

Unlocking the Benefits of On-Device Generative AI

Discover the advantages of on-device generative AI, including privacy, performance, personalization, cost savings, and energy efficiency.

"Empowering devices with on-device generative AI is not just about efficiency; it's about safeguarding privacy, ensuring instantaneous responses, and granting access even when offline."



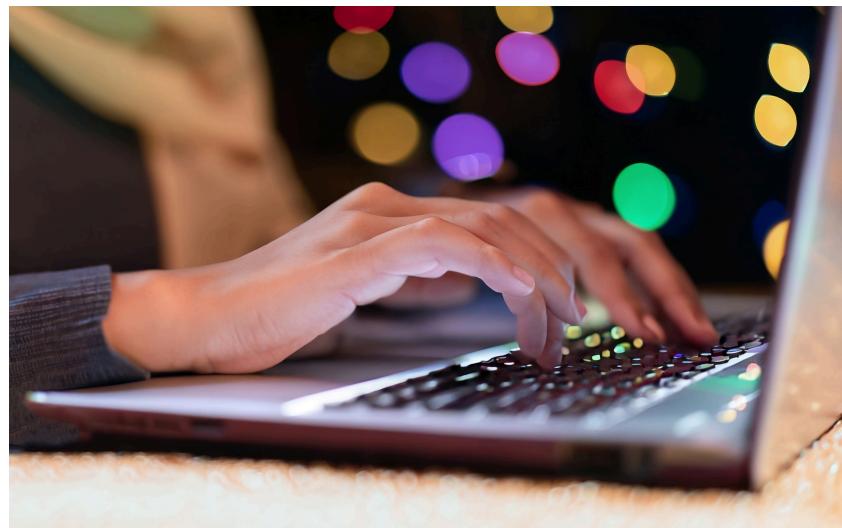
CONTRIBUTOR
BAREN BARUNA HRP
NIM: 23/519317/NUGM/01057

INTRODUCTION

Dalam dunia kecerdasan buatan (AI) yang berkembang pesat, AI generatif pada perangkat menonjol sebagai teknologi transformatif. Teknologi ini membawa kekuatan AI langsung ke perangkat *edge*, memungkinkan pemrosesan waktu nyata tanpa memerlukan koneksi *cloud* yang konstan. Artikel ini membahas pentingnya AI generatif pada perangkat dan menyoroti berbagai manfaatnya, termasuk privasi yang ditingkatkan, kinerja yang dioptimalkan, pengalaman pengguna yang dipersonalisasi, penghematan biaya, dan efisiensi energi.

ON-DEVICE GENERATIVE AI BENEFITS - ACCORDING TO QUALCOMM

Qualcomm, nama terkemuka dalam industri teknologi, telah banyak menyuarakan tentang berbagai keuntungan dari AI generatif pada perangkat. Salah satu manfaat utamanya adalah peningkatan privasi yang ditawarkannya. Dengan memproses data pada perangkat itu sendiri, risiko pelanggaran data dan akses yang tidak sah berkurang. Optimalisasi kinerja adalah keuntungan signifikan lainnya.



Sumber ilustrasi: Canva Elemen foto karya Md Rafiqul Islam

AI di perangkat memastikan pemrosesan waktu nyata, sehingga menghasilkan respons yang lebih cepat dan pengalaman pengguna yang lebih lancar. Selain itu, hal ini memungkinkan pengalaman yang dipersonalisasi yang disesuaikan dengan preferensi pengguna individu, sekaligus hemat biaya dan hemat energi.

MEDIATEK'S COLLABORATION WITH META FOR ON-DEVICE GENERATIVE AI

MEDIATEK, raksasa teknologi lainnya, telah menyadari potensi AI generatif pada perangkat dan telah berkolaborasi dengan Meta untuk mendorong batas-batas teknologi ini. Upaya bersama mereka bertujuan untuk menciptakan ekosistem komputasi *edge* yang komprehensif, mempercepat pengembangan aplikasi AI pada smartphone dan perangkat *edge* lainnya. Kemitraan ini menandakan sebuah langkah maju dalam membuat AI lebih mudah diakses dan efisien untuk perangkat sehari-hari.

