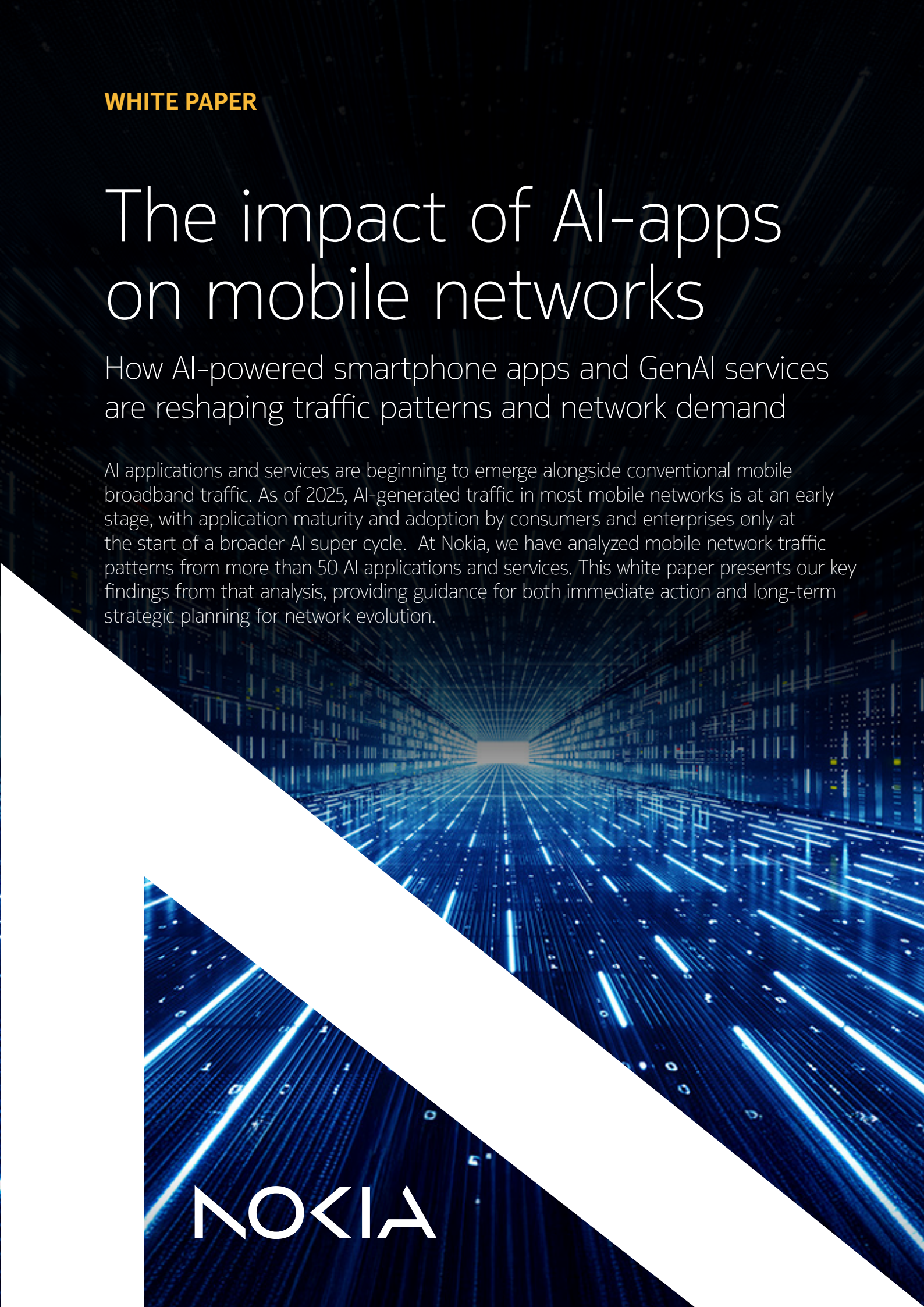
WHITE PAPER

The impact of AI-apps on mobile networks

How AI-powered smartphone apps and GenAI services are reshaping traffic patterns and network demand

AI applications and services are beginning to emerge alongside conventional mobile broadband traffic. As of 2025, AI-generated traffic in most mobile networks is at an early stage, with application maturity and adoption by consumers and enterprises only at the start of a broader AI super cycle. At Nokia, we have analyzed mobile network traffic patterns from more than 50 AI applications and services. This white paper presents our key findings from that analysis, providing guidance for both immediate action and long-term strategic planning for network evolution.



Summary

The telecommunications industry stands on the brink of a profound shift. Generative AI (GenAI) applications are rapidly moving from novelty to necessity, redefining how users interact with digital services. While much of the conversation around AI has centered on its capabilities, less attention has been paid to how this traffic behaves - and more importantly, how mobile networks must evolve to support it.

At Nokia, we have conducted an extensive analysis of GenAI applications and services to understand how they are changing requirements on mobile networks. In short there are four key findings that every mobile network operator should know:

- **Shifting UL/DL ratios**: Uplink data is growing faster than downlink traffic, driven in part by conditioning inputs such as images transmitted to AI inference factories.
- **Rising data volumes**: Multi-modal, user-friendly experiences are increasing overall traffic. People are starting to “talk with their data”, interact with AI assistants, and share photos and videos from their smartphones to refine prompts.
- **Sensitivity to latency**: Conversational AI services respond non-linearly to extended latency. Today, this can disrupt AI voice conversations. In the future, immersive applications will be unusable without consistently low latency.
- **Agentic AI opportunities**: AI agents (software agents that perform tasks on a user’s behalf) can shift inference loads and related network traffic away from peak hours, which are defined by the level of human activity. By operating in scheduled, off-peak cycles, they ensure results are ready when needed while avoiding congestion.

As AI apps and services approach perfection and become more valuable, user expectations (both consumer and professional) will evolve from best-effort mobile broadband to premium quality, high-speed, low-latency connectivity. The good news: Advanced 5G and upcoming 6G networks are designed to deliver exactly that.

AI-generated traffic is still in its infancy, coexisting with conventional mobile broadband usage. As adoption accelerates across consumer and enterprise segments, this shift marks the early stage of a much broader AI super cycle.

Conducting our analysis

To understand the impact of AI applications on mobile networks, we performed detailed throughput, data volume and latency measurements on commercial iOS and Android smartphones, operating in over-the-air radio cells.

We analyzed more than 50 popular AI apps and services. These included the dominant models and services from leading AI providers, as well as AI-powered apps of smaller companies that have gained strong traction among (typically younger) users.

Our testers engaged in sessions designed to reflect typical consumer or professional smartphone behavior. The median session lasted about 7.5 minutes.

In total, the analysis covered more than 20 use cases, which we grouped into four main categories:

- **Chat and conversation**: Text-based chats with general-purpose chatbots (e.g. ChatGPT) and voice conversations with AI services. Also included task-specific use cases, such as scene recognition or solving handwritten math problems. Some of these use cases take images as conditioning input, increasing the data volumes and data rates in uplink.
- **Document generation**: Creation of longer texts and formatted documents like PDFs or presentations. Prompts included text and voice to input documents and images.
- **Image generation**: Creation of images from scratch based on prompts, and AI-powered image manipulation, reflecting the growing entertainment-driven adoption among younger demographics. These applications, while fun, have a heavy traffic impact on the network.
- **Video generation**: AI-based video creation. Throughput-intensive in the downlink, while image inputs also drive relatively high uplink volumes.

We also extended our tests to agentic AI use cases.

To ensure accurate results, we removed bandwidth constraints in the test network. This allowed us to measure how quickly AI endpoints could send and receive data without the radio cell acting as a bottleneck. For latency, we artificially introduced delays to test how tolerant apps and servers are to network lag.

All measurements were performed in the Helsinki area, Finland, and we observed that AI apps and services were also communicating with endpoints in North America, Europe and Asia.

We ensured that bandwidth was not a limiting factor, so throughput and latency measurements reflected only endpoint performance and network path efficiency, not radio constraints.

