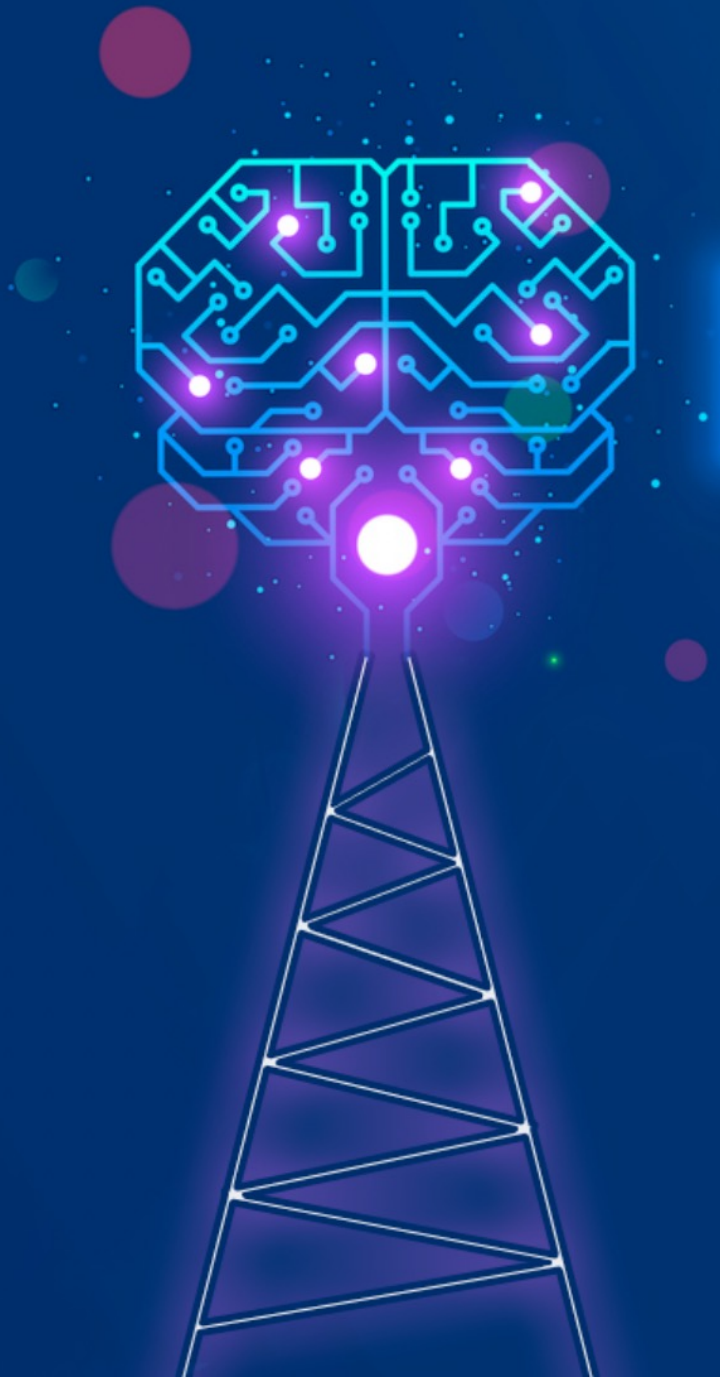


Generative AI in Telecom Systems and Business

2024 Networked AI Systems course University of Helsinki

Jari Arkko

Ericsson Research

Acknowledgments to Jaime Jiménez, Athanasios Karapantelakis, and Vincent Huang, all at Ericsson

Agenda

- Introduction
- What is Ericsson & Telecom
- Generative AI at Ericsson
- Generative AI in Telecom
- Example technical topics
- Evolving Generative AI
- Summary

Generative AI

Artificial intelligence that can generate new content: text, images, or other media.

Language Models (LMs)

A probabilistic model of a natural language that can generate probabilities of a series of words.

Large Language Models (LLMs)

An LM able to achieve general-purpose language understanding and generation.

Transformers, Attention

Technologies that made current LLMs feasible.

Today's Question

Generative AI has had and is predicted to have a big impact on many fields – starting from writing, art, search, etc.

What is the impact in *our* fields?

Specifically, is Generative AI useful to the telecom Industry?

Dall-e 2: "An astronaut walking in Stockholm"

Today's Question

Generative AI has had and is predicted to have a big impact on many fields – starting from writing, art, search, etc.

What is the impact in *our* fields?

Specifically, is Generative AI useful to the telecom Industry?

… and how, where, when?

Dall-e 2: "An astronaut walking in Stockholm"

Today's Question

Generative AI has had and is predicted to have a big impact on many fields – starting from writing, art, search, etc.

What is the impact in *our* fields?

Specifically, is Generative AI useful to the telecom Industry?

… and how, where, when?

… any limitations?

Dall-e 2: "An astronaut walking in Stockholm"

Today's Question

Generative AI has had and is predicted to have a big impact on many fields – starting from writing, art, search, etc.

What is the impact in *our* fields?

Specifically, is Generative AI useful to the telecom Industry?

… and how, where, when?

… any limitations?

… the impact on our companies?

Dall-e 2: "An astronaut walking in Stockholm"

Agenda

- Introduction
- What is Ericsson & Telecom
- Generative AI at Ericsson
- Generative AI in Telecom
- Example technical topics
- Evolving Generative AI
- Summary

What is Ericsson?

Ericsson key numbers

101,000
employees worldwide

29,000
R&D employees worldwide

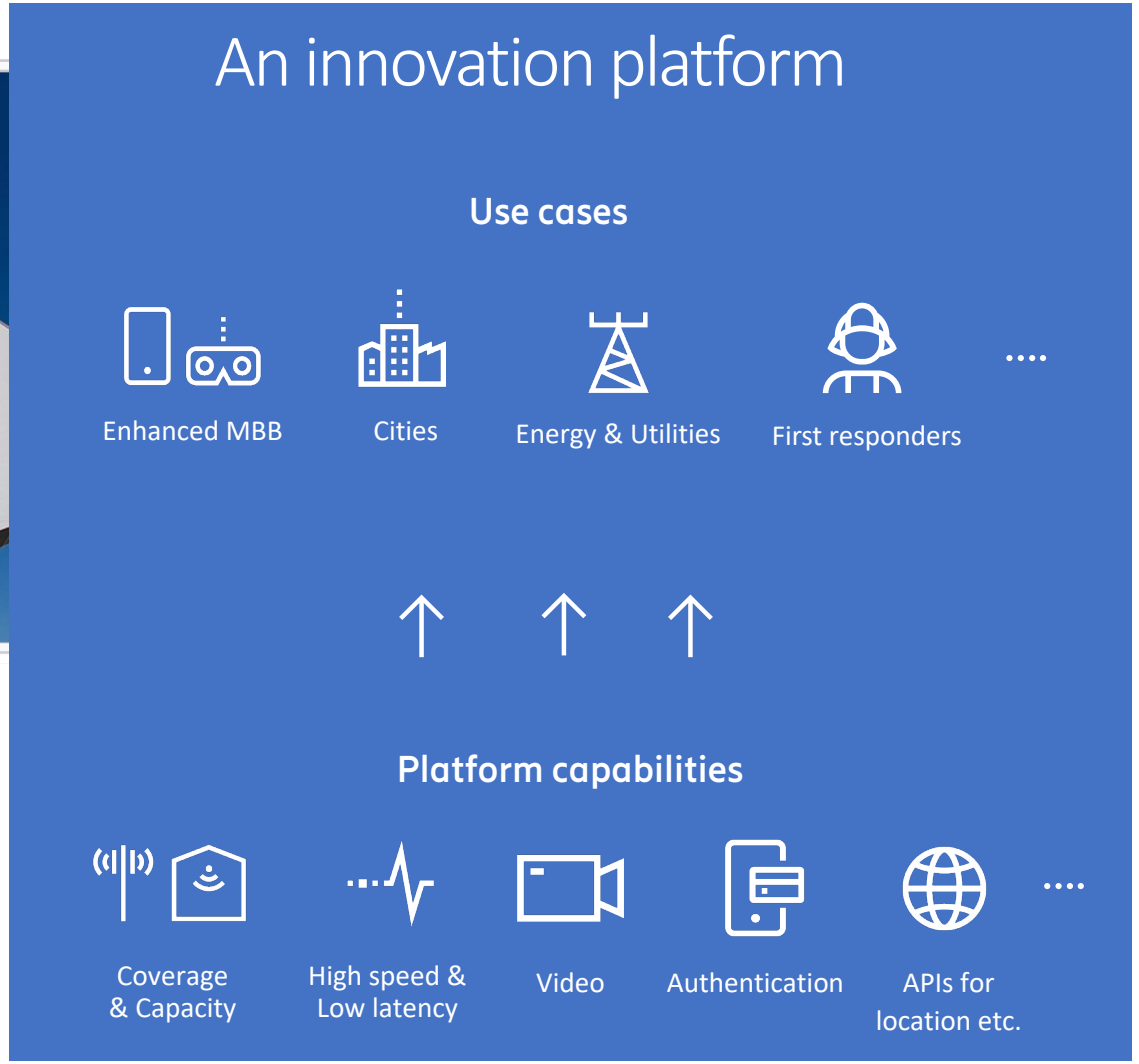
Ericsson key offerings

- 5G Core
- 5G Radio Access Network
- 5G Transport
- Business and Operations Support Systems
- Cloud Communications and Network APIs
- Managed Services
- Mission Critical Communications
- Network Automation and AI
- Network Services
- Private Networks
- Wireless WAN

What is telecommunications?

Telecommunication, often used in its plural form, is the transmission of information by various types of technologies over wire, radio, optical, or other electromagnetic systems.

- Transport of bits but also APIs, control, authentication, management, resilience, ….
- Technologies and standards
- Products, services, development, operations
- Business, marketing, customer relations
- Deeply connected ecosystems

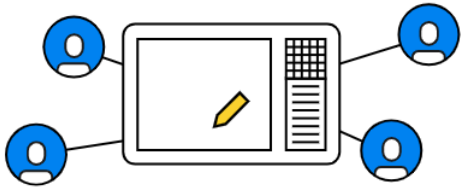


The forces shaping the future

Society and geopolitics

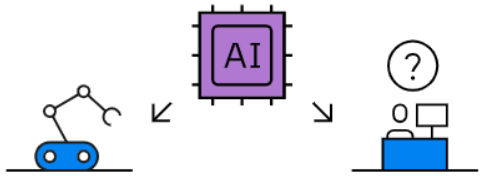
The climate crisis

Moves from 'a challenge' to 'the challenge' that will dominate the next decade



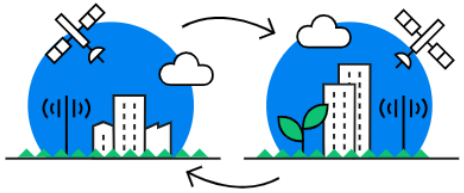
Post pandemic society

De-urbanization; increasing virtual interactions and collapsed cycle of innovation



Socio-economic shifts

AI & automation employment disruption and persistent economic uncertainty



Global fragmentation

Higher emphasis on national digital sovereignty – threatening global collaboration

Redefining trust

Trust in the digital world is challenged. The landscape of who and what we trust is changing

Implications for the ICT industry

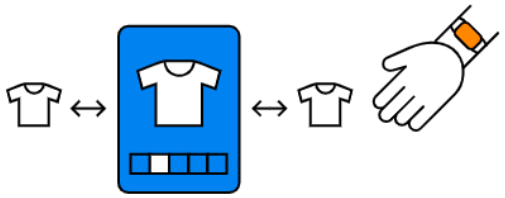
- More will be demanded to address sustainability. Net zero becomes a table-stake – companies will be held accountable to the difference their technologies *could* make
- The business environment will be turbulent – trust issues challenge take-up; fragmentation creates silos; an uncertain economy clouds investment decisions
- But – changes in societal dynamics & a strong desire for continual innovation will drive new opportunity development

The forces shaping the future

Technology

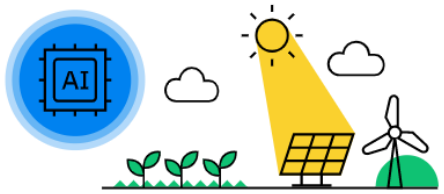
Connectivity in everything

Trillions of sensors in everything – from devices, to smart materials, to nature



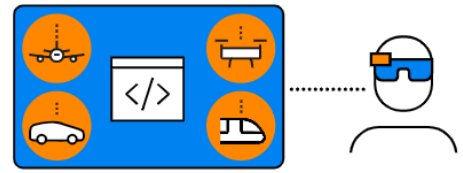
Beyond smartphones

From XR wearables, to devices offering a full spectrum of senses, to the first 'bio-interfaces'



Artificial gets real

AI supports all aspects of life, and becomes a primary competitive differentiator



The mirror world

Sophisticated digital and spatial mapping creating a virtual replica of the 'real world'

Implications for the ICT industry

- The arrival of 'connectivity in everything' will necessitate the need for new networks of 'limitless connectivity' – offering complete coverage, vastly increased speed & capacity, and near zero latency
- Networks will need to be continually optimized for AI & processing in the core, the edge, and everywhere in between
- Smartphones will no-longer be the dominant form of consumer device connectivity – we'll see an explosion of new consumer device formats which will consolidate towards the end of the decade
- New applications of the 'mirror-world' will drive fast paced evolution of automation & digital twin technologies

New network attributes

Extensive and dynamic

- From population to full geographic and 3D coverage
- Supporting trillions of sensors
- Managing extreme and dynamic performance
- …at reduced energy consumption

Highly intelligent

- Operated at scale without human intervention
- Centralized and decentralized AI for service availability & energy use
- AI-native design: Learning, cognitive design, and data-oriented
- …intent driven and explainable

Capable of processing anywhere

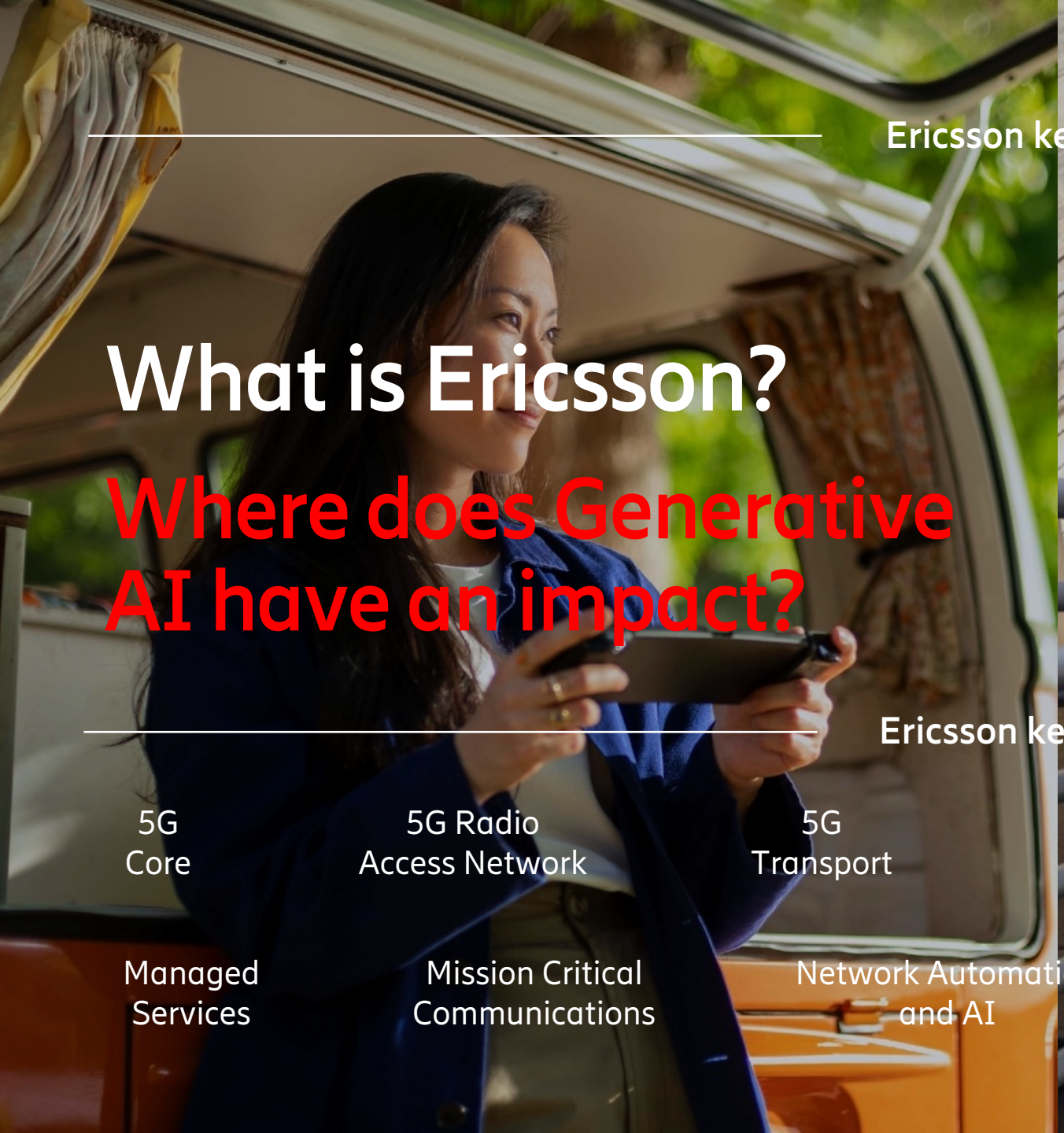
- Complex & real-time processing distributed & tightly integrated through-out the network
- Multi-cloud philosophy & continual evolution of processor technology (towards quantum)

Resilient and trustworthy

- Increased focus on capabilities inherent in networks today – reliability, availability, resilience, security & privacy
- Extensions into digital twins and AI for risk modelling and predictive action

Agenda

- Introduction
- What is Ericsson & Telecom
- Generative AI at Ericsson
- Generative AI in Telecom
- Example technical topics
- Evolving Generative AI
- Summary



What is Ericsson?

Where does Generative AI have an impact?

Ericsson key numbers

101,000
employees worldwide

29,000
R&D employees worldwide

Ericsson key offerings

- 5G Core
- 5G Radio Access Network
- 5G Transport
- Business and Operations Support Systems
- Cloud Communications and Network APIs
- Managed Services
- Mission Critical Communications
- Network Automation and AI
- Network Services
- Private Networks
- Wireless WAN

What is Ericsson?

Where does Generative AI have an impact?

Ericsson key numbers

101,000
employees worldwide

29,000
R&D employees worldwide

Ericsson key offerings

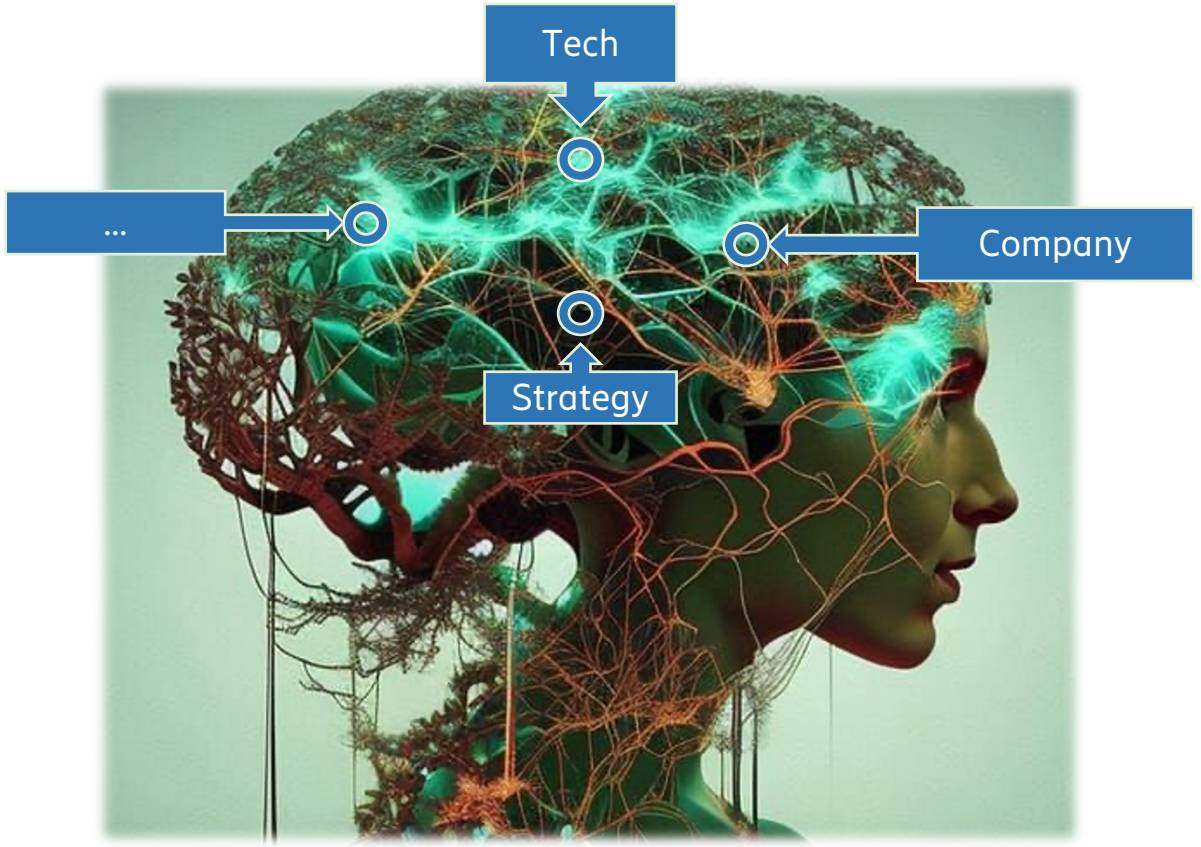
- 5G Core
- 5G Radio Access Network
- 5G Transport
- Business and Operations Support Systems
- Cloud Communications and Network APIs
- Managed Services
- Mission Critical Communications
- Network Automation and AI
- Network Services
- Private Networks
- Wireless WAN

New Technology Introduction

Before looking at the tech changes, let's zoom out a bit:

What does one do in a company with new tech?

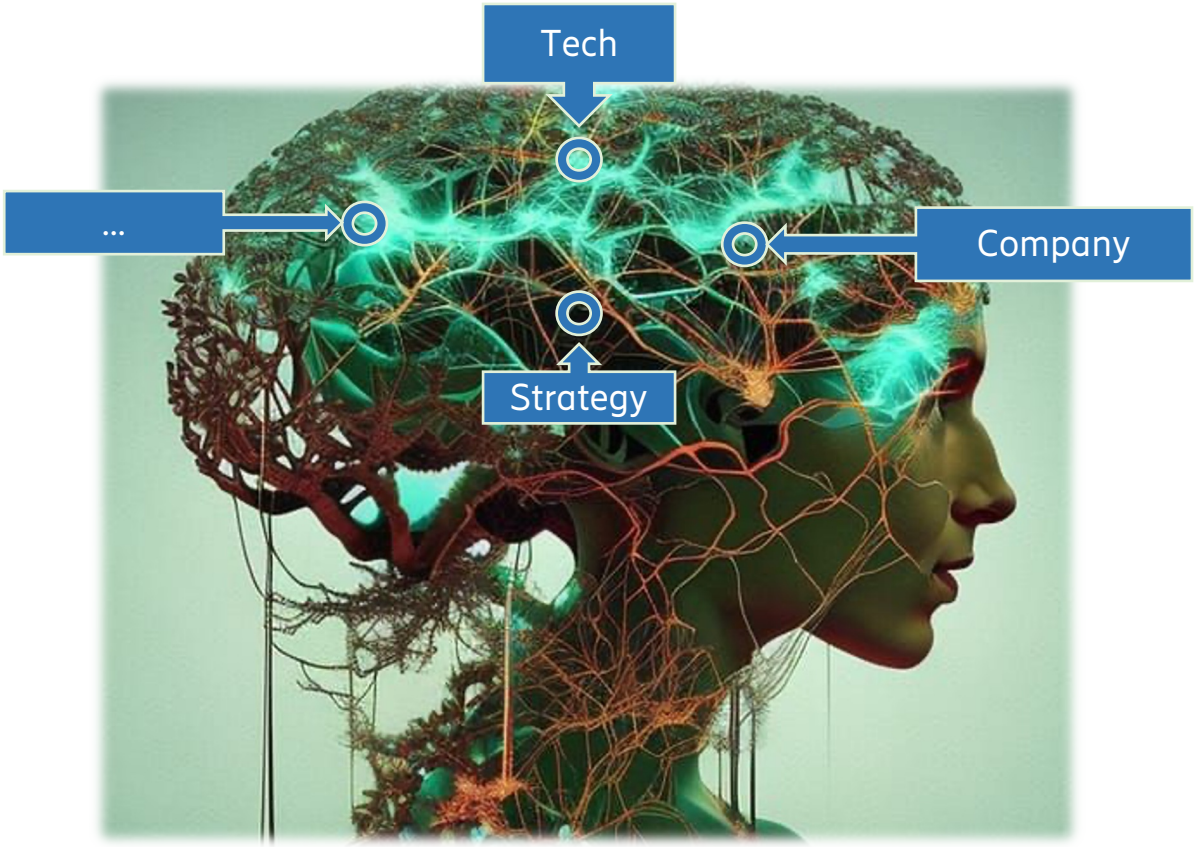
- Many pieces have to fall in place, in addition to the tech