QUALCOMM'S HYBRID APPROACH TO ON-DEVICE AND EDGE AI

Pendekatan Qualcomm terhadap AI adalah unik dan berfokus pada masa depan. Mereka percaya pada model *hybrid* yang menggabungkan kekuatan AI di perangkat dan AI *edge*. Pendekatan *full-stack* ini bertujuan untuk memanfaatkan potensi penuh AI, memastikan kinerja yang mulus di seluruh perangkat. Dengan mengintegrasikan pemrosesan di perangkat dengan komputasi *edge*, Qualcomm membayangkan masa depan di mana AI lebih responsif, efisien, dan berpusat pada pengguna.

“Dalam ranah AI, *edge* bukan hanya sebuah lokasi; melainkan sebuah janji. Janji akan privasi dalam interaksi kita, kesegeraan dalam respon kita, dan kemandirian dari batasan koneksi.”



Unlocking the Benefits of On-Device Generative AI

Discover the advantages of on-device generative AI, including privacy, performance, personalization, cost savings, and energy efficiency.

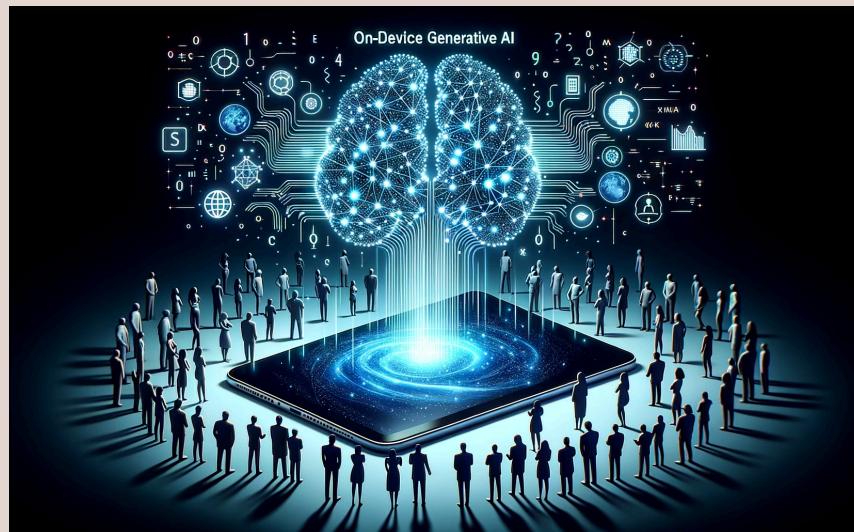
THE FUTURE OF AI: SCALING GENERATIVE AI WITH ON-DEVICE AI

AI generatif memiliki potensi yang sangat besar untuk masa depan, dan skalabilitasnya terkait erat dengan teknologi AI pada perangkat. Dengan memproses tugas AI pada perangkat, tantangan yang terkait dengan latensi, kinerja, dan privasi dapat diatasi secara efektif. Kemampuan pemrosesan di perangkat ini memastikan bahwa AI generatif dapat meningkatkan skalabilitas secara efisien, memenuhi berbagai macam aplikasi dan permintaan pengguna.

Evolusi pesat dari Artificial Intelligence (AI) telah membuka banyak sekali peluang di berbagai sektor. Salah satu kemajuan tersebut adalah Edge AI, yang mengacu pada pemrosesan algoritme AI pada perangkat keras lokal, lebih dekat dengan tempat data dihasilkan, daripada mengandalkan gudang pemrosesan data terpusat. On-device Generative AI, bagian dari Edge AI, semakin menonjolkan kemajuan ini dengan tidak hanya memproses tetapi juga menghasilkan data baru secara langsung pada perangkat lokal. Pergeseran paradigma ini membawa implikasi yang signifikan terhadap privasi, latensi, dan aksesibilitas offline, menjadikannya sebagai solusi ideal di dunia yang peka terhadap data.