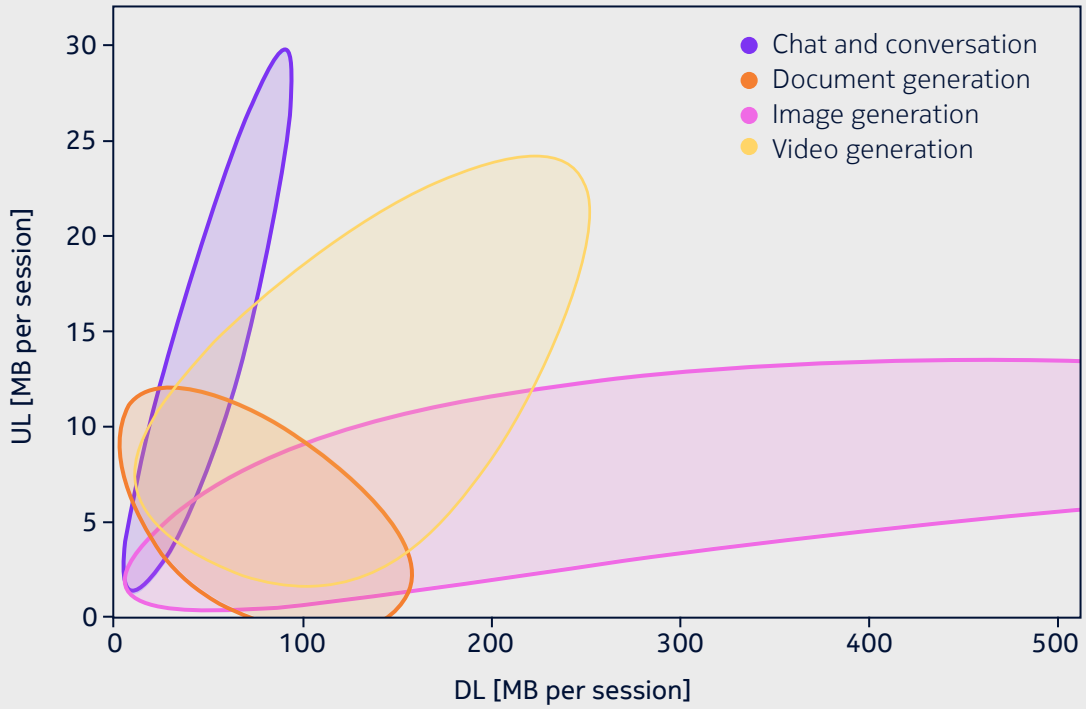
Results of the analysis

Data volumes

AI app data volumes increased with modality from text to voice, image and video. Unlike conventional broadband traffic, each app used different amounts of conditioning inputs and outputs, resulting in a wide range of uplink and downlink data volumes per session.

This variance in data usage reflects the multi-modal nature of many AI applications, which rely on combinations of text, image and audio inputs to refine responses.

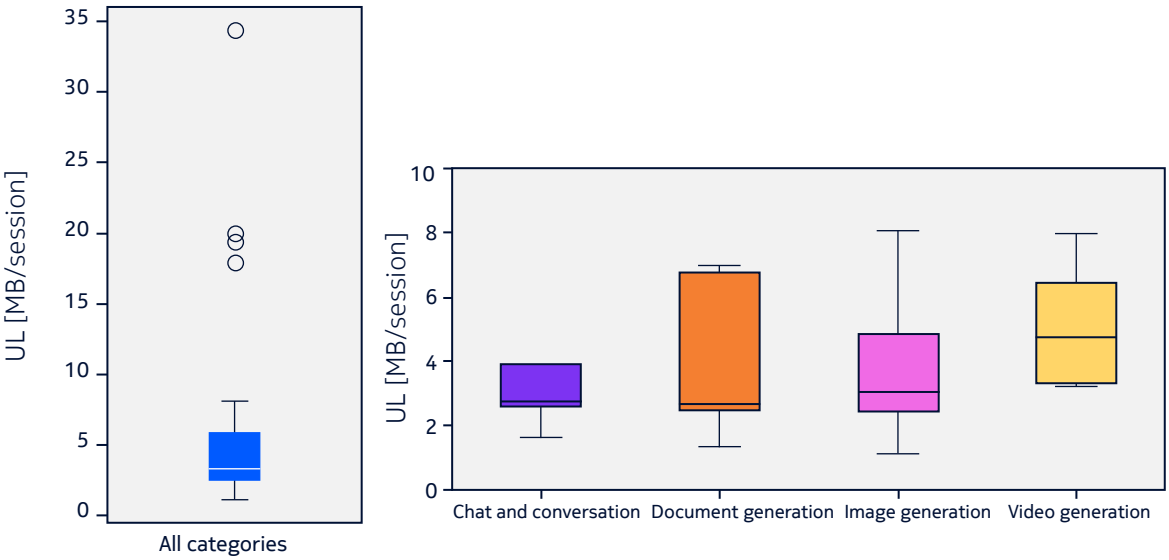
Downlink and uplink data volumes per AI-app session