Non-Technical Aspects

These are important, too:

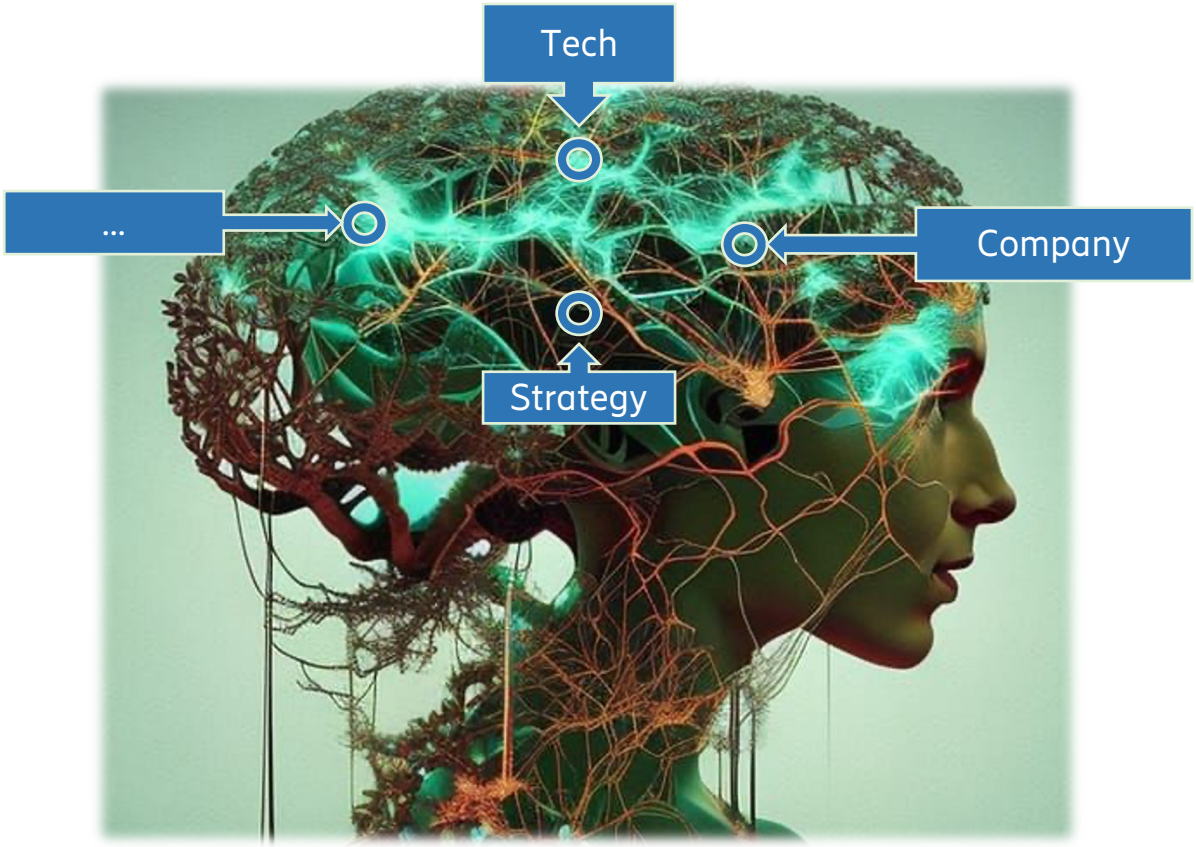
- Getting people involved
  - Competence
  - Top-down & bottom-up
- Your data
- Strategy
- Managing uncertainty
- Expectations
- Legal
- Sourcing & partners
- Roles in the ecosystem



Non-Technical Aspects

The lesson is:

A technology change can not exist alone, it needs to fit a broader picture



Non-Technical Aspects

The lesson is:

A technology change can not exist alone, it needs to fit a broader picture

Even if we talk only about technical changes in products, we still need to look at development, those products, and operations together

Agenda

- Introduction
- What is Ericsson & Telecom
- Generative AI at Ericsson
- Generative AI in Telecom
- Example technical topics
- Evolving Generative AI
- Summary

Generative AI in Telecom

Very broad: part1 vs. part2 vs. …

And products vs. operations vs. development

Generative AI is also not our only tool

- In the toolbox there's also "old-fashioned" AI (predictive, discriminative), all of software engineering, hardware engineering, signal processing, math, …

So where **does** it fit in?

Impacts

1. Generative AI & AI as an app
   - What traffic will be carried by networks?
   - Present day usage not much different from other applications
   - Training requires new arrangements, but it is largely done in specialized data centers
   - But looking ahead, we can see some potential impacts:
     - Results to the needed place
     - Size/delay of inputs & training
     - Gen. AI for data reduction

Impacts

1. Generative AI & AI as an app
2. Generative AI inside telecom
   - What can we use it for?
   - Inside products, in their construction, in operations

Categories of Uses

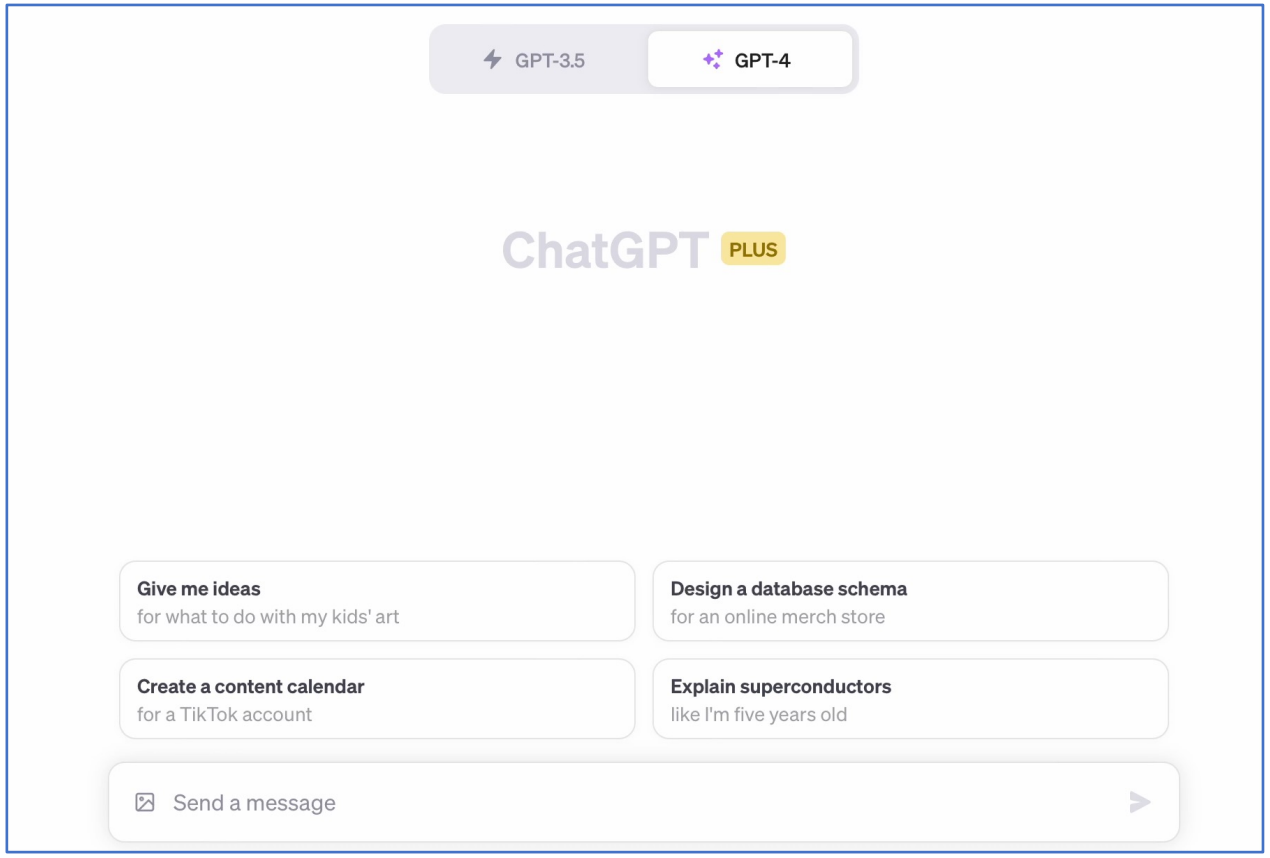
1. General assistance
    - Writing, drawing, slides, normal business routines