PRIVACY

Salah satu keuntungan utama dari AI generatif pada perangkat adalah peningkatan privasi yang ditawarkannya. Ketika model AI berjalan langsung di perangkat, data pengguna tidak perlu dikirim ke server eksternal atau *cloud* untuk diproses. Ini berarti informasi sensitif tetap berada di perangkat, sehingga mengurangi risiko pelanggaran data atau akses yang tidak sah.



Di era saat ini, masalah privasi data menjadi sangat penting, pemrosesan di perangkat memberikan ketenangan bagi pengguna, karena mengetahui bahwa data mereka tidak dibagikan atau disimpan secara eksternal.

LOW LATENCY

Latensi mengacu pada penundaan antara tindakan pengguna dan respons sistem. Dengan AI generatif di perangkat, pemrosesan data terjadi secara real-time di perangkat itu sendiri, sehingga tidak perlu mengirim data bolak-balik ke server eksternal. Hal ini menghasilkan respons yang hampir seketika, sehingga meningkatkan pengalaman pengguna.

Untuk aplikasi yang membutuhkan umpan balik waktu nyata, seperti augmented reality, game, atau terjemahan waktu nyata, latensi yang rendah sangat penting. AI di perangkat memastikan bahwa aplikasi ini berjalan dengan lancar dan efisien, sehingga pengguna dapat berinteraksi tanpa hambatan.

OFFLINE ACCESSIBILITY

Keuntungan signifikan lainnya dari AI generatif pada perangkat adalah kemampuannya untuk berfungsi secara offline. Karena model AI disimpan dan dijalankan pada perangkat itu sendiri, maka model ini tidak bergantung pada koneksi internet yang aktif untuk berfungsi. Hal ini sangat bermanfaat di area dengan koneksi yang buruk atau ketika pengguna berada dalam mode pesawat. Baik itu untuk pengenalan suara, pemrosesan gambar, atau tugas berbasis AI lainnya, pemrosesan di perangkat memastikan bahwa aplikasi tetap berfungsi meskipun tanpa koneksi internet.

Hal ini tidak hanya meningkatkan pengalaman pengguna, tetapi juga memastikan bahwa fitur-fitur penting yang digerakkan oleh AI selalu dapat diakses, terlepas dari ketersediaan jaringan.



FREQUENTLY ASKED QUESTION



ON-DEVICE
GENERATIVE AI

How does on-device generative AI offer benefits in privacy, performance, personalization, cost, and energy?

On-device generative AI memproses data secara langsung di perangkat, memastikan privasi yang lebih baik dengan mengurangi pelanggaran data. AI ini menawarkan pemrosesan waktu nyata untuk kinerja yang dioptimalkan, memungkinkan pengalaman pengguna yang dipersonalisasi, serta hemat biaya dan hemat energi.

1. AI PRIVACY AND SECURITY

- AI di perangkat memberikan privasi dan keamanan yang lebih baik bagi pengguna.
- Transfer, penyimpanan, dan penggunaan data di berbagai platform dan layanan *cloud* dapat meningkatkan risiko pelacakan, manipulasi, dan pencurian data.
- Dengan AI di perangkat, pertanyaan dan informasi pribadi pengguna hanya ada di perangkat, sehingga memberikan perlindungan terhadap potensi pelanggaran data.
- Hal ini sangat penting untuk aplikasi yang sensitif seperti data medis, perusahaan, dan pemerintah.
- Misalnya, aplikasi asisten pemrograman yang menghasilkan kode dapat beroperasi pada perangkat tanpa mengekspos informasi rahasia ke *cloud*.



