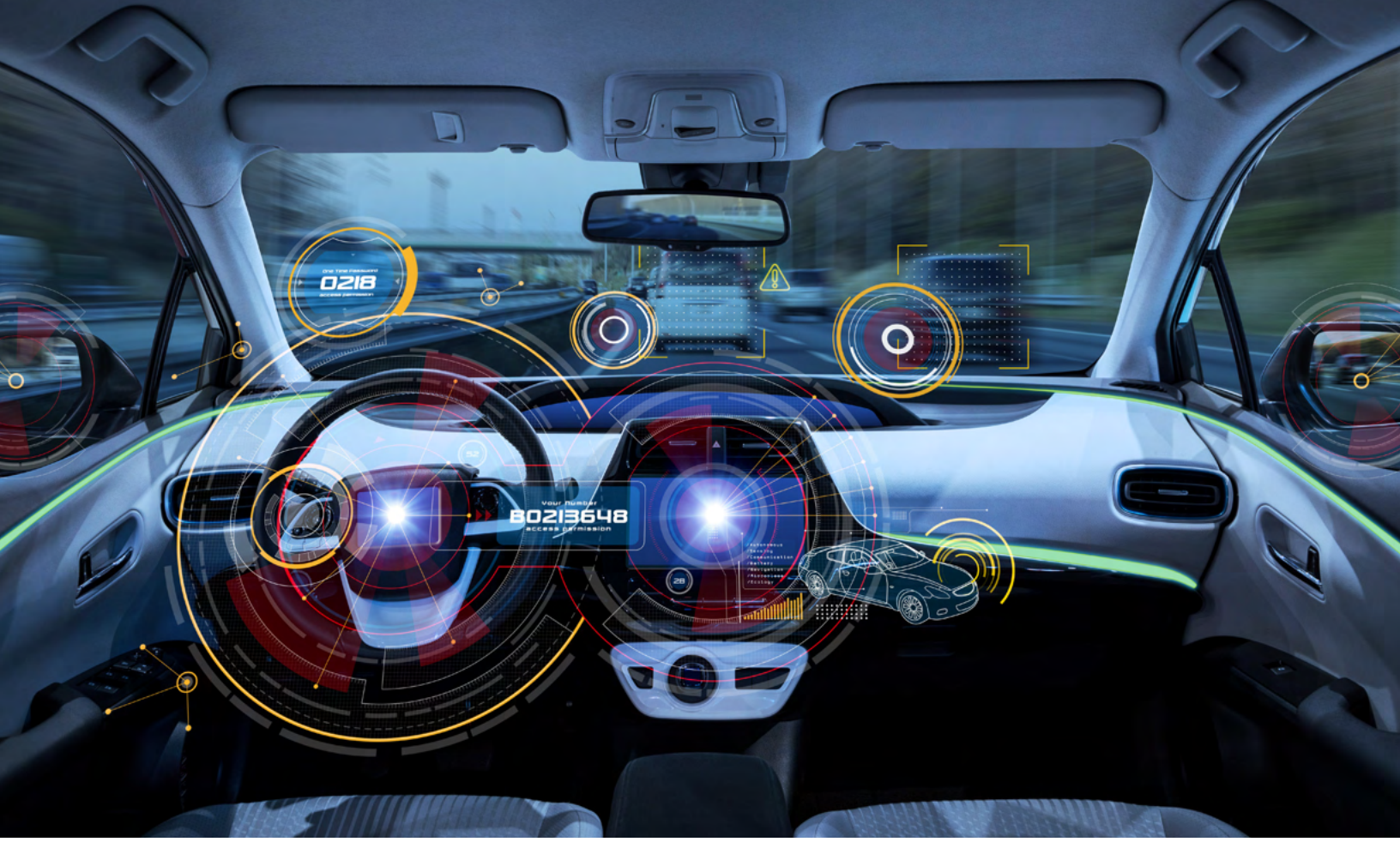
L&T Technology Services

Software Defined Vehicle, Generative AI, and LLM: A Perspective for OEMs

Table of Contents



GLOBAL LANDSCAPE: RISE OF THE SDV

The Software Defined Vehicle (SDV) has emerged as an area of priority investment for OEMs worldwide. While most new-generation OEMs are on an SDV-first approach, major legacy participants are adopting roadmaps to migrate from legacy to SDV architectures. The transition, by enabling a 'continuous revenue and services delivery' platform, is expected to drive significant value for OEMs, their suppliers, and end customers.