Uplink

Uplink data volumes were surprisingly high – even for text chats and voice conversations - as many apps used photos and videos from the smartphone as conditioning inputs. This reflects the multi-modal nature of many apps. On-device AI apps also synced with the cloud, further driving uplink traffic. Without adequate capacity planning, congestion could limit AI app performance and lead to suboptimal uplink data rates as a consequence of congestion.

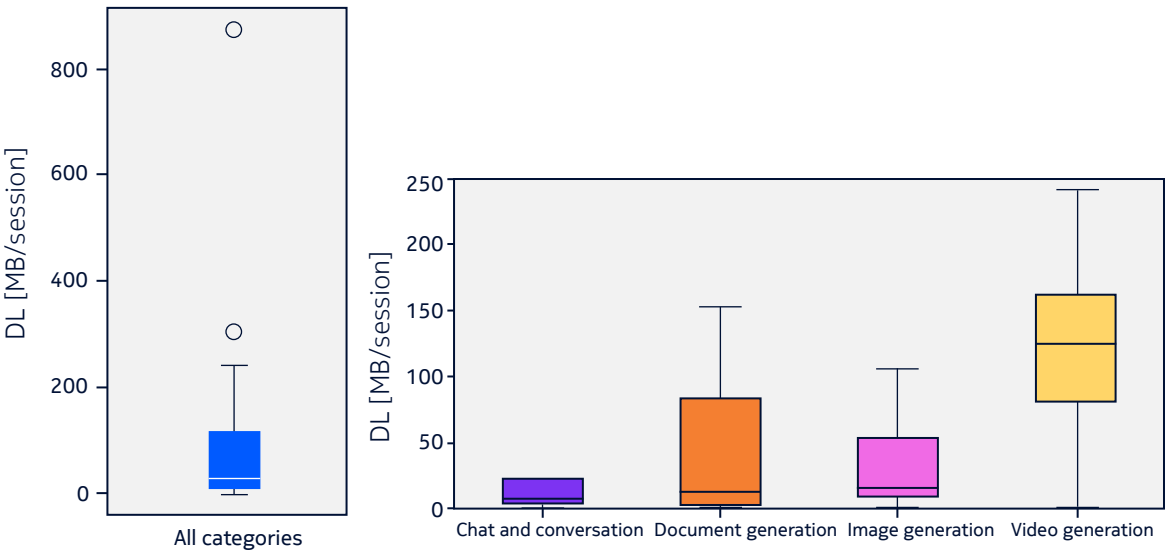
When comparing the data volumes for sharing screen content as live stream with GenAI services, we noted up to 9x higher uplink data volumes compared to sharing the screen with established services for human interaction like Facetime or Zoom.



Downlink

Some apps went beyond a one-shot answer, repeatedly consulting sources and updating responses. With each consultation, they increased the downlink data volume beyond what was required for the final answer. While some users may enjoy this iterative process, it inevitably generates additional downlink traffic.

Users might appreciate seeing that the app is “working hard” on their behalf, but for the network this iterative experience translates into additional, sometimes redundant traffic.



