Main Circle 2024

It was a lot of fun …

> ~ 70 Teilnehmer an 3 Terminen und tollen Locations – danke nochmal an die Sponsoren/Hosts
  > Sandra Janicijevic und Oliver Naegele, FinTech Headquarter
  > Martina Williams, JLL und
  > Natalie Wehrmann, CBRE
> Einen großen Dank auch an unsere ‘Facilitators’ von Konii – Andreas Söntgerath und Heiko Danz!
> 4 Fachvorträge – danke HC Fuchs und Ramin Moktari!
> 3 × AI Newflash vom ‘Program Committee’
> Eine Menge Diskussionen zwischen den Teilnehmern

> **Insgesamt genau das, was wir hatten erreichen wollen - und mehr!**
> **Danke nochmal von Sebastian, Christian und Marc!**

Main Circle 2025

… and the fun continues

> Es geht los am **29. Januar, 19.00 Uhr bei Ardian** mit einem Vortrag von **Sebastian Schirber, artefact** – und natürlich wieder dem AI News Flash

> Weitere Termine
  > **27.2. 19.00 Uhr**
  > **27.3. 19.00 Uhr**

> Da die Zahl der Interessenten wächst, werden wir eine Mitgliedschaft einrichten und Mitglieder bei der Anmeldung bevorzugen

> Ihr als Teilnehmer ‘der ersten Stunden’ seid natürlich Mitglieder ☺

> Für die Mitglieder möchten wir eine (geschlossene) Diskussionsgruppe auf LinkedIn einrichten, um uns auch jenseits der Termine auszutauschen und Neuigkeiten zu posten

> **Was denkt ihr über einen 1-Tages-Event mit Program im Q2/25?**

AI News Flash

2024 Highlights

Textverbesserungen, -vorschläge, –korrekturen, -zusammenfassungen etc. sind bereits Alltag.

Staunen, Bewunderung, Ehrfurcht vor dem was die generativen Systeme bereits alles können …

Es gibt bereits eine gewissen produktive Nutzung in den meisten Unternehmen.

Modelle sind in verschiedendsten Varianten und mit verschiedener Privacy zu attraktiven Preisen in der Cloud verfügbar.

Die 2024 Downsites

Datenanalysen im Prompt sind fehlerhaft und ändern sich.

Ernüchterung darüber was alles nicht funktioniert

Ergebnisse lassen sich nicht ohne Überprüfung verwenden.

LLMs liefern keine Anwendungen, ihre Verwendung muss man wie jedes andere Softwaresystem in eine Systemlandschaft einbetten – und meistens auch noch programmieren.

2024 was das Jahr des RAG GenAI

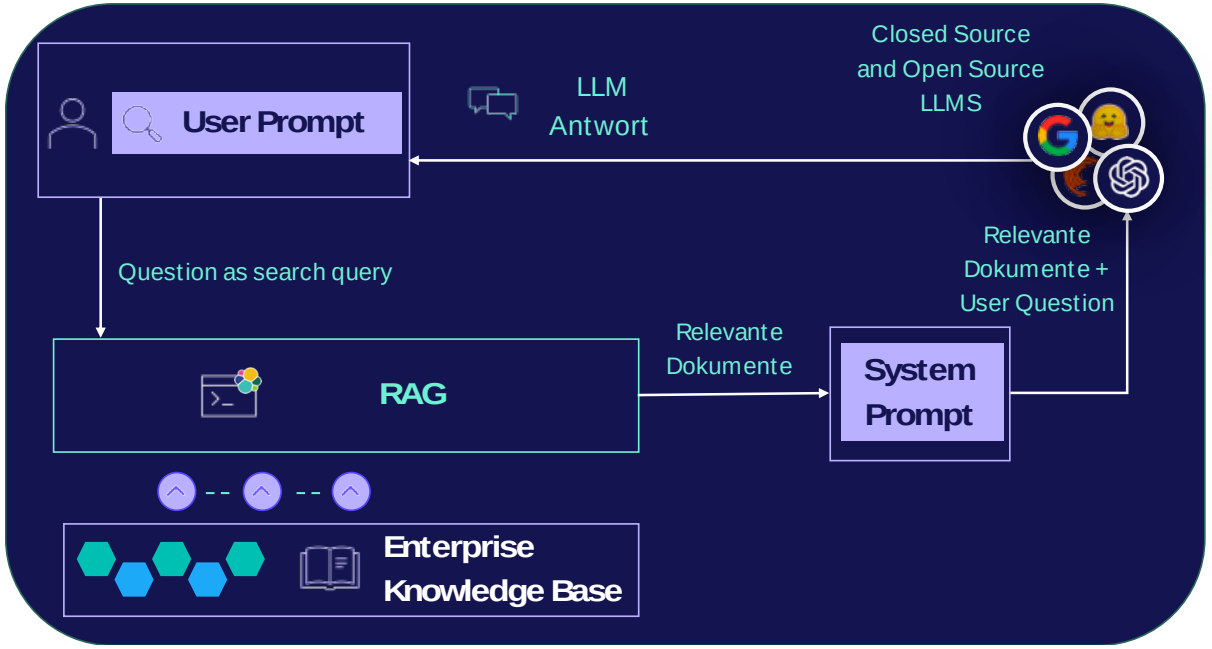
Der Einsatz von RAG fuer GenAI Uses Cases in Unternehmen hat ein Auf und Ab erlebt

