White Paper

intel ai

Decentralizing Generative AI (GenAI) Inference

On-Device Deployment of Lightweight Open Source GenAI Models, Including Large Language Models (LLMs), Can Improve Accessibility and Latency

Author

Bharath G. Srivats

Product Marketing Engineer

Credits

Sanjay Addicam

Ilia Efimov

Pinkesh Shah

Balaji Srinivasan

Michael Campbell

Helena Kloosterman

Michael Hansen

Abstract

Bringing AI inference closer to the data source offers significant advantages in cost, privacy and performance. Recent advancements in lightweight GenAI models (i.e., 1-8B parameters) provide a disruptive opportunity to shift GenAI deployment from the cloud to the edge, but alternatives to cloud-based GenAI need to be practical and efficient. This white paper outlines a strategic approach to shift GenAI deployments from cloud-native (i.e., GPU based) solutions to edge (i.e., hardware based) solutions using the built-in compute acceleration of CPU-GPU-NPU (e.g., Intel® Core™ Ultra processors, Intel® Arc™ GPUs) and open source GenAI models. On-device deployment offers low total cost of ownership (TCO), offline capabilities, data sovereignty and reduced latency, making powerful GenAI models accessible across regions and sectors that may previously have faced barriers to deployment.

Table of Contents

Introduction

Large, sophisticated GenAI models have immense computational needs, and as a result, cloud-based GenAI solutions have dominated. However, these models require substantial infrastructure investment or recurring end-user API costs, which limits access to their advanced capabilities.

Recent advances in GenAI open ecosystem development are changing the landscape; models are optimized to be smaller, faster and less power hungry, enabling edge-based, decentralized deployments with the ability to run GenAI models locally. This shift from cloud-based AI towards offline, edge-based AI presents a compelling deployment alternative — one that democratizes GenAI access away from centralized infrastructure and offers benefits like reduced costs and increased data privacy by leveraging hardware (CPU-GPU-NPU).

On-device GenAI could have wide-reaching positive impacts, increasing access to GenAI tools in circumstances, locations or sectors that face challenges like lack of internet, low bandwidth, shortage of skilled people, etc.

Why it Matters

Low Cost

Localized deployments of on-device GenAI applications do not require incremental subscriptions or fees. By optimizing lightweight models for existing underlying hardware configurations (i.e., leveraging sunk costs), organizations can achieve competitive GenAI performance at a fraction of the cost associated with cloud-based instances.

Education:

UNICEF estimates that nearly 1 billion children[1] may face unreliable internet access and a shortage of qualified teachers. Low cost, offline access to GenAI-powered educational tools presents an opportunity to bridge this gap.

Data Privacy

On-device GenAI is complementary to agentic AI workflows[2], offering a local inference tier that can function offline at low latency and act upon private data enclaves for sovereign data intelligence.

Healthcare:

In crisis response situations or remote locations without reliable internet, healthcare providers could access GenAI tools that support diagnostics, triage or secure access to critical patient records.

Latency

While cloud-based GenAI scales for high-end workloads, it often entails network latency. Lightweight models with low latency and on-device compute offer a hybrid approach; AI workloads can be allocated between the cloud and on-device depending on user requirements, resource utilization and desired quality or user experience.

Automotive:

Low-latency, on-device GenAI tools could automate certain workflows at the point of decision-making or help streamline repairs at the point of service.

Technical Performance Benefits of On-Device GenAI

Edge solutions, in contrast to compute-intensive advanced cloud-based models, offer a balance between performance and computational efficiency. The following table demonstrates how recent advancements in lightweight, open-source GenAI models and Intel OpenVINO™ toolkit-based optimizations enable efficient deployment and orchestration on Intel® Core™ Ultra processors[3].

Example Solution Ingredients:

To demonstrate how on-device GenAI deployment can replicate cloud-based deployment and functionality, we ran the following scenarios on OpenVINO: chatbot, image generation and live captioning/audio transcription.

- **LLM (Large Language Model):** Llama, Phi, Qwen, SmolLM, Mistral, etc.
- **LVM** (Large Vision Model):** MiniCPM-V, Qwen-VL, Llava, SmolVLM, etc.
- **Image Generation**:** Stable Diffusion, Flux
- **Audio Transcription**:** Whisper
- **Console**:** OpenVINO Test Drive

Benchmarking GenAI on OpenVINO

As the following table illustrates, we found satisfactory performance (i.e., 20-30 TPS) on real-life use cases such as conversing with a chatbot.

Our benchmarking demonstrates that Lunar Lake (LNL) is approximately 1.6x more power efficient at equivalent* iGPU AI inferencing compared to Meteor Lake (MTL); and LNL is approximately 1.4x faster (i.e., throughput) at equivalent* iGPU AI inferencing compared to MTL. Power draw and other utilization results are shown below.

iGPU*^	Models, HW:	Throughput TPS (2nd tkn+)	Latency (ms) (2nd token)	At Full iGPU Utilization*, Measured:	Power Draw*
Llama	Llama 3.1-**8B**-int4, **MTL**	**14.20*/16.48^**	**70.50*/60.67^**	16.4% CPU utilization 59000 MB/s Mem BW Idle Mem: 12GB Load Mem: 21GB	Idle: 5.5W **Load: 29W**
	Llama-3.1-**8B**-int4, **LNL**	**20.38*/20.91^**	**49.07*/47.80^**	21.8% CPU utilization 78401 MB/s Mem BW Idle Mem: 6GB Load Mem: 15GB	Idle: 2.4W **Load: 25W**
	Llama 3.2-**1B**-int4, **MTL**	**66.06***	**15.13***	16.8% CPU utilization 49698 MB/s Mem BW Idle Mem: 11GB Load Mem: 13.5GB	Idle: 5.1W **Load: 29W**
	Llama 3.2-**1B**-int4, **LNL**	**79.33***	**12.59***	22% CPU utilization 64676 MB/s Mem BW Idle Mem: 6.2GB Load Mem: 8GB	Idle: 2.4W **Load: 23.5W**
iGPU^	**Models & Generations:**	**Throughput^ TPS** (2nd tkn+)	**Latency^** (ms) (2nd token)		
Llama	Llama-3.2-**3B**-instruct-int4, **MTL**	**31.76**	**31.47**		
	Llama-3.2-**3B**-instruct-int4, **LNL**	**40.69**	**24.57**		
Phi	Phi-3-mini-**4k**-instruct-int4, **MTL**	**24.59**	**33.95**		
	Phi-3-mini-**4k**-instruct-int4, **LNL**	**35.21**	**28.4**		
Qwen	Qwen2.5-**7b**-int4, **MTL**	**17.68**	**56.53**		
	Qwen2.5-**7b**-int4, **LNL**	**21.62**	**46.24**		
DeepSeek	Deepseek-r1-distill-qwen-**7b** int4, **MTL**	**17.82**	**56.09**		
	Deepseek-r1-distill-qwen-**7b** int4, **LNL**	**21.68**	**46.12**		
NPU*	**Model, HW:**	**Throughput TPS** (2nd tkn+)	**Latency*** (ms) (2nd token)	**At Full NPU Utilization*, Measured:**	**Power Draw***
NPU data is preliminary and not fully optimized for					
Llama	Llama 3.2-**1B**-int4, **MTL***	**22.91***	**43.5***	**2.9% CPU utilization** 13375 MB/s Mem BW Idle Mem:11GB Load Mem: 13.4GB	Idle: 5.1W **Load: 16.45W**
	Llama 3.2-**1B**-int4, **LNL**	**25.97***	**38.5***	**3% CPU utilization** 27030 MB/s Mem BW Idle Mem: 6.2GB Load Mem: 8.9GB	Idle: 2.4W **Load: 15.0W**

Sources & Notes:

Open-source instruct-tuned models, OV-quantized/optimized, please see sources for additional details
***Config 1:** SKUs tested: Meteor Lake Core Ultra 7 **MTL 165H**, Lunar Lake Core Ultra 7 **LNL 268V**, OpenVINO 2024.6.0.0
*Memory and power usage metrics were measured under a custom Q&A inference workload implemented with OpenVINO GenAI, reflecting measured system resource performance defined test parameters
*Input token size = <25, output token size = 512

^Config 2: SKUs tested: Meteor Lake Core Ultra 7 **MTL 155H**, Lunar Core Ultra 7 **LNL 288V**, OpenVINO 2025.0
^input token size = 32, output token size = 1024
^https://docs.openvino.ai/2025/index.html
^https://docs.openvino.ai/2024/openvino-workflow/model-preparation.html
^https://docs.openvino.ai/2024/openvino-workflow/model-optimization.html

**Coming soon - Whisper audio transcription, image time-to-generate, multimodal models benchmarking, Openvino Test Drive console (beta)
**https://github.com/openvinotoolkit/openvino_testdrive (beta)

Figure 1. Benchmarking GenAI on Intel MTL, LNL and OpenVINO

Cost, Privacy and Accessibility Benefits of Edge-Based GenAI

With Intel® Core™ Ultra processors, Intel® Arc™ GPUs and OpenVINO, enterprises can meet their edge and on-prem GenAI application needs.