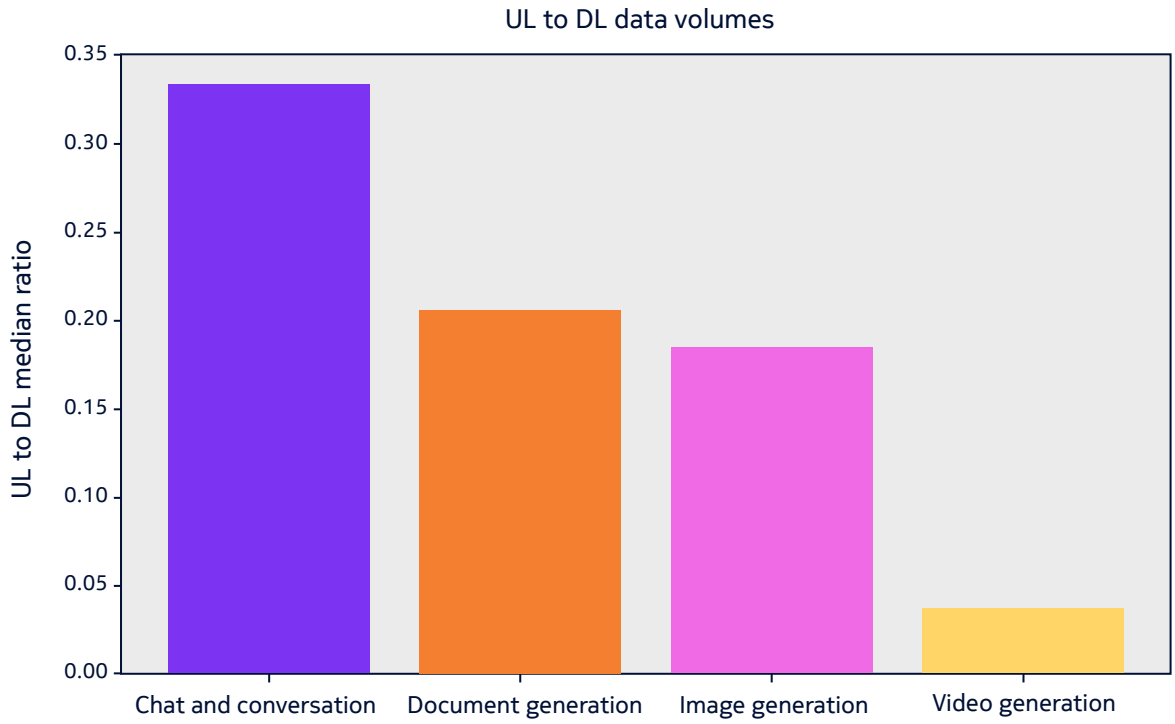
Observed distribution

Session data volume distributions showed that while some apps consumed many times the median, most fell within comparable ranges, reflecting a certain maturity. However, interquartile ranges of uplink and downlink data volumes also indicate a developing equilibrium among app categories, suggesting maturing usage patterns.

The spread of the data volumes per app session strongly evidences that overall network traffic prediction becomes more challenging with AI traffic stepping alongside conventional human-generated traffic.

Downlink to uplink ratio per category

The most uplink-centric AI apps analyzed sent more than 30 times more data than they received, especially when handling photos or videos.



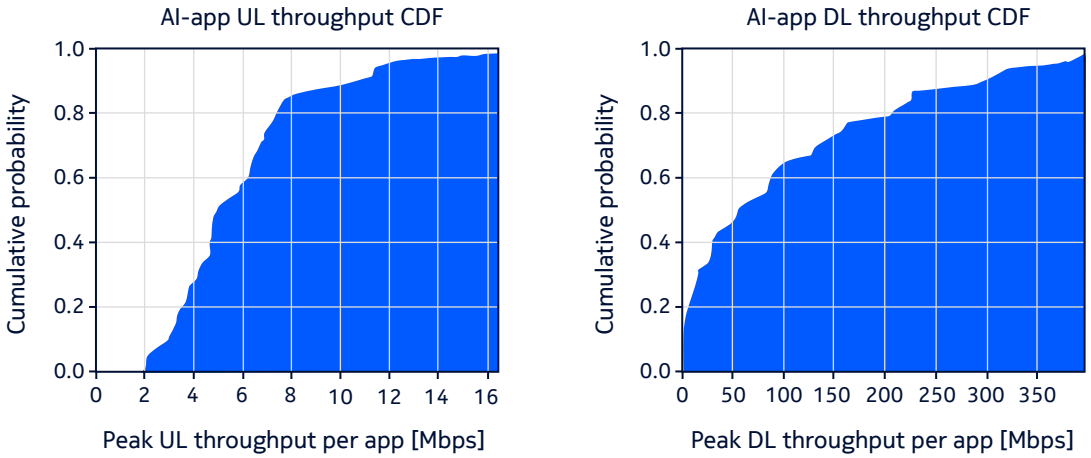
The ratio of downlink and uplink median data volumes is in the range of only 3:1 to 5:1 for chat, conversation, document and image generation. Video generation showed a much higher ratio, as longer videos are not yet widely used as conditioning inputs. As of 2025, many 5G networks show downlink to uplink data volumes of 12:1 or above.