Nach intensiver Diskussion über die Schwächen und Stärken von RAG steht Ende 2024 fest, dass RAG eine wichtige Rolle im Einsatz von GenAI in Unternehmen spielt

Das waren die wichtigsten Punkte in der RAG Diskussion 2024:

- RAG versus LLM fine tuning: in den meisten Anwendungsfällen ist RAG kostengünstiger und performanter bei gleicher Qualität
- Opensource LLMs schliessen in der Qualität auf und finden vermehrt Einsatz bei RAG Anwendungen
- Long Context LLMs und RAG co-existieren am besten zusammen. Long-Context LLM erlauben ein besonders grosses Kontextfenster im Systemprompt („to RAG or not to RAG“)
- RAG basierte Anwendungen waren auf textuelle Dokumente beschränkt, in 2024 nahmen sogenannte multi-modale Parser Fahrt auf, die PDFs, PPTs, Tabellen, Images und andere Formate verarbeiten können

RAG steht fuer Retrieval Augmented Generation und dient zur Verbesserung des Einsatzes von GenAI in Unternehmen durch Einbindung von unternehmensspezifischen Wissensquellen, die die Genauigkeit und Relevanz der Antworten von LLMS verbessert.



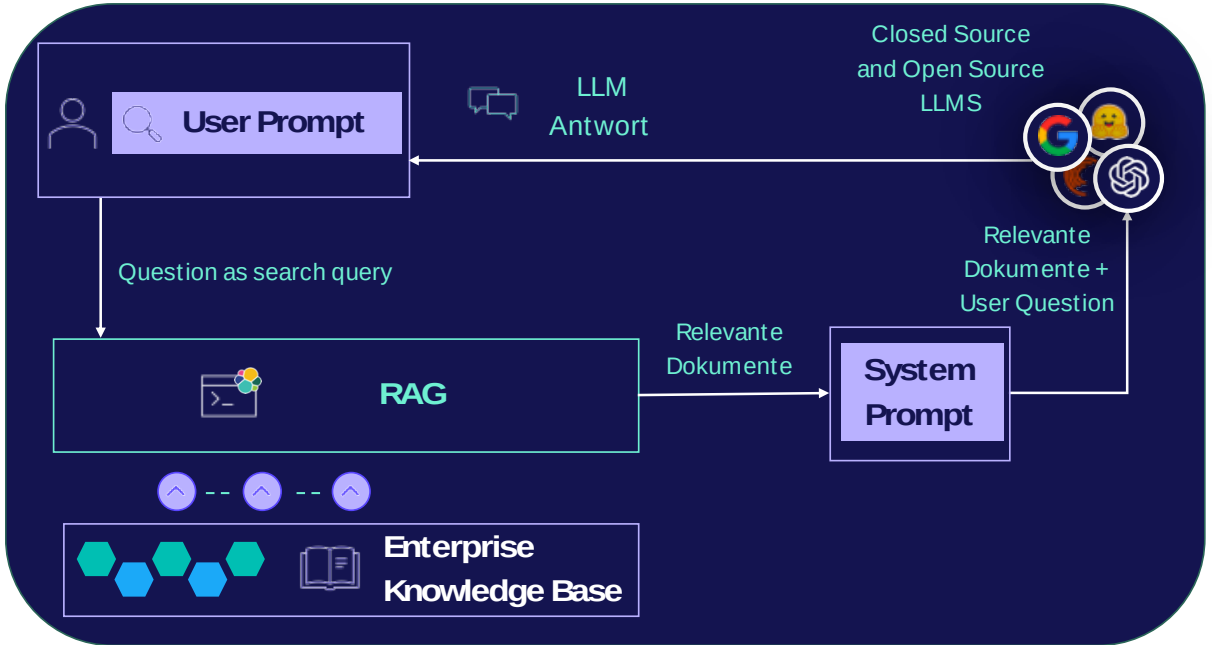
2024 was das Jahr des RAG GenAI

Der Einsatz von RAG fuer GenAI Uses Cases in Unternehmen hat ein Auf und Ab erlebt