A single SDV in operation generates large volumes of data, which can be stored in a repository by the OEM, or an intermediatory, for analysis and use. The scenario becomes complex, as well as throws up immense possibilities, when we consider the millions of SDVs that are set to become operational over the next three to five years. The potential use cases that can be built on this huge dataset available is a long list, and can provide valuable insights for OEMs, vehicle owners, and other participants.

From a technology perspective, for some time now, the rising SDV ecosystem has emerged as an obvious target for AI intervention, given the size of the data and the use cases possibilities.

A DYNAMIC ECOSYSTEM: GENERATIVE AI AND LLMS

Developments over the last two years, especially the rapid advancements being made in the field of Artificial Intelligence (AI) holds significant promise for a dynamic technology ecosystem. With the advent of Generative AI (Gen AI), Foundation Models, and Large Language Models (LLMs) we are witnessing the emergence of newer avenues where voluminous data streams, such as the ones available from the SDVs, can be leveraged for maximum impact in the shortest possible time.

At L&T Technology Services (LTTS), initial studies and pilots have led us to believe that there are at least five key areas where the application of these new techniques can result in transformative outcomes within a short span of time.

01. Software Development Life Cycle (SDLC)

We feel that currently, there is very less repeatability and reuse of codes within teams and departments, even if the functions that these teams are writing are the same. Code duplication is not the only effort that goes in, the associated effort and time invested in quality, testing and security is a waste of valuable resources.

There is a huge potential in this area for leveraging Gen AI and LLMs.

How can one reduce effort by reusing proven, tested, and approved code? With companion tools like Copilot from Microsoft, Duet AI from Google, and Code Whisperer from AWS, it is now possible to do so. These tools assist developers in software development and enable code reuse.

In an initial experimentation with a customer, we observed that structured and monitored code reusage resulted in approximately 35-40 percent of reduction in efforts and time. This results in considerable cost savings, helps streamline operations, and enables a faster go to market.

The need here is to ensure that guidelines and processes are in place for developers to leverage these tool, including, monitored access to code repositories, defining policies on usage of artefacts from open source, enabling real-time interventions, etc.

SDLC is a promising area where Gen AI and companion tools can be leveraged given the ease of use and value unlocked across effort, quality, and time.

02. LLM's for Test Automation/ Validation

Test Automation has been a priority for everyone with a keen interest in reducing costs, minimizing efforts, and shortening cycle time. Emergence of LLMs like LRM and others are opening possibilities of transforming the entire spectrum of automation.

LLMs can be leveraged to create test cases directly from requirement documents in any language via simple, two-step processes. This involves the requirement document being interpreted directly to build a list of test scenarios. Each test scenario has the required test steps detailed in specific formats. This can even be extended to automated test execution on the target device.

At LTTS, we have undertaken three successful pilots in this area with very promising results. Initial indications point to a marked reduction of efforts across the whole process in the range of 30-45%. This also gives rise to new possibilities wherein over a period time, LLMs trained for OEM-specific data could drive higher accuracies, deliver consistent quality, and more importantly, increase test coverage and drive greater automation from the initial phase.

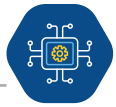
03. LLMs for AD/ADAS

With the emergence of foundation models like Florence for object detection in AD/ADAS use cases, we can start looking at how to increase the quality of images captured during poor lighting

conditions. While the damage to sensors, poor image quality, bad weather, poor lighting condition, and corner cases continue to be a challenge, LLMs like Florence, with expertise in object detections can be fine-tuned for driving significantly improved outcomes. This scenario is especially effective for the growing fleet of SDVs worldwide.

For better results, the LLM will have to be developed, trained, and fine-tuned based on voluminous OEM data streams. Despite the need for training, we feel that it is a promising model to address existing challenges and drive future opportunities and avenues for improvement. We are exploring on the modalities of a pilot using Florence for object detection.

04. Embedded AI



As we move toward inferencing and processing at the edge (involving a controller or a camera for SDVs) it is important to look at potential avenues for driving model optimizations. This includes leveraging low compute and memory footprint, key features enabled by edge devices. New techniques in quantization and pruning can prove to be a major differentiator in this regard.

Leading global SOC vendors are now delivering more advanced toolkits that enable deeper model optimization pathways. At LTTS, we are investing in an embedded AI framework that enables a two process model optimization technique. The first stage involves running it through our framework to get a model that is platform agnostic. In stage tow, we can undertake further optimization with the tools provided by the target device vendor. We have already recorded a reduction of 70% in compute power needs in the optimized model.

As OEMs continue to build models for enabling different functions and features in the SDV ecosystem, one can easily envisage a large catalogue of use cases to be leveraged across various scenarios. This underscores the need to focus on model optimizations as an inherent requirement for effective model development.

A challenge that comes across is finding the acceptable balance of optimization versus accuracy. While sometimes, a highly optimized model does show a drop in accuracy by a small percentage, we feel that this is a decision that will always have to be exercised based on the criticality of the feature in enabling the overall capabilities of the offering within the SDV.

05. NLP LLMs

LLMs like Llama, GPT, T5, etc., are NLP (Natural Language Programing) based models and can address different text-based requirements. These LLM's, when trained adequately, can be leveraged across various use cases like enabling interactive question/answer interactions between the user and the user manual on any topic that a user might want to know and is covered in the manual.

The LLMs can also be trained to address complex use cases like AutoSAR configuration where it can extract unstructured values from specification documents, combine with tabular data, and generate AutoSAR configuration in JSON format.

Our experience further indicates that the true impact of AI is evident when it is allowed to cut through the layers of computing. LTTS feels that for effective leverage, we need to operate across all layers of computing, as is illustrated in Figure 1 below:

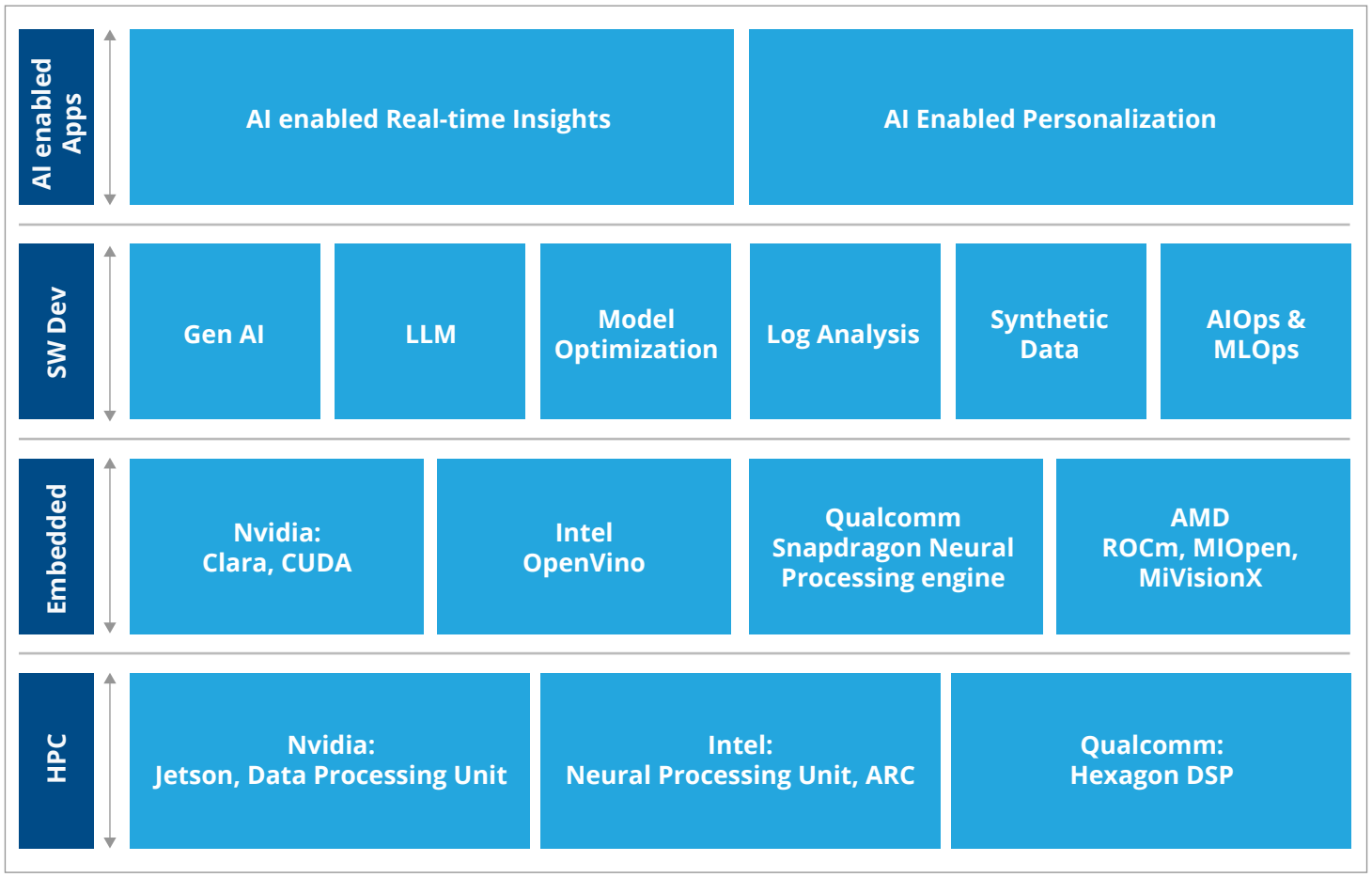


Figure 1: AI: Increasing Influence across layers of Computing - LTTS Footprint

CHARTING THE FUTURE: A PERSPECTIVE FOR OEMS

As technology evolves and greater computing power is made available in a more efficient and accessible manner, we can identify several areas that present themselves as potential ground for adopting Gen AI and LLM-based approaches. However, there are two very important aspects to be considered while considering LLM's.

Choosing the right LLM is a very serious activity and this requires a very good understanding of the proposed model as well as the areas where it will be used. One needs to be very detailed about this research. Our work in this area shows that there is a very clear difference in performance timings and accuracies at a subtask level when one compares different LLMs. Table 1 illustrates some of our findings in this direction.

Task Category	Task Name	Baseline	FLAN T5-large - 750m (Google)	FLAN T5-xxl - 13B (Google)	FLAN UL2 -20B (Google)	LLAMA2-7B (Meta)	Falcon 7B Instruct (Technology Innovation Institute)
Information Extraction	Title Identification	GPT 3.5 - OpenAI (ChatGPT)	Good	Good	Good	Very Good	Average
Information Extraction	Extractive QA		Good	Good	Good	Good	Average
Information Extraction	Abstractive QA		Very Bad	Very Bad	Very Bad	Good	Bad
Reasoning Tasks	Arithmetic Reasoning		Very Bad	Very Bad	Very Bad	Very Bad	Very Bad
Reasoning Tasks	Commonsense Reasoning		Average	Good	Good	Very Good	Very Bad
Reasoning Tasks	Explanation Generation		Bad	Bad	Bad	Very Good	Bad
Reasoning Tasks	Cause effect identification		Very Bad	Very Good	Very Good	Very Good	Very Good
Text Classification	Topic Classification		Average	Good	Good	Good	Very Bad
Text Classification	Toxic lang Identification		Good	Good	Good	Good	Very Bad
Text Generation	Summary Generation		Bad	Bad	Bad	Very Good	Very Bad
Text Generation	**Context Generation**		Bad	Bad	Bad	Very Good	Bad
Text Generation	Code Explanation		Very Bad	Very Bad	Very Bad	Very Good	Very Bad
Text Generation	Code Explanation		Very Bad	Very Bad	Very Bad	Good	Very Bad
Translation	Translation		Very Good	Bad	Bad	Very Bad	Very Bad

Table 1: Open Source LLM Performance on Various NLP Tasks*

Subjective Score	0 - 0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8 - 1
Rating	Very Bad	Bad	Average	Good	Very Good

"Subjective Score" is human evaluation score after comparing output from ChatGPT & listed LLMs on various generic NLP Tasks

Our teams created comparison charts for various NLP tasks with popular, open source LLMs that can be deployed on moderate GPU configurations. We found that:

- Smaller model (Flan-T5) perform well with non-generative tasks such as "Information Extraction" , "Text Classification," and "Language Translation" at acceptable levels, while
- LLAMA2-7B performed well on most of NLP tasks except Arithmetic reasoning and translation.

The second area of importance is around ensuring that LLMs are trained properly for improved accuracy. We need be remember here that LLMs are particularly effective in capturing the complex relationships between words and phrases in natural language, and can achieve significant results across a variety of NLP tasks. The training of LLMs is data-intensive, and it requires significant computing resources and specialized algorithms to process large amounts of data.

Table 2 (below) illustrates the evaluation of LLM performance using foundation models. Our teams created a comparison chart for summarization and Q&A task with popular LLMs that can be deployed on moderate GPU configurations. The focus was on striking a balance between inference time and accuracy, and we found LLAMA2 8 bit as being more memory efficient with acceptable accuracy.

LLM Name	GPU Memory	Bench mark against OpenAI ChatGPT	
		Summarization	Q&A
flan-ul2 (20B)	76372MiB / 81920MiB	Bad	Average
flan-xxl (13B)	43332MiB / 81920MiB	Bad	Good
llama2-7b-chat-hf-32bits	26700MiB / 81920MiB	Good	Good
llama2-7b-chat-hf-16bits	13772MiB / 81920MiB	Good	Good
llama2-7b-chat-hf-8bits	8068MiB / 81920MiB	Good	Good
llama2-13b-chat-hf-32bits	50765MiB / 81920MiB	Very Good	Very Good
llama2-13b-chat-hf-16bits	25983MiB / 81920MiB	Very Good	Very Good
llama2-13b-chat-hf-8bits	25783MiB / 81920MiB	Very Good	Very Good
wizardML-13B	50685MiB / 81920MiB	Very Bad	Very Bad
TheBloke_stable-vicuna-13B-HF	25943MiB / 81920MiB	Good	Good
Open-Orca_OpenOrca-Platypus2-13B	43740MiB / 81920MiB	Very Bad	Very Bad
palm 2 (hat-bison@001)	API based	Very Good	Very Good

Table 2: Foundation Models Quantized – Evaluating LLM performance*

Subjective Score	0 - 0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8 - 1
Rating	Very Bad	Bad	Average	Good	Very Good

"Subjective Score" is human evaluation score after comparing output from ChatGPT & listed LLMs on various generic NLP Tasks

As indicated earlier, improving the accuracy of LLMs can have a significant impact on the performance of natural language processing applications, including helping improve the accuracy of machine translation systems. This can have a significant impact on cross-border communication and global businesses, besides helping improve the accuracy of search and text classification systems, and streamlining and revitalizing information retrieval and knowledge discovery paradigms.

As LLM's continue to evolve rapidly, another area where we can effectively leverage the capabilities unlocked by these models is in the development of vision computing for enabling Autonomous Driving/Autonomous Driving Assistance Systems (AD/ADAS). This includes using Segment Anything Models (SAM) for image segmentation, or leveraging Contrastive Language-Image Pre-training (CLIP) or Florence for image classification, segmentation, and generation. The span of these models is clear from the size of the data sets used, ranging from about 100 million parameters for SAM to about 1 billion for Florence. Figure 2 illustrates the Foundation Models for Vision Computing.