Categories of Uses

1. General assistance
   - Writing, drawing, slides, normal business routines
   - Summarization and other semantic tasks

Categories of Uses

1. General assistance
2. Software engineering

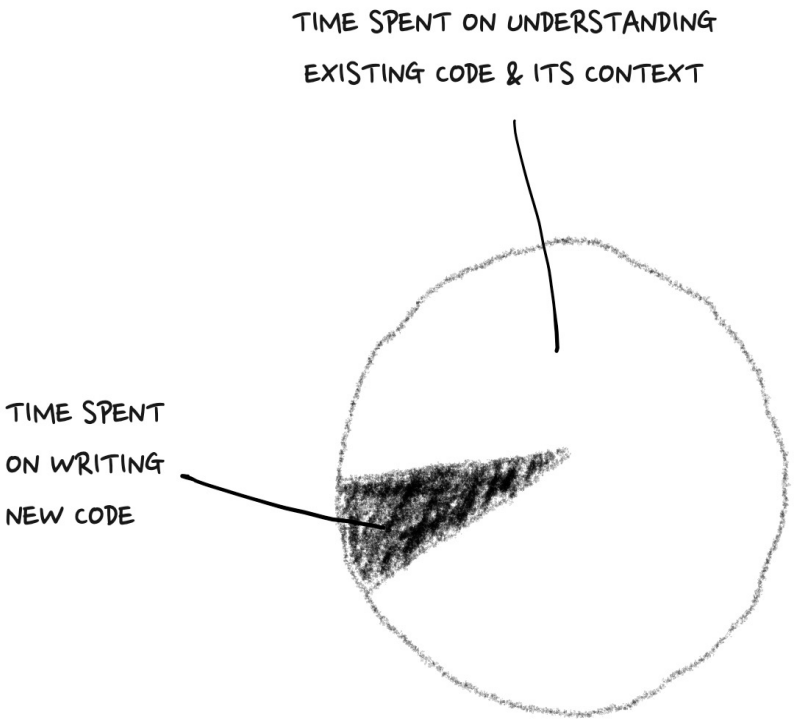


Categories of Uses

1. General assistance
2. Software engineering
   - Programming
   - API assistance

Categories of Uses

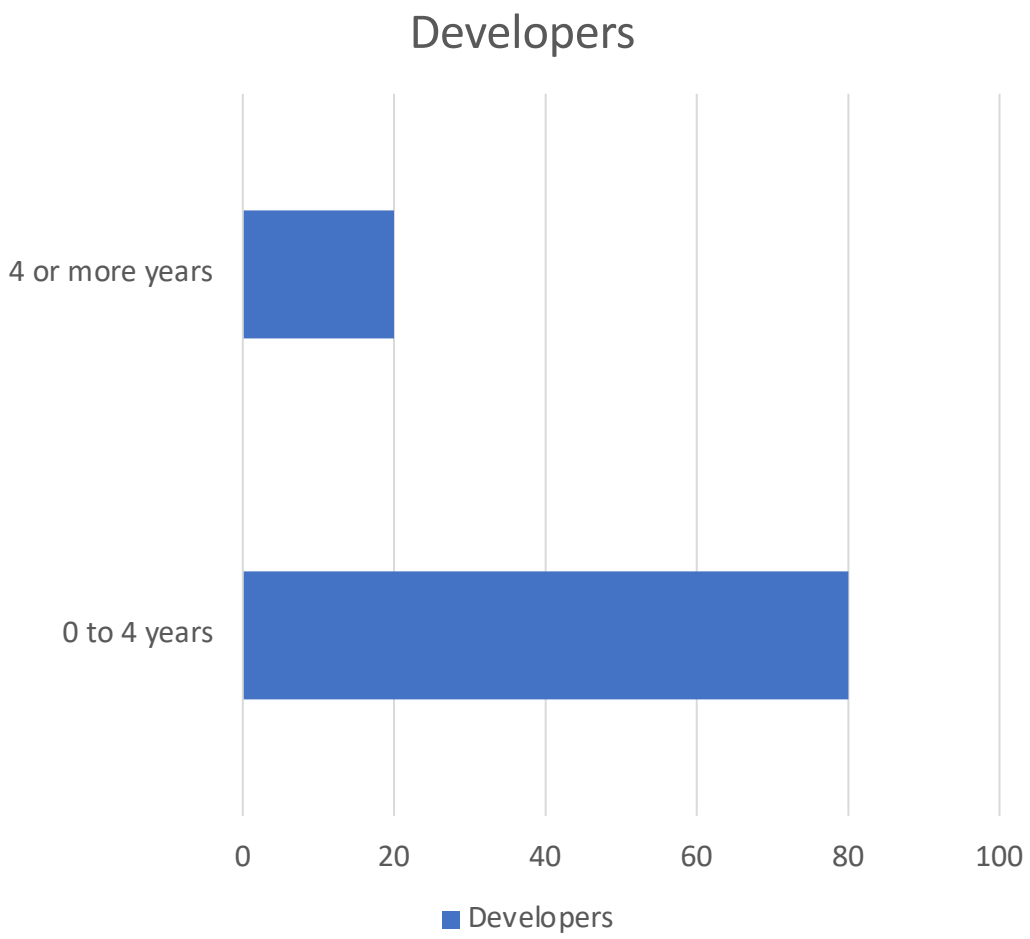
1. General assistance
2. Software engineering
    - Programming
    - API assistance
    - Test generation



Hindered productivity

Categories of Uses

1. General assistance
2. Software engineering
   - Programming
   - API assistance
   - Test generation
   - Code comprehension



Categories of Uses

1. General assistance
2. Software engineering
   - Programming
   - API assistance
   - Test generation
   - Code comprehension
     - What code implements this feature? What changed for bug X?

Categories of Uses

1. General assistance
2. Software engineering
3. Chatbots & knowledge bases

What do LLMs know?

LLMs are trained with large amounts of Text

Domain Knowledge (Telecom Text)

Organization or domain-specific knowledge

What is UE?

- **User Equipment**
- Unreal Engine
- User Experience
- Union Européenne
- University of Edinburgh

Telecom Dataset for Language Modeling

Large **unstructured** textual data is the key ingredient to recent advances

3GPP specifications

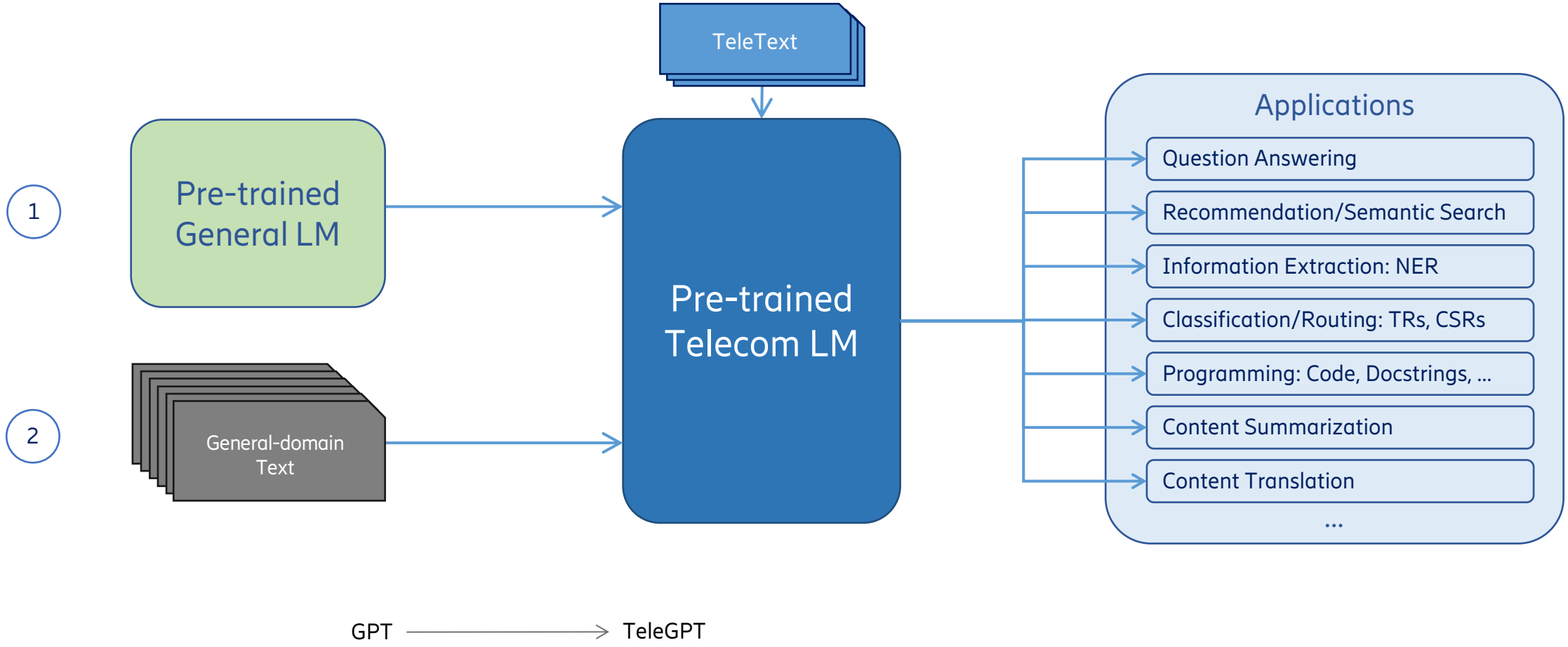
+

ERICSSON

Ericsson Internal Technical Documents

Pre-Training Telecom Language Models

Incorporate telecom-domain knowledge into the models in the pre-training phase.



GPT: Generative Pre-trained Transformer

Blog Post: https://www.ericsson.com/en/blog/2022/1/neural-language-models-telecom-domain

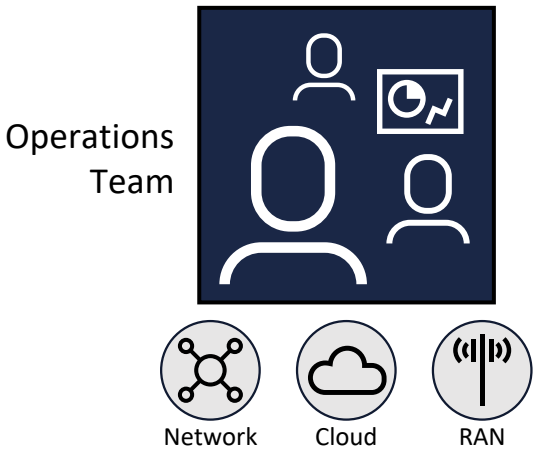
Categories of Uses

1. General assistance
2. Software engineering
3. Chatbots & knowledge bases
4. Data generation
   - Missing data (e.g., XR)
   - Compression/semantic comm. (e.g., radio CSI)
   - Training data (e.g., other AIs)

Categories of Uses

1. General assistance
2. Software engineering
3. Chatbots & knowledge bases
4. Data generation
5. Network management
    - Creating intents
    - Creating configurations
    - Deployment plans

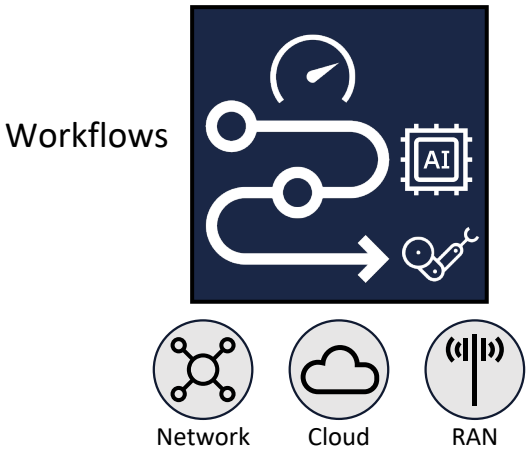
Manual → Automated → Autonomous



Manual operations

Task execution
Human

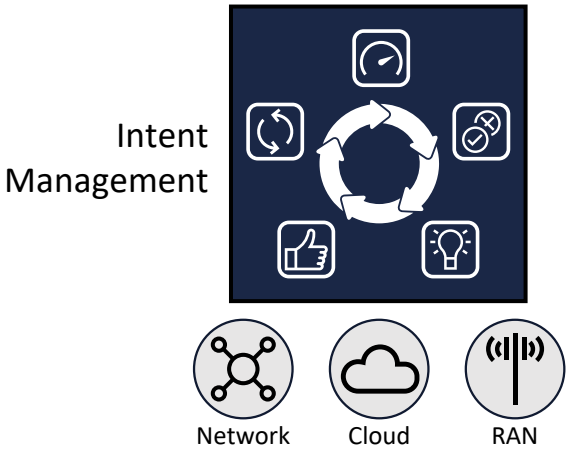
Which tasks to perform?
Human



Automated execution

Task execution
Automated

Which tasks to perform?
Still human



Adaptive autonomy

Task execution
Automated

Which tasks to perform?
Gradually more automated

What is an intent?

"Intent is the formal specification of all expectations including requirements, goals, and constraints given to a technical system"

TM Forum

Example:

I want a URLLC service;

the users shall experience

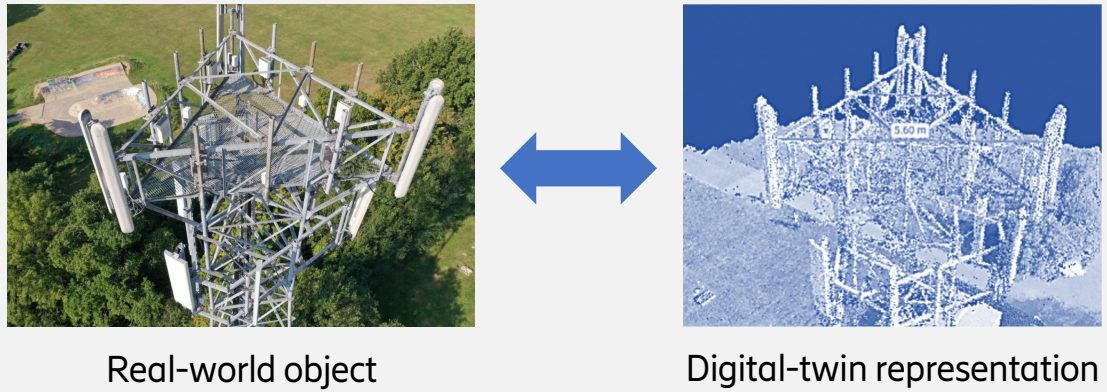
- a max latency of 15 ms, and
- a max packet loss of 0.001%

```
zt:latencyCrititicalServiceIntent1
  a zt:MyIntent ;
  zt:incurringPenalty 1 ;
  cc:hasExpectation
  [ a cc:PropertyEqualExpectation ;
    cc:target ne:urllc-service-instance;
    cc:params [ zt:serviceType ne:urllc;
                tel:usrGrp ne:urllc1_usrg_group_instance
              ];
  ],
  [ a zt:ThresholdedMetricLessThanExpectation;
    cc:hasContext ne:urllc-service-instance;
    cc:percent 1.0;
    cc:target ne:urllc1_usrg_group_instance;
    cc:params [ tel:latency 15
              ];
  ],
  [ a zt:ThresholdedMetricLessThanExpectation;
    cc:hasContext ne:urllc-service-instance;
    cc:percent 1.0;
    cc:target ne:urllc1_usrg_group_instance;
    cc:params [ tel:packetLoss 0.001
              ];
].
```