Das waren die wichtigsten Punkte in der RAG Diskussion 2024:

- Die Qualität der Antworten in einem RAG basierten GenAI Anwendung hängt stark von der Retrieval Komponente ab, die möglichst die optimalen Unternehmensdokumente zu einer Frage identifiziert. Hybride Suchverfahren setzen sich durch (z.B. semantische Vektorsuche und BM25)
- Ein grosser Einflussfaktor auf die Qualität der Antworten auf eine Frage an eine RAG basierte Anwendung ist das sogenannte Chunken, d.h., das Aufteilen eines Dokumentes in kleiner Teile. Eine Reihe neuer Verfahren (Contextual Retrieval, Meta Chunking, dsRAG e.a.) wurden untersucht und werden in die mainstraim RAG tools und Systeme einfliessen

RAG steht fuer Retrieval Augmented Generation und dient zur Verbesserung des Einsatzes von GenAI in Unternehmen durch Einbindung von unternehmensspezifischen Wissensquellen, die die Genauigkeit und Relevanz der Antworten von LLMS verbessert



Neuigkeiten aus dem GenAI RAG Umfeld

Die Weiterentwicklung von LLMs ist ungebrochen sowohl als closed und vermehrt auch open source

- InternVL 2.5. is the First open-source model to surpass 70% accuracy on the MMMU benchmark. This achievement brings open-source capabilities closer to commercial systems like GPT-4V. To be noted though that the benchmark caputre only a subset of overall capabilities
- Amazon is reportedly preparing to unveil “Olympus,” a multimodal large language model (LLM) capable of processing text, images, and videos. The rumored Olympus model, potentially boasting 2 trillion parameters, may represent a new evolution or iteration of Amazon’s previous LLMs
- QwQ: Reflect Deeply on the Boundaries of the Unknown | Qwen The QwQ-32B-Preview, developed by the Qwen Team, explores AI’s reasoning capabilities through reflective problem-solving and self-questioning. Designed for tasks like mathematics, coding, and logical reasoning, it achieves notable benchmarks: 65.2% on GPQA, 50.0% on AIME, 90.6% on MATH-500, and 50.0% on LiveCodeBench
- The Allen Institute for AI (Ai2) has launched OLMo 2, its latest open-source language model series, featuring 7B and 13B parameter models. These models, built entirely with publicly accessible tools and data, outperform comparable models like Meta’s Llama 3.1 in tasks like text summarization, coding, and Q&A.

https://www.linkedin.com/pulse/ai-research-weekly-roundup-genai-works-gk0ce/
https://medium.com/@eugina.jordan/gen-ai-for-business-newsletter-33-f39f7c264356

Einsatz von GenAI in Unternehmen

Some lighthouse deployments of GenAI in 2024

- McKinsey's generative AI platform, **"Lilli,"** streamlines access to it's century-long knowledge base and enhance cleint service: orchestrates large and small AI models securely and efficiently, with applications such as transforming writing into high-quality insights
- Fujitsu has developed "Policy Twin," a generative AI-powered digital twin solution for Japanese municipalities to optimize healthcare policy-making
- Walmart deployed a Generative AI platform to create personalized marketing content tailored to individual customer preferences and previous purchasing behavior. This AI analyzes customer data to generate customized product recommendations, promotional emails, and dynamic website content, creating unique homepages for each shopper
- Toyota has embraced Generative AI to revolutionize its vehicle design process. This technology allows for the automatic generation of numerous potential vehicle designs, considering critical parameters such as fuel efficiency, aesthetics, and cost
- HSBC implemented an AI-driven system to enhance its anti-money laundering (AML) efforts. This system uses highly developed machine learning algorithms to evaluate unusual patterns and potentially illegal activities across various transactions, distinguishing them from normal activity more effectively than traditional methods.