Low Total Cost of Ownership (TCO)

Offline edge and enterprise GenAI deployments can supplement reliance on expensive cloud infrastructure, thereby reducing long-term API token costs. Once the hardware is deployed (i.e. sunk cost), open-source GenAI models can run offline, significantly reducing costs over an estimated five year depreciation period. Based on assumptions regarding task-specific performance and precision requirements, enterprises can optimize workload distribution between edge and cloud to best match the demands of each application. In the long run, enterprises may be able to realize significant savings by adopting a hybrid strategy to leverage more on-device functionality relative to cloud-based. As lightweight models' portability and quality continues to rapidly improve, they can be customized to fit in a hybrid framework and utilized for local intelligence.

The decentralization of GenAI through edge devices creates a more accessible entry point for developers and enterprises alike, lowering both capital expenditure (CAPEX) and operational expenditure (OPEX).

No Internet Needed

On-device GenAI models do not require internet to work. This is especially important in emerging markets where intermittent connectivity and poor bandwidth impact operations. For example, an on-device chatbot can be initialized in a matter of seconds from the time the system is powered on and remain available thereafter with near-instantaneous response times and full offline functionality. Furthermore, the ability to deploy multimodal* GenAI models — text, audio, vision inputs and text, audio, vision outputs — on-device expands sophisticated offline use cases and could include image recognition*, image generation* or language translation, to name a few.

Optimized Performance

OpenVINO's sophisticated optimization framework maximizes performance through Intel GPU XMX acceleration and NPU capabilities, allowing maximum flexibility for developers through intelligent workload distribution across architectures. Independent Software Vendors (ISV) can create solutions with models across the GPU+NPU mix. Advanced quantization techniques reduce memory requirements while accelerating inference speeds, making 1-8B parameter models viable for on-device deployments that prioritize privacy protection, performance optimization and cost efficiency.

Improved Latency

Computation (e.g., TOPS) residing next to the data source or device offers fundamental latency benefits for running GenAI. Processing GenAI models directly on prem and local edge deployments reduces latency to milliseconds — a critical advantage over cloud-based GenAI systems where latency can be affected by network bandwidth and remote server loads.

Examples (see Figure 1): Under 100ms* iGPU latency on lightweight LLM enables near-instantaneous intelligence, such as a yes/no that can be derived as alerts or decisions (i.e., agentic). Through sophisticated pipeline engineering (e.g., preprocessing, context loading), a response can be extracted in a few milliseconds, fundamentally accelerating inference for specialized scenarios. This low-latency verdict is ideal for simple yes/no or pass/fail classifications, providing an on-device gating superpower that outpaces cloud latency. This initial low-latency assessment can revolutionize system responsiveness in carefully selected, high-impact applications where immediate categorical decisions drive operational success, such as in healthcare and automation.

Data Sovereignty

On-premise and local GenAI application deployments ensure that sensitive data remains on the device, preserving privacy and adhering to regulations like the General Data Protection Regulation (GDPR) or Health Insurance Portability and Accountability Act (HIPAA). Because they don't need an internet connection to process GenAI in real-time, organizations gain enhanced control over their data while avoiding the risks of data breaches and leaks associated with cloud-based systems.

Cost-Effectiveness

On-premise and edge GenAI solutions present compelling economic and privacy advantages compared to cloud-based solutions. Organizations can significantly reduce or eliminate subscription and API expenses and achieve precise operational cost management with local inference capabilities.

Low Maintenance

Edge and on-premise GenAI models deployed on local infrastructure require only periodic manual updates to stay current. Given that lightweight models evolve rapidly — often in six-month cycles — an annual maintenance schedule is recommended for edge-deployed GenAI. This allows for critical updates to models and optimizations through OpenVINO release updates, ensuring the models remain efficient and effective for their intended use cases.

Across multi-year deployment cycles, model maintenance and updates become increasingly predictable, simplifying maintenance cycles and reducing total expenditure compared to cloud-centric approaches that demand continuous network connectivity and consumption-based pricing models.