2. AI PERFORMANCE

- Performa AI mencakup performa pemrosesan dan latensi aplikasi.
- Performa pemrosesan pada perangkat seluler telah meningkat secara signifikan dengan setiap generasi teknologi.
- Hal ini memungkinkan penggunaan model AI generatif yang lebih besar dari waktu ke waktu, terutama karena model tersebut menjadi lebih optimal.
- Untuk AI generatif, latensi aplikasi sangat penting. AI di perangkat memastikan latensi rendah, kinerja tinggi, dan keandalan, menghindari potensi masalah latensi yang disebabkan oleh jaringan yang padat atau server *cloud*.



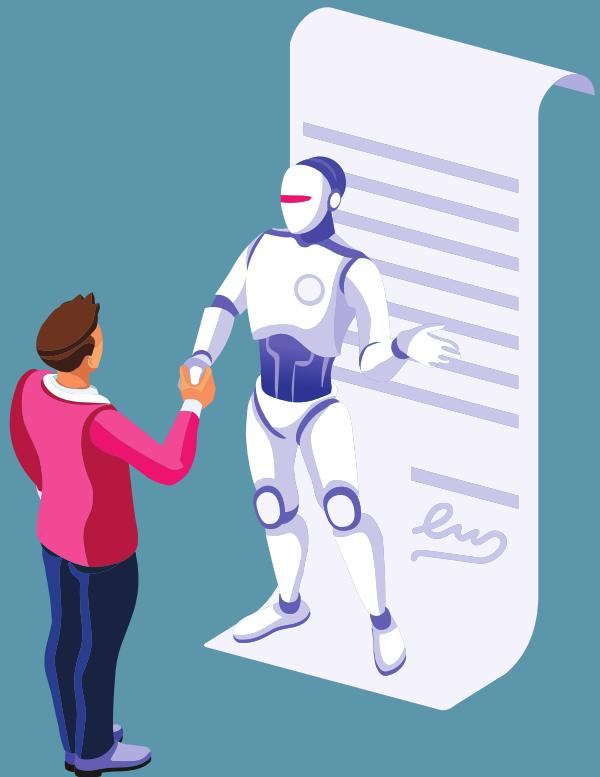
How does on-device generative AI offer benefits in privacy, performance, personalization, cost, and energy?

3. AI AND PERSONALIZATION

- AI generatif di perangkat menawarkan personalisasi yang lebih baik bagi pengguna.
- Hal ini memungkinkan penyesuaian model dan respons berdasarkan pola unik pengguna, lingkungan, dan data eksternal, sehingga menciptakan persona digital yang unik untuk setiap pengguna dari waktu ke waktu.
- Kemampuan ini juga dapat diterapkan pada kelompok atau organisasi untuk menciptakan respons yang kohesif.

4. COST OF AI

- Penyedia layanan *cloud* menghadapi peningkatan biaya yang terkait dengan menjalankan model AI generatif, yang mengarah ke biaya konsumen untuk layanan yang pada awalnya gratis.
- Menjalankan AI generatif pada perangkat dapat mengurangi biaya bagi konsumen dan penyedia layanan *cloud*, sehingga sumber daya dapat dialokasikan untuk tugas-tugas prioritas tinggi lainnya.



5. AI AND ENERGY

- Energi yang dibutuhkan untuk menjalankan model AI generatif pada perangkat jauh lebih sedikit daripada di *cloud*.
- Pemrosesan AI di perangkat lebih hemat energi, terutama ketika mempertimbangkan transportasi data.
- Perbedaan dalam konsumsi energi ini membantu penyedia layanan *cloud* memenuhi tujuan lingkungan dan keberlanjutan mereka.

What is the role of edge devices in enhancing on-device generative AI?



Perangkat edge memainkan peran penting dalam AI generatif di perangkat dengan menyediakan dukungan perangkat keras dan perangkat lunak yang diperlukan. Perangkat ini memungkinkan pemrosesan waktu nyata, memastikan respons yang lebih cepat dan pengalaman pengguna yang mulus.

MediaTek, perusahaan semikonduktor global, memanfaatkan perangkat *edge* untuk meningkatkan AI generatif di perangkat. Mereka berkolaborasi dengan Meta's Llama 2 untuk menjalankan aplikasi AI langsung di perangkat, bukan melalui *cloud*. Keuntungan dari pendekatan ini meliputi:

- Performa yang lebih cepat
- Privasi yang lebih baik
- Keamanan yang ditingkatkan
- Latensi yang berkurang
- Biaya operasional yang lebih rendah

Dengan chipset terbaru dari MediaTek, mereka berharap dapat mempercepat pengembangan dan inovasi dalam AI generatif di perangkat.

MediaTek, melalui inovasi dan kolaborasi, berupaya memimpin transformasi ini, membawa AI generatif lebih dekat ke kehidupan sehari-hari kita. Dengan demikian, peran perangkat *edge* tidak hanya sebagai alat komunikasi atau hiburan, tetapi juga sebagai pusat inovasi dan pemrosesan AI yang canggih.

What are the advancements in on-device and edge AI?

Qualcomm dan MediaTek telah membuat langkah yang signifikan dalam AI pada perangkat dan edge. Pendekatan hibrida Qualcomm menggabungkan AI pada perangkat dan edge untuk kinerja optimal, sementara kolaborasi MediaTek dengan Meta bertujuan untuk mempercepat pengembangan aplikasi AI pada perangkat edge.

Open Access | Review

Advancements in On-Device Deep Neural Networks

by  Kavya Saravanan 1,2 and  Abbas Z. Kouzani 1,* 

Dalam dunia kecerdasan buatan yang berkembang pesat, pergeseran ke arah AI pada perangkat dan edge telah menandai langkah signifikan dalam kemajuan teknologi. Transisi ini tidak hanya menjanjikan pengalaman pengguna yang lebih baik, tetapi juga mengatasi masalah yang berkaitan dengan latensi, privasi, dan konektivitas. Beberapa kemajuan yang dapat disimpulkan dari artikel komprehensif dari MDPI.

- **Arsitektur ResNet:** Menggunakan koneksi residual untuk memudahkan optimasi bobot dan menghindari masalah *vanishing gradient*.
- **Kerangka NNStreamer:** Memberikan pipa efisien untuk pengembangan AI di perangkat.
- **Arsitektur Hierarki untuk Video Streams:** Menggunakan perangkat *edge* untuk pemrosesan awal dan server *cloud* untuk pemrosesan yang lebih kompleks.
- **Visual-Inertial Odometry (VIO):** Metode navigasi yang menggabungkan data visual dan inersia untuk navigasi drone *real-time*.
- **Arsitektur Perangkat Keras untuk Accelerators:** Menggunakan elemen pemrosesan yang terhubung melalui jaringan di chip.





How does on-device AI enable the scaling of generative AI?

Kecerdasan buatan (AI) generatif mengalami adopsi yang cepat dan seiring dengan itu, terjadi peningkatan permintaan komputasi. Untuk mengatasi hal ini, arsitektur AI hibrida muncul sebagai solusi. Arsitektur ini mendistribusikan dan mengoordinasikan beban kerja AI antara cloud dan perangkat edge, seperti ponsel pintar, mobil, PC, dan perangkat IoT. Motivasi utama di balik pendekatan ini adalah penghematan biaya. Sebagai contoh, biaya pencarian berbasis AI generatif diperkirakan 10 kali lebih tinggi daripada metode tradisional. Dengan memanfaatkan kemampuan komputasi perangkat edge, AI hibrida dapat mengurangi biaya ini. Arsitektur ini tidak hanya menawarkan penghematan biaya tetapi juga meningkatkan kinerja, personalisasi, privasi, dan keamanan dalam skala global.

Berdasarkan pada kompleksitas model dan kueri, pemrosesan dapat didistribusikan antara *cloud* dan perangkat atau dijalankan sepenuhnya di perangkat. Potensi AI hibrida semakin meningkat seiring dengan semakin kecilnya model AI generatif yang kuat dan meningkatnya kemampuan pemrosesan di perangkat. Pendekatan ini sangat penting bagi AI generatif untuk meningkatkan skala dan memenuhi kebutuhan global, menekankan bahwa masa depan AI memang bersifat hibrida.