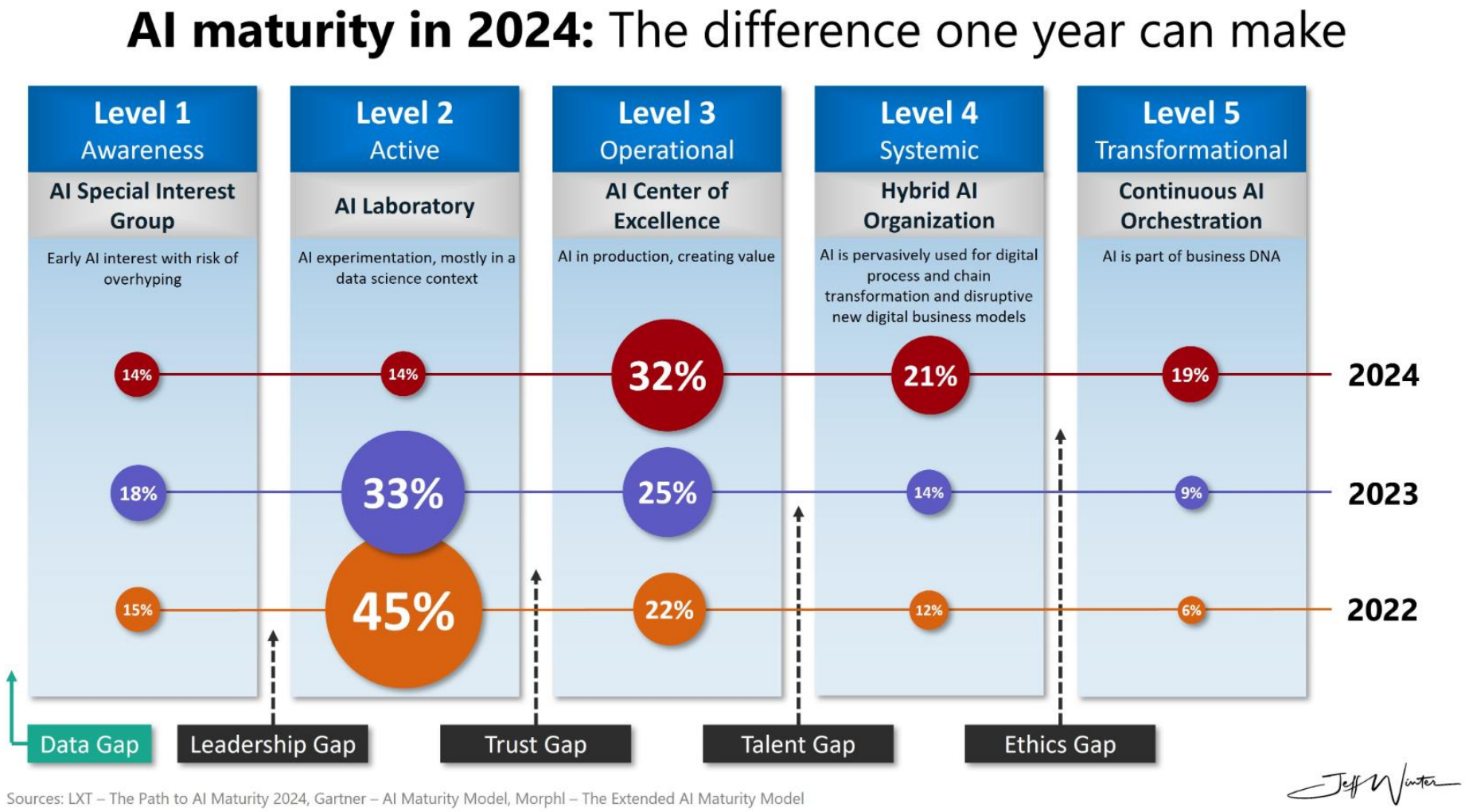
Einsatz von GenAI in Unternehmen

10 Trends to watch when applying GenAI

1. Data-driven organizations are best placed to take advantage of GenAI
2. Organizations are scaling GenAI carefully
3. There is a strong awareness of risks
4. GenAI is improving productivity but some organizations are unsure what to do with the time freed
5. Improving the quality of work is another important driver for deployment of GenAI
6. People are not always comfortable with the outcomes of using the technology
7. GenAI cannot be implemented without change management
8. Most organizations don’t know exactly what percentage of their workforce is using GenAI
9. Few organizations have developed a strategy for sustainable use of AI
10. Removing humans from the loop is still considered to be a mistake

KI Reifegrad in 2024

Der Unterschied, den ein Jahr ausmachen kann



Quelle: Jeff Winter

KI Reifegrad in 2024

Der Unterschied, den ein Jahr ausmachen kann

> **Rasante Entwicklung der KI:** Der Artikel hebt die schnellen Fortschritte und die zunehmende Integration von KI in Unternehmensstrategien hervor.
> **Generative AI-Marktprognose:** Bis 2032 soll der Markt ein Volumen von 1,3 Billionen USD erreichen, mit einer jährlichen Wachstumsrate von 42%.
> **Gartners AI Maturity Model:** Dieses Modell hilft Unternehmen, ihren aktuellen Reifegrad in der KI-Entwicklung zu bestimmen.
> **Fortschritte bei der KI-Integration:** Ein 24%iger Anstieg von Unternehmen zeigt, dass viele vom Experimentierstadium in die Reifephase übergehen.
> **Veränderte Prioritäten:** KI wird zunehmend im Risikomanagement eingