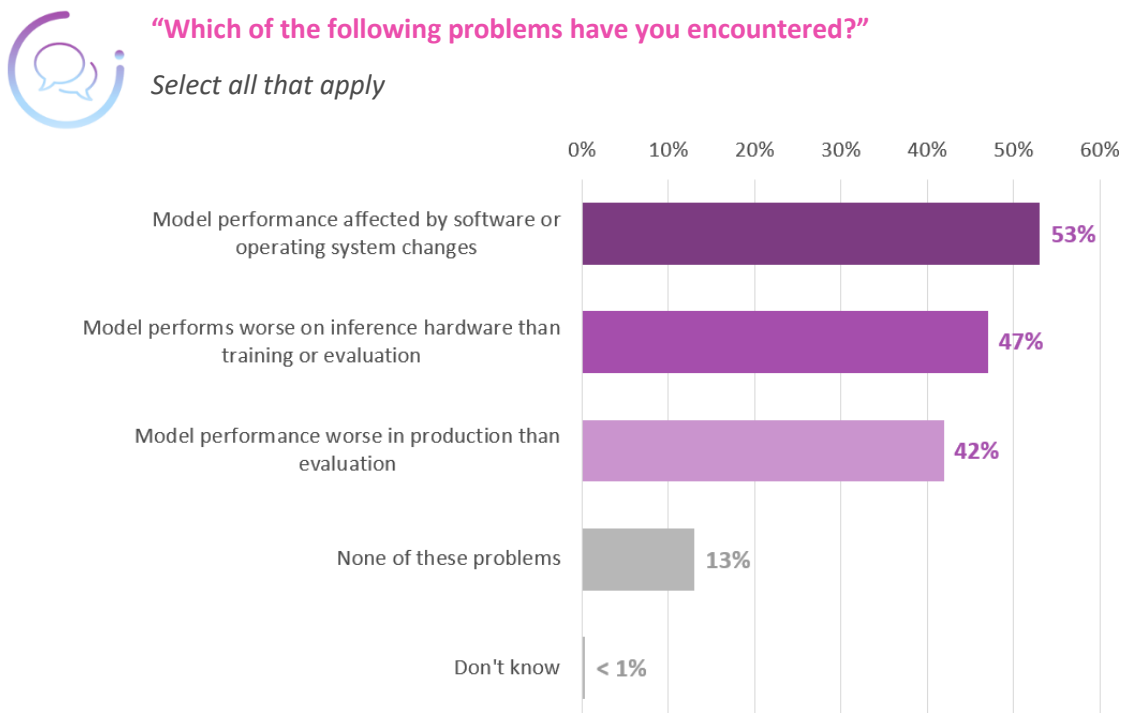
Source: Omdia

Hardware to beat the implementation gap

It is notoriously common that AI models fail evaluation or perform worse in production than they do in evaluation, for a variety of reasons. One of the most common of these reasons is that the model proves to be sensitive to operating system, implementation, compiler, or hardware quirks, and behaves differently after transfer from a development environment to a production environment. This may just be an incremental loss of accuracy or of operational performance such as inference throughput or latency, or it may mean something systematic, for example, the emergence of an edge case where an image classifier reliably misclassifies certain examples.

“Models losing performance in the transfer into production was very common indeed...it’s therefore important to think about hardware early on in the process, and to see the choice of hardware as part of your software and model development process.”

Figure 10. Nearly half our respondents have experienced problems transferring models from the development environment into production