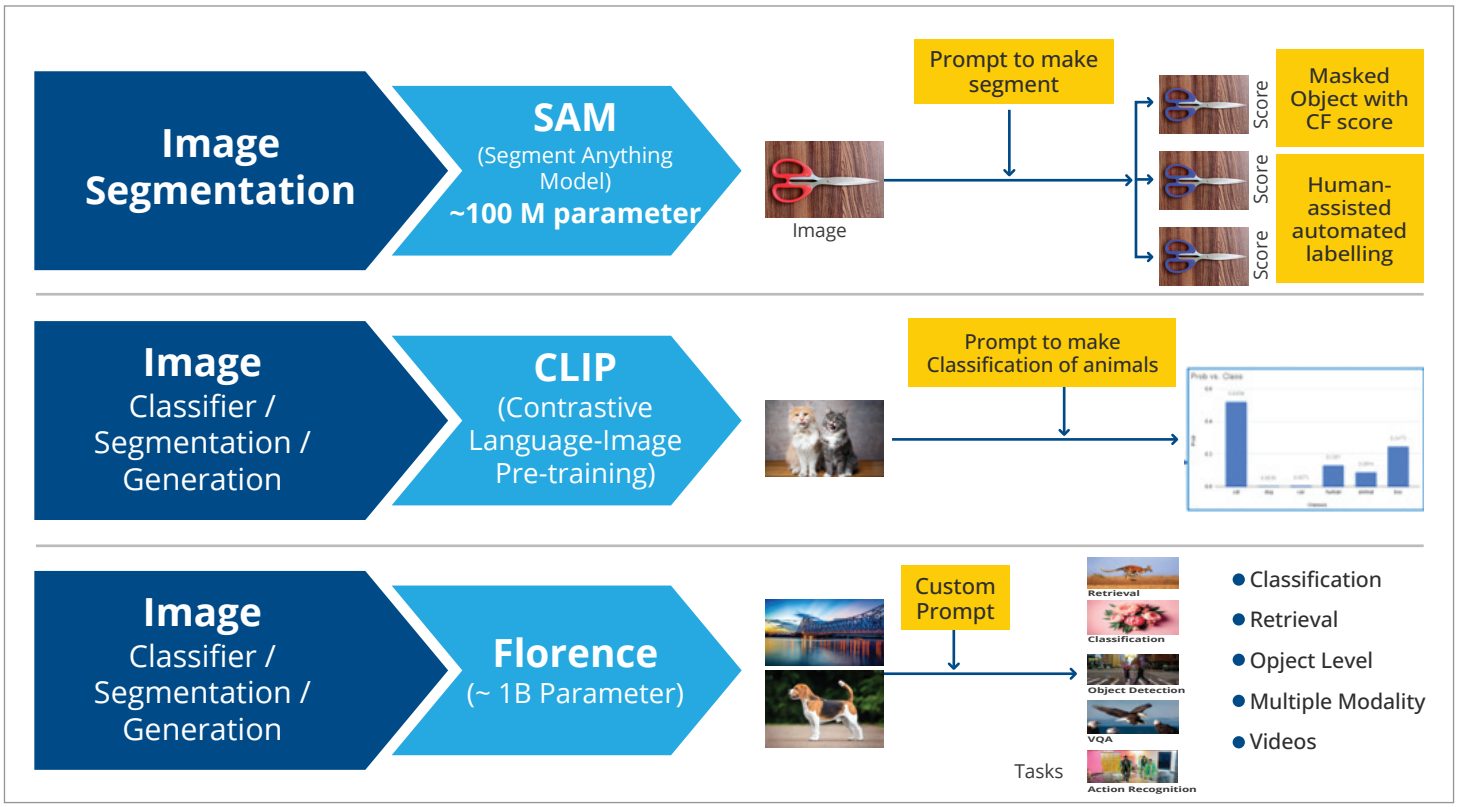


Figure 2: Foundation Model for Vision Computing: AD/ADAS

All these models are significantly better than Classical Dense CNN/GAN models in most of the tasks needed for enabling the next level of AD/ADAS performance.

LTTS EDGE: ENABLING NEXGEN BUSINESS SUCCESS JOURNEYS

At LTTS, we see the emergence of these Gen AI and LLMs as a major development that will have a significant impact in the way we write software. The rapid maturity of these tools promises faster adoption across the board, especially in the area of SDVs.

It is important to start exploring these tools now, simply for the reason that the effectiveness of these tools will completely depend on the OEM willing to invest in making them meaningful by providing the requisite data. The training of the LLMs, combined with the investments in computing resources for the right Gen AI tool, will enable the creation, delivery, and sustenance of a transformative experience and revenue paradigm. In turn, this would help unlock new value streams for both customers and OEMs across the global spectrum.

And finally, we should continue to be aware of the importance of model optimizations for the LLMs to be ported onto edge devices or controllers. Increasingly, large amounts of compute power is now being made available on the edge devices. As a result, it is now possible to look at how this additional computing power can be harnessed for performing intelligent tasks at the mode level itself. To achieve this, pruning, quantization and optimization techniques are maturing and are available for model optimizations to take place. The focus therefore, going ahead, will be on the next level of digital engineering in unlocking the true value from the models.

AUTHOR

Ashish Khushu

Chief Technology Officer, L&T Technology Services

Ashish Khushu is Chief Technology Officer (CTO) at L&T Technology Services (LTTS) and is responsible for the company's technology roadmap, innovation charter, and investments in solutions and platforms.

L&T Technology Services

ENGINEERING **THE CHANGE**

www.LTTS.com

info@LTTS.com