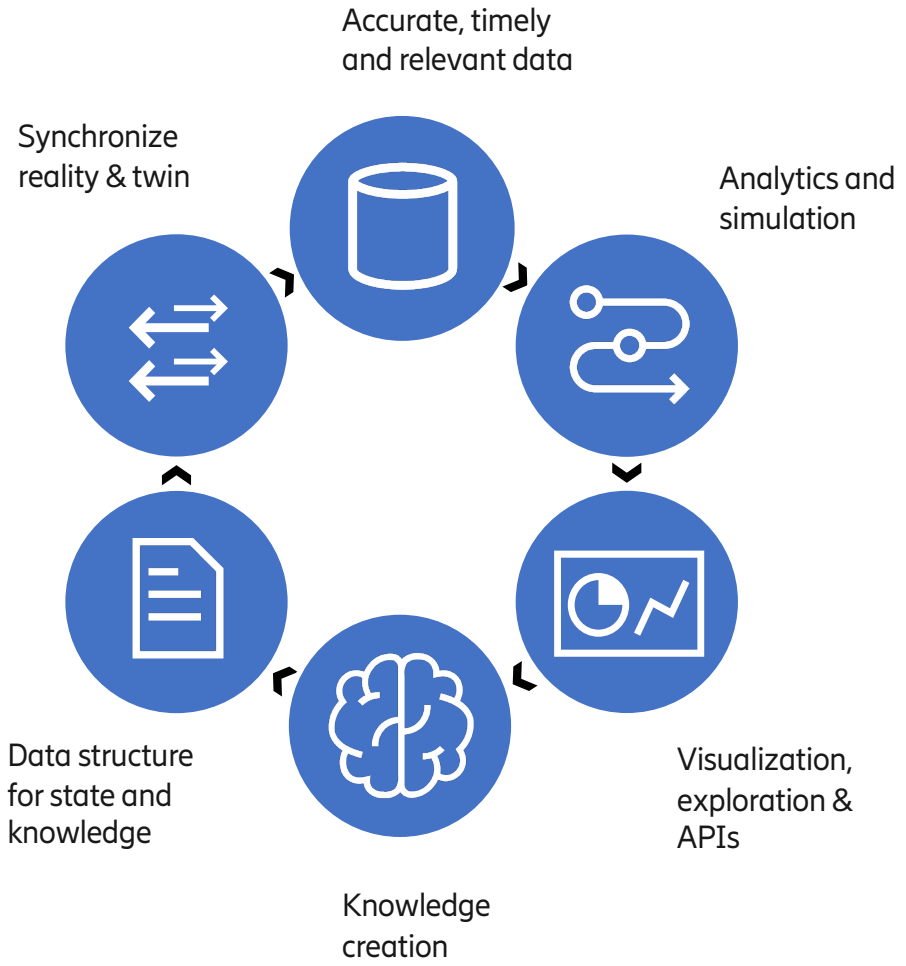
Categories of Uses

1. General assistance
2. Software engineering
3. Chatbots & knowledge bases
4. Data generation
5. Network management
6. Digital twins

Network digital twins – overview

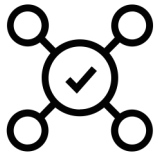


A network digital twin creates a mirror image of the network, augmented with tools for specific purposes.

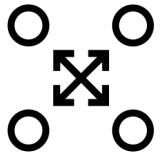


Digital Twin Use Cases

Modelling of traffic and user behaviors



Verification of new software / algorithms



Impact of service changes on the end-to-end network

Cost and performance of new network services

"What-if" scenarios – equipment failures, disaster scenarios

Sandbox training environment

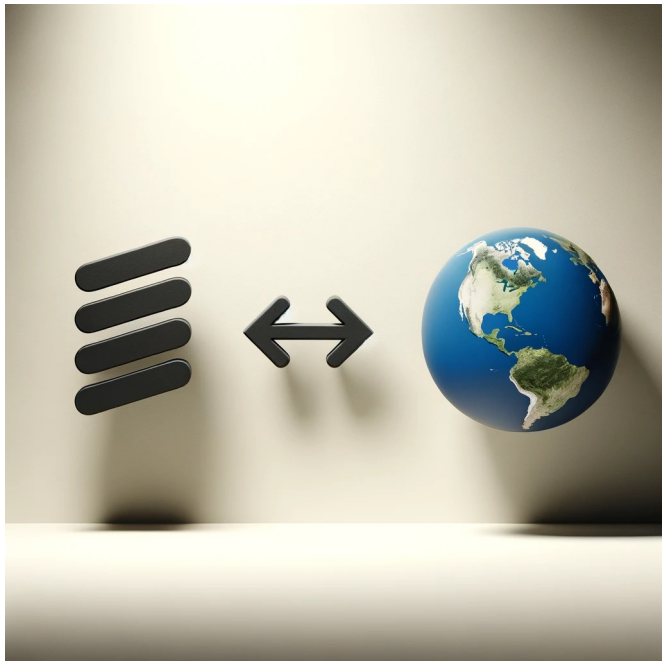
Agenda

- Introduction
- What is Ericsson & Telecom
- Generative AI at Ericsson
- Generative AI in Telecom
- Example technical topics
- Evolving Generative AI
- Summary

Some interesting topics:

- Limitations
- Safety
- Explainability
- Beyond human languages
- Reasoning
- New things

Limitations



- Hallucination
- Plausible-but-false results

Fundamental issue, not just an artefact of current versions

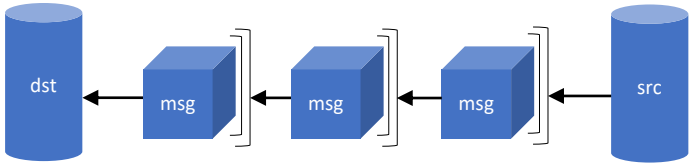
Safety of trusting the results

- Generative AI is not at the level where we can trust the results in situations where the consequences of an error are severe
- Human in the loop (e.g., programming assistance)
  - Most of the use cases discussed earlier fall in this category
- Guardrails
  - In the AI model itself
  - Hard limits in the system around it, e.g., cannot influence system-wide configuration, firmware, limited influence only on task AI is given
  - Intent-based management and configuration
- Explainability

Other safety issues

- Ethical vs. misuse
- Data leaks, data feeding
- Dependencies and gatekeepers

Beyond human languages



- Much of telecommunications is about software systems or nodes talking to each other over standardized protocols or APIs
  - RFCs, 3GPP specifications, web APIs, open-source toolkit APIs, …
  - However, fully machine-readable formal specifications of protocols are rare
- LLMs are reasonably well aware of the basics in common protocols
  - Even more detailed awareness of the protocol specification is important in many situations, e.g., programming, debugging, fault analysis
  - Multi-modality: awareness of specifications, program code, packets
  - Could LLMs “understand” protocols, like they “understand” other languages?

Achieving bigger things

Having the LLMs assist in small tasks is good, but … what about bigger tasks?

- Writing or debugging larger pieces of software
- Proposing improvements in some arrangement, e.g., a protocol design

To some extent such tasks can be handled stepwise or through bringing a selected information to the LLM's context (prompt), etc. LLMs have produced some impressive step-by-step "logical thinking" results when correctly prompted.

Ultimately, better tools may be needed

- Reasoning, planning, state & exploration, multiple agents working together

What is reasoning?

A cognitive process that draws conclusions from data and knowledge using logical techniques such as deduction, abduction and induction.

In the simplest form this can be just nudging an LLM into a “Chain-of-Thought”:

- In the simplest form, just add “Let’s think step-by-step” at the end of a prompt ☺
- Or through examples

Good survey paper on this is Jie Huang, Kevin Chen-Chuan Chang, Towards Reasoning in Large Language Models: A Survey, arXiv:2212.10403

New Things

- Much of what you've seen discussed is mimicking human tasks
- "Make our product better"
- "A new way to interact with <X>"
- "Assist humans"
- But usually, technology revolutions create entirely new things, things that we didn't even know we needed – or in this case perhaps we can do things that humans never could or will be able to do
- What is the new thing enabled by Generative AI?