As expected, AI apps running directly on the smartphone reduced immediate data exchange with the cloud. However, cloud-based synchronization or data backups of documents, images or videos increased overall data volumes. In the future, as AI-empowered users become more productive, synchronization traffic, especially from smartphone or smart glasses data, will accelerate in uplink more than in downlink.

A broad adoption of AI-apps will shift the ratio towards a relatively higher uplink capacity need.

Throughput

In our AI app measurements with smartphones on a high-speed 5G network under low cell load, throughput was not constrained by the RAN. Instead, it depended on the capabilities of the AI endpoints and IP routes connecting to them. Endpoints located in Finland or elsewhere in Europe delivered the highest throughput.



In contrast to the observations of overall data volumes, the downlink to uplink ratio of measured peak data rates remained closer to traditional broadband patterns, underscoring that throughput limitations often stem from endpoint performance rather than radio interface constraints. For most apps analyzed, the observed data rates were within levels that can be efficiently supported by Advanced 5G features, such as slicing, to ensure excellent user experience. Observed throughput levels align with Advanced 5G capacity envelopes, meaning existing RAN designs can accommodate AI workloads provided sufficient backhaul and IP path optimization.

Latency

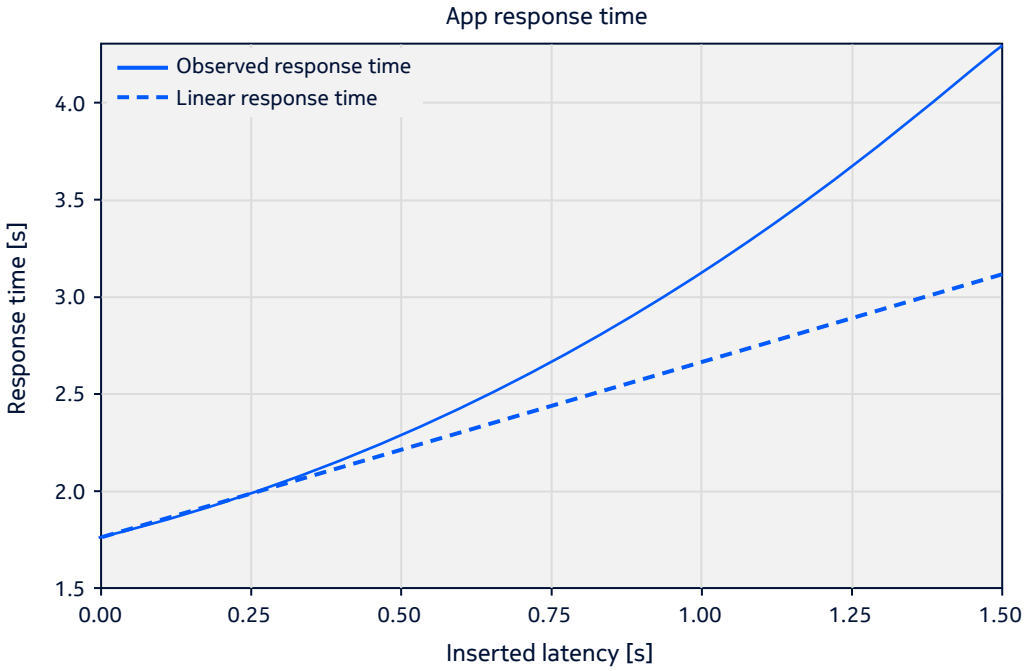
For AI apps with models running in the Cloud, user experience is determined by the combined effect of network latency and computation time. In Advanced 5G networks, RAN latency is typically much smaller than the computation time of a generative AI model. While some apps responded almost instantly, others took several seconds. One app required nearly 20 seconds to respond, underscoring significant variation in responsiveness.