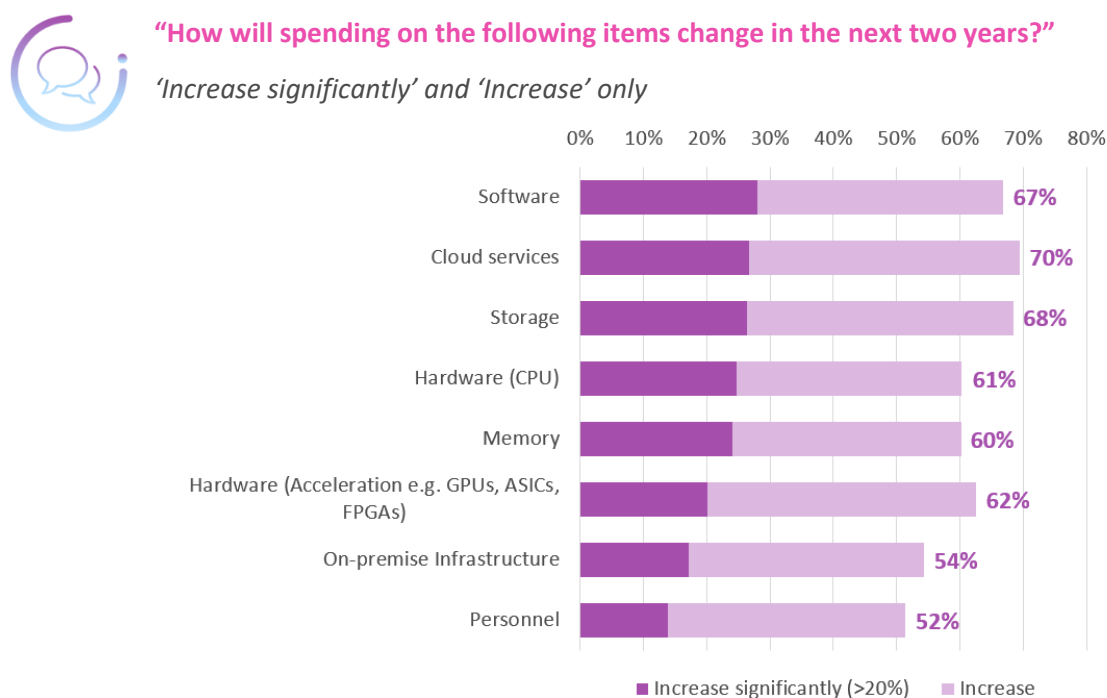


Source: Omdia

This phenomenon is especially problematic, as one of the main strategies AI developers have adopted to cope with the increasing size of the models is hardware specialization – for example, using dedicated inference accelerators, or configuring specific training and inference instance types. By contrast, some engineers prefer a “uniform fleet” strategy that minimizes differences between the training, evaluation, and inference environments. This strategy trades off ultimate performance in favor of a more efficient software and model development process. Indeed, the survey also found that whilst chip performance is the top overall consideration in selecting AI hardware, compatibility and software tools are close behind and even ahead in some verticals (Government and Financial Services). In Figure 10, we ask respondents about their experiences of AI model performance problems.

“Indeed, the survey also found that whilst chip performance is the top overall consideration in selecting AI hardware, compatibility and software tools are close behind and even ahead in some verticals.”

Figure 11. Enterprise spending plans by category



Source: Omdia

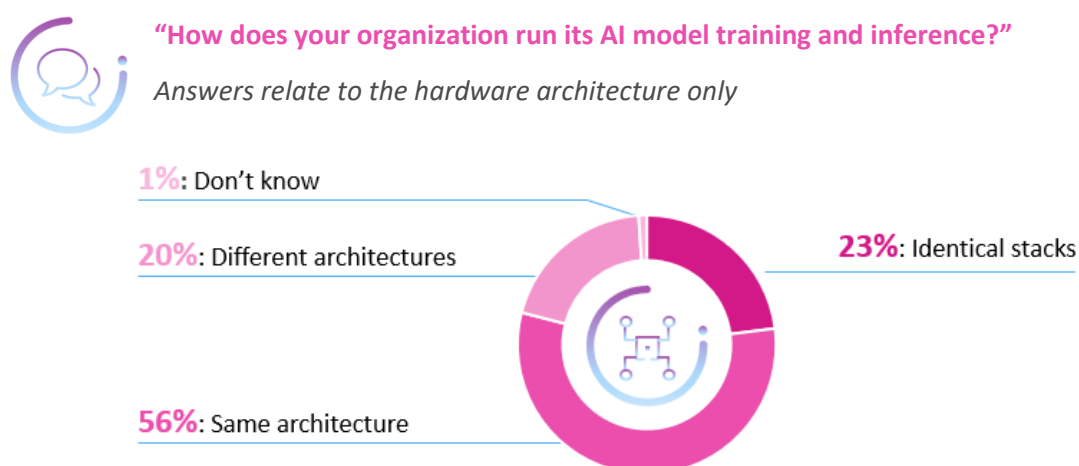
Models losing performance in the transfer into production was very common indeed, and it is worth noting that one reason to have a uniform fleet of machines is precisely in order to have a uniform software stack. It’s therefore important to think about hardware early on in the process, and to see the choice of hardware as part of your software and model development process. An important consequence of this is that despite everything, our respondents are still more likely to increase CPU investment by over 20% than any other logic category.

Interestingly, personnel came last even though enterprises with over 5,000 employees gave a lack of right people and skills as their joint top barrier to greater AI adoption – perhaps reflecting a difficult

hiring market, as if you can't hire you're unlikely to be increasing your budget much. This skills gap is very likely to be a major contributor to the implementation gap we identified earlier. Cnvr.io offers a marketplace of AI Blueprints - curated, tested machine learning pipelines that can be deployed from within the cnvr.io MLOps platform, to close that gap, and enable more developers to deliver value from AI.

Building the infrastructure for an efficient development process needs the widest possible selection of CPU hardware with a common software ecosystem. Figure 12 shows that our respondents were most likely to use a common microarchitecture across training and inference, but not necessarily the same processor.