Agenda

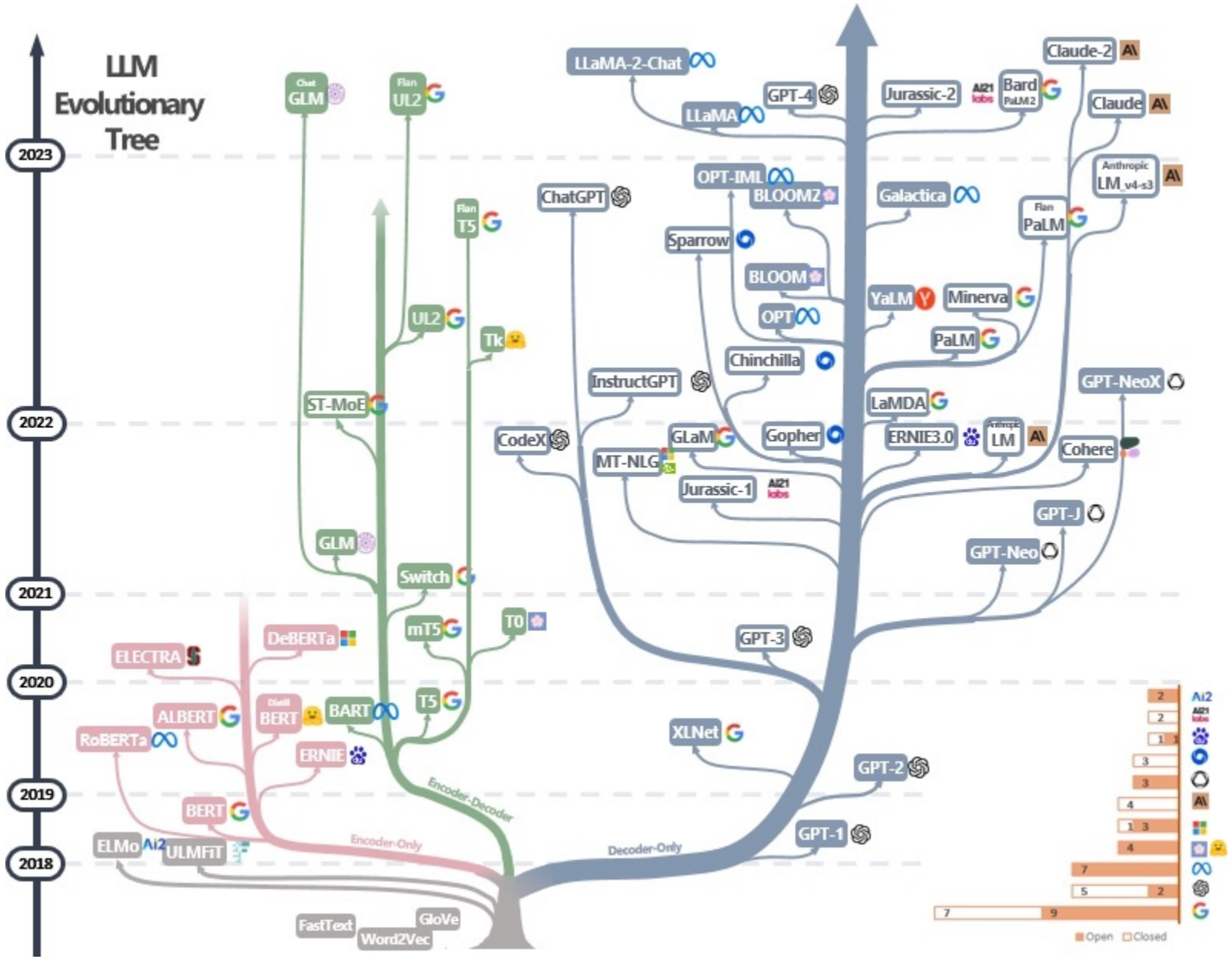
- Introduction
- What is Ericsson & Telecom
- Generative AI at Ericsson
- Generative AI in Telecom
- Example technical topics
- Evolving Generative AI
- Opportunities & challenges

Evolution

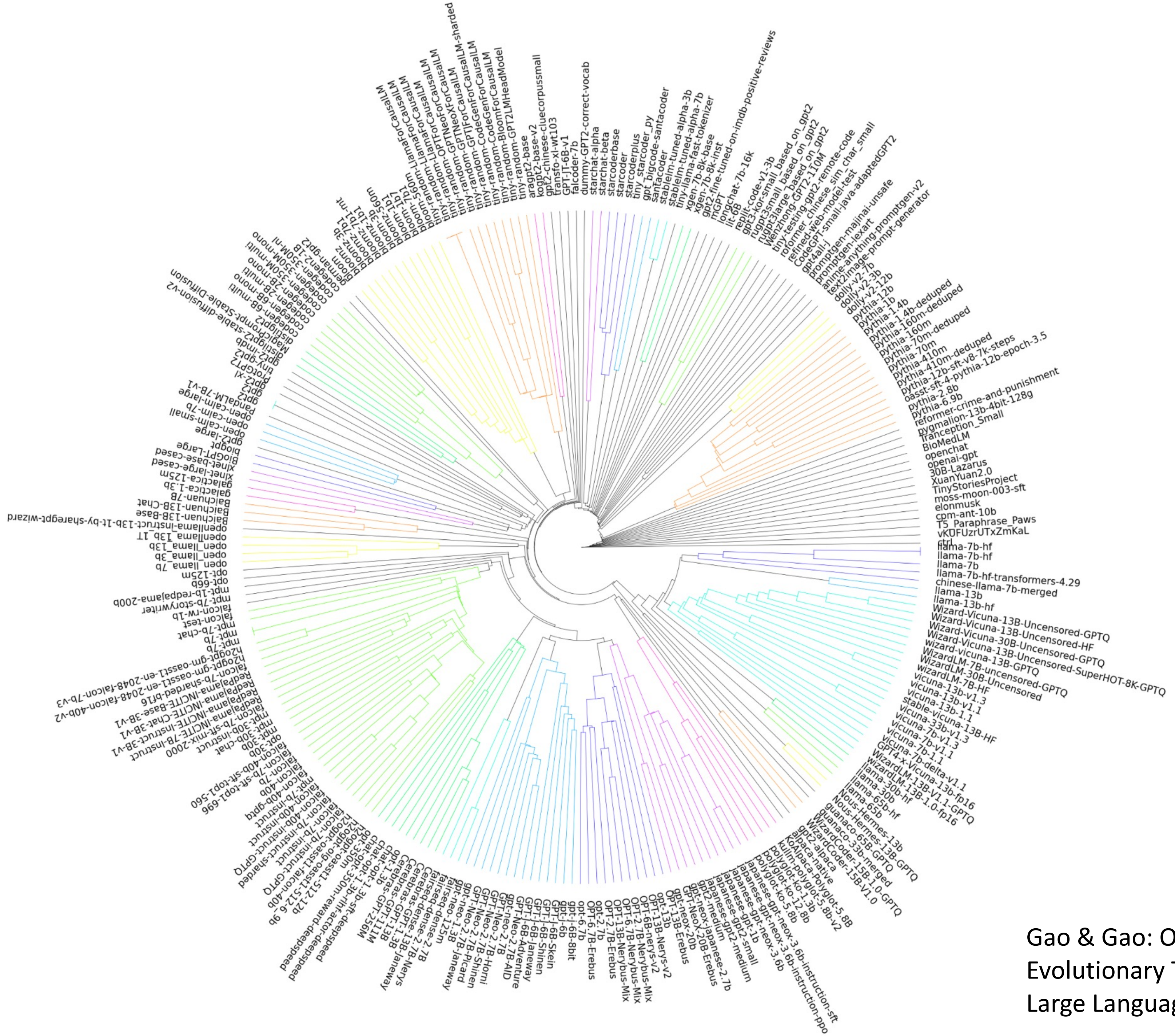
Generative AI is evolving very rapidly

Better technology and training methods, not just adding more GPUs

Example: Open-source data sources and LLMs developing rapidly during 2023



Yang et al: Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond



Gao & Gao: On the Origin of LLMs: An Evolutionary Tree and Graph for 15,821 Large Language Models

Open Source or Local LLMs (LLLMs?)

- It is also possible to run your own LLM.
- Some are competitive to best commercial ones
- Benefits
  - Avoid leaking
  - Tweak the software or model
  - Inexpensive forking
  - Avoid any guardrails or limitations
  - Can run without connectivity
  - Transparency & learning
- Risks
  - Licensing
  - Performance
  - Misuse

Model	Developer	Size	Context	License
Llama2	Meta	7B to 70B	4K	Meta
Mistral	Mistral.ai	7B	8K	Apache 2.0
Falcon	TII UAE	7B to 180B	2K	Apache 2.0 or Falcon
MPT	MosaicML	7B to 30B	8K to 65K	Apache 2.0
BLOOM	BigScience	1B to 176B	2K	RAIL License

Source: 8n8.io

Agenda

- Introduction
- Generative AI at Ericsson
- Generative AI in Telecom
- Example technical topics
- Evolving Generative AI
- Summary

Summary

- Generative AI will have a significant impact
- Influences way we work as well as the products we build
- Thinking about where it applies is important!

  Chat, co-pilot, etc. today

  Something else entirely tomorrow

Summary 2

- Avoid overestimating
  - Comes with many limitations
  - Understand the costs
- Just one tool among many – not a solution for all problems
  - A good model is to apply CS + AI + …
  - Hire AI *and* domain experts
- The field continues to rapidly evolve, we know some of the implications but not all
  - There’s potential for much bigger capabilities