Tren Research and Conclusion

Penelitian tren terbaru menggarisbawahi momentum pertumbuhan Edge AI. Menurut laporan dari Grand View Research, ukuran pasar perangkat lunak edge AI global diperkirakan akan tumbuh dengan laju pertumbuhan tahunan gabungan (CAGR) sebesar 21,0% dari tahun 2023 hingga 2030. Pertumbuhan ini didorong oleh meningkatnya masalah privasi dan meningkatnya adopsi perangkat IoT.

Selain itu, sebuah publikasi dari Deloitte Insights menyoroti aplikasi dunia nyata dari AI generatif di perangkat, yang menekankan dampak transformatifnya pada bisnis dan teknologi pribadi. Laporan ini membahas bagaimana perusahaan memanfaatkan teknologi ini untuk meningkatkan pengalaman pelanggan, merampingkan operasi, dan menciptakan produk dan layanan baru.

Kesimpulannya, Edge AI, terutama AI generatif pada perangkat, bukan hanya tren sesaat, melainkan sebuah kebutuhan yang lahir dari kebutuhan mendesak akan privasi, pemrosesan waktu nyata, dan akses tanpa gangguan. Adopsi di seluruh industri menggarisbawahi potensinya untuk mengatasi beberapa tantangan paling mendesak yang dihadapi oleh sistem AI berbasis cloud tradisional. Seiring dengan langkah kita ke depan, hal ini akan berperan penting dalam membentuk masa depan di mana AI dan privasi hidup berdampingan dengan mulus, mendorong dunia yang lebih aman dan efisien.

REFERENCES

Deloitte. (2023). Deloitte Launches Generative AI Practice to Help Clients Harness the Power of Disruptive New AI Technology. Diakses pada 13 Oktober 2023, dari <https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-launches-generative-ai-practice-to-help-clients-harness-the-power-of-disruptive-new-ai-technology.html>

Forbes. (2023). Qualcomm is Right About AI: It Requires Strong Edge Computing. Diakses pada 12 Oktober 2023, dari <https://www.forbes.com/sites/patrickmoorhead/2023/06/27/qualcomm-is-right-about-ai-it-requires-strong-edge-computing/?sh=6d538971c95c>

Grand View Research. Edge AI Market Size, Share & Trends Analysis Report By Component (Hardware, Software, Edge Cloud Infrastructure, Services), By End-use Industry, By Region, And Segment Forecasts, 2023 - 2030. Diakses pada 13 Oktober 2023, dari <https://www.grandviewresearch.com/industry-analysis/edge-ai-market-report>

MediaTek. (2023). MediaTek Leverages Meta's Llama 2 to Enhance On-Device Generative AI in Devices. Diakses pada 13 Oktober 2023, dari <https://corp MEDIATEK.com/news-events/press-releases MEDIATEK-leverages-metas-llama-2-to-enhance-on-device-generative-ai-in-edge-devices>

NVIDIA. (2022). What is Edge AI and How Does It Work? Diakses pada 29 September 2023, dari <https://blogs.nvidia.com/blog/2022/02/17/what-is-edge-ai/>

Qualcomm. (2023). 5 benefits of on-device generative AI. Diakses pada 10 Oktober 2023, dari <https://www.qualcomm.com/news/onq/2023/08/5-benefits-of-on-device-generative-ai>

Qualcomm. (2023). How on-device AI is enabling generative AI to scale [The future of AI is hybrid]. Diakses pada 10 Oktober 2023, dari <https://www.qualcomm.com/news/onq/2023/05/how-on-device-ai-is-enabling-generative-ai-to-scale>

REFERENCES

Saravanan, K., & Kouzani, A. Z. (2023). Advancements in On-Device Deep Neural Networks. Information, 14(8), 470. <https://doi.org/10.3390/info14080470>