Figure 12. Over half the respondents are using the same architecture for training and inference, but not the same chips



Source: Omdia

One way to access a wide selection of processors and to scale up in the CPU domain quickly is of course to use the cloud. Cloud services are the second-highest growth item in Figure 11, and this comes as no surprise when we look at the interviewees' remarks. As one of them says, “scaling up in CPU is a great idea in terms of time, if you're not afraid of Amazon Web Services bills”. This interviewee even specified their favorite AWS EC2 instance type, c5d.9xlarge, based on Intel Xeon Platinum Cascade Lake processors.

“Scaling up in CPU is a great idea in terms of time, if you're not afraid of Amazon Web Services bills.”

*Data Scientist
Airline*

The lesson here is that drawing on more CPU compute from the cloud permitted our interviewee to save on developers' and data scientists' time, by having a more efficient development process and fewer problems transferring models across the implementation gap into production, while retaining flexibility to cope with a diverse variety of problems, data sets, and model architectures. Further, the cloud business model, as usual, allowed them to substitute OPEX for CAPEX and to use upward and downward scalability to manage costs.

Conclusions



As the enterprise AI market shifts gears and adoption ramps up, we see that developers are working with more models, more complex/diverse models and more model training projects than ever before. Good AI outcomes therefore come from considering these needs from the outset, in enabling developers to deliver.

Real enterprises deploying AI need, above all, to make it part of an efficient software development process, running on infrastructure that can cope with a high ratio of training and development work to inference. This implies they need a wide selection of performant CPU-based instance types to support their developers and data scientists. Forthcoming CPUs are increasingly likely to include AI inference acceleration as part of the core instruction set – using technologies such as Intel’s AMX (Advanced Matrix Extensions).

Asked about the criteria they use to select processors, respondents to Omdia’s survey put software compatibility and support as either a close second to performance or an effective tie. This should be no surprise, seeing as over half the respondents had experienced an AI model that performed worse in production due to software or operating system quirks and just under half had experienced similar problems porting their models from training to inference hardware. Projects such as cnvrg.io and Intel OneAPI exist to provide a better software interface layer over CPUs and indeed other processor types.

In the enterprise, AI is not just about high-end server GPUs or exotic accelerators – for many, many users, dense cloud instances with high performance x86 CPUs therefore remain a strong choice.

“Real enterprises deploying AI need, above all, to make it part of an efficient software development process, running on infrastructure that can cope with a high ratio of training and development work to inference. This implies they need a wide selection of performant CPU-based instance types to support their developers and data scientists.”

Appendix

Methodology

Omdia surveyed 304 respondents from enterprises, in the categories “AI App Developer”, “Data Scientist”, “Data/ML Engineer”, “Infrastructure Engineer”, and “Executive”, evenly split between EMEA, North America, and Asia Pacific. Omdia further carried out a number of short semi-structured interviews and made use of results from other Omdia products such as the *Enterprise AI Contracts Tracker*. Enterprises included firms in the following fields:

- Telecommunications
- Industrial/Manufacturing
- Financial Services
- Healthcare
- Retail
- Energy/Utilities
- Government

Author

Alexander Harrowell

Principal Analyst, Advanced Computing
alexander.harrowell@omdia.com

About Intel

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com.

Cnvr.io

Find the range of Cnvr.io AI Blueprints [here](#).

Get in touch

www.omdia.com
askananalyst@omdia.com

Omdia consulting

Omdia is a market-leading data, research, and consulting business focused on helping digital service providers, technology companies, and enterprise decision-makers thrive in the connected digital economy. Through our global base of analysts, we offer expert analysis and strategic insight across the IT, telecoms, and media industries.

We create business advantage for our customers by providing actionable insight to support business planning, product development, and go-to-market initiatives.

Our unique combination of authoritative data, market analysis, and vertical industry expertise is designed to empower decision-making, helping our clients profit from new technologies and capitalize on evolving business models.

Omdia is part of Informa Tech, a B2B information services business serving the technology, media, and telecoms sector. The Informa group is listed on the London Stock Exchange.

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help your company identify future trends and opportunities.

Copyright notice and disclaimer

The Omdia research, data and information referenced herein (the “Omdia Materials”) are the copyrighted property of Informa Tech and its subsidiaries or affiliates (together “Informa Tech”) or its third party data providers and represent data, research, opinions, or viewpoints published by Informa Tech, and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice and Informa Tech does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an “as-is” and “as-available” basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness, or correctness of the information, opinions, and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa Tech and its affiliates, officers, directors, employees, agents, and third party data providers disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa Tech will not, under any circumstance whatsoever, be liable for any trading, investment, commercial, or other decisions based on or made in reliance of the Omdia Materials.