Generative AI models are improving rapidly, seeking the right balance between response quality and response speed. Improvements target both time-to-first token and steady token throughput. In robotic AI use cases, generative models already deliver sub-100 ms responses, enabling end-to-end control cycles at 10 Hz.

Some applications begin returning partial outputs while inference continues, masking latency from users but not reducing network load.

As conversational AI evolves toward fully multimodal immersive services, end-to-end latency requirements will shrink further. Every millisecond saved at the RAN level could reduce computing costs or enable the use of more advanced models.

In our analysis we artificially introduced network latency to observe app behavior.



In one measurement series, the AI-app responded linearly for inserted latency times up to 0.5 s. Then the non-linear response starts. At around 1.5 s inserted latency, the end-to-end response time of the AI-app has grown by almost twice the inserted latency.

This non-linear sensitivity poses a serious risk to user experience in conversational AI apps and requires close monitoring as immersive AI services are introduced.

Agentic AI

Our measurements included several agentic AI apps. The agentic tasks spanned multiple steps, ranging from data search and analysis to the generation of documents in defined formats. One example was the generation of PDF-format travel plans, including flights, accommodation, meeting schedules and budget limitations.

The AI agents typically operated for 10 to 20 minutes, reflecting a relatively large amount of AI inference work, yet the data volumes were roughly in line with those of the other AI applications analyzed. The resulting outputs were often more data -rich, though this was partly offset by interim step results not being sent to the smartphone.

Scheduling agentic AI tasks can also help shift mobile network data traffic and AI inference workloads away from peak hours of human activity.

Results summary

Overall, the analysis showed that AI applications are reshaping mobile traffic patterns, with uplink demands rising sharply, throughput staying within manageable ranges and latency becoming a decisive factor for user experience. These arising traffic patterns have the potential to challenge traditional network planning assumptions, making Advanced 5G (and eventually 6G) capabilities essential to sustain both performance and user expectations.

Looking forwards

As artificial intelligence continues to evolve, AI apps will shape how we interact with data, both in the cloud and in our physical surroundings. These apps and services are developing rapidly, though adoption is still in its early stages.

The impact of AI apps on any given mobile networks will depend heavily on local factors such as demographics and attitudes towards new technologies .

In the near to mid-term, uplink network capacity might be the area in which network requirements can change fastest. Broader adoption of today's AI apps by smartphone users could drive uplink traffic growth by more than 50%. On top of that, new waves of AI applications and services will further shift traffic patterns, especially in metropolitan areas where mobile broadband networks are already the most loaded.

Media-enriched, conversational AI apps are already available 24/7 and may soon supplement human-to-human calls or complement other media consumption. Immersive AI apps and services will add additional uplink and downlink traffic demands and will only succeed when end-to-end latency remains within tight delay budgets.

For users, the AI experience on their smartphones will soon become as important as mobile broadband itself. To elevate this experience beyond best-effort transmission, 5G slicing provides an ideal solution.

For operators, this shift also signals new business opportunities, from monetizing premium AI experiences with network slicing to capturing value through differentiated low-latency and uplink capacity services.

While this white paper has focused on human users, robotic AI will bring further network requirements. Affordable robots will depend on foundational AI models running in the cloud, as these cannot be supported by the robot's processor or battery alone.

The AI super cycle has only just begun. Over time, consumers, enterprises and society will come to view AI applications not as novelties, but as necessities.

Meeting these data exchange requirements will be both a key network differentiator and a major business opportunity for network operators, much as mobile broadband was more than a decade ago.

Glossary

AI	Artificial Intelligence
GenAI	Generative Artificial Intelligence
UL	Uplink
DL	Downlink
PDF	Portable Document Format
RAN	Radio Access Network
IP	Internet Protocol
Hz	Hertz

Nokia OYJ
Karakaari 7
02610 Espoo
Finland

Tel. +358 (0) 10 44 88 000

CID: 215147

nokia.com

About Nokia

Nokia is a global leader in connectivity for the AI era. With expertise across fixed, mobile, and transport networks, we're advancing connectivity to secure a brighter world.