The Impact of Decentralized AI on Education in Emerging Markets

Use Case Deep Dive

The deployment of offline LLMs through low-cost on-device GenAI solutions can have a transformative impact on education in emerging markets. With an estimated 1 billion[1] students in regions such as South America, India and Africa facing unreliable internet access and a shortage of qualified teachers, GenAI-powered educational tools present an opportunity to bridge this gap. On-device LLMs can contextualize student requirements, understand their strengths and weaknesses and provide individually tailored responses — all in a matter of seconds while preserving student privacy. By combining local and cloud-based models for a hybrid approach, educational institutions could:

Address the Teacher Shortage with Offline GenAI

In regions where access to quality education is limited, on-device GenAI can be a proxy to learning resources that anyone can access, anytime, for free. LLMs running on low-cost edge devices can offer personalized tutoring, generate content, answer student queries and facilitate self-paced learning, complementing the role of teachers in areas with inadequate educational infrastructure and democratizing access to high-quality learning.

Build Localized Solutions

Language and vision GenAI models can provide diverse educational tools from interactive learning experiences to creative content support while running efficiently on low-power edge devices. Independent Software Vendors (ISVs) and Systems Integrators (SIs) play a critical role in this ecosystem by providing the front-end UI/UX wrappers, ensuring they are user-friendly, intuitive and tailored to specific educational needs. ISVs can develop custom interfaces that cater to both teachers and students, while SIs can focus on the setup, maintenance and activation of the edge GenAI systems. Through customized interface design and user experience optimization, these lightweight LLMs effectively serve diverse educational environments with adaptations such as the ability to translate learning materials into anyone's first language. And when integrated with retrieval-augmented generation (RAG)[4] or extended context workflows[5], these lightweight models align precisely with specific curricular requirements and moderation protocols like safety, ethics, age appropriateness for K-12 segments and others.

Reduce Operational Costs

By using on-premise GenAI, schools can reduce or eliminate subscription or API fees and gain tighter control over operational costs. In other words, they can implement the above solutions in a cost-effective way.

In summary, the deployment of optimized lightweight LLMs on edge devices offers a scalable and cost-effective solution to improving education in underserved regions. By leveraging the rapid evolution of open-source GenAI models and relying on ISVs and SIs for customization and maintenance, emerging markets can take significant strides toward bridging the education gap, bringing high-quality, GenAI-driven learning to millions of students.

Conclusion

Decentralized, on-device GenAI architectures can transform both operational performance and cost efficiency in myriad real-world deployments, from enterprises to emerging markets. Together, Intel® Core™ Ultra processors, Intel® Arc™ GPUs and OpenVINO offer an accessible route to democratizing GenAI by accelerating the development of optimized on-device GenAI solutions.

- Learn more about the AI PC powered by Intel and deployment with the OpenVINO toolkit
- Try OpenVINO GenAI examples on your hardware

Disclaimer

This whitepaper analyzes various AI models, including fully open-source models (e.g., Apache 2.0, MIT-licensed models) and openly available models (e.g., Meta's LLaMA). Each model remains the intellectual property of its respective creators and is governed by its original license. Fully open-source models such as Qwen, Mistral, and Phi allow unrestricted use, modification, and distribution under their respective licenses. Models like LLaMA are available under specific usage terms that may include restrictions. This document does not redistribute, modify, or alter any model weights, nor does it claim ownership over them. Users should consult the official licensing terms of each model before use. The inclusion of specific models in this whitepaper does not imply endorsement by their respective developers. The recommendations presented are based on independent research and do not constitute official guidance from Meta, OpenAI, Google, or any other model provider.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure. Your costs and results may vary. Intel technologies may require enabled hardware, software or service activation. Current characterized errata are available on request.

Statements in this document that refer to future plans or expectations are forward-looking and subject to change without notice. Actual results may vary.

intel ai

Sources

[1] How many children and young people have internet access at home? UNICEF.

[2] Agentic AI and Confidential Computing. Agentic AI is: "a paradigm that involves deploying agents or other programs that act autonomously on behalf of humans or other AI systems to make decisions and take actions to achieve specific goals. Unlike current agent models, which are rules-based, agentic AI uses sophisticated reasoning, knowledge, and intelligence to help drive process automation by making decisions and taking actions rather than just responding to questions."

[3] Optimizing Large Language Models with the OpenVINO™ Toolkit. Intel. April 2024.

[4] What is retrieval-augmented generation? RAG is an AI framework that "helps LLMs deliver more-accurate and -relevant AI responses. RAG supplements an LLM with data from an external knowledge base, ensuring LLMs can access the most-reliable and -current information. This additional data helps LLMs deliver up-to-date and contextually meaningful responses."

[5] What is a context window? McKinsey & Company. December 2024. "A context window is how information is entered into LLMs. The larger the window, the more information an LLM is able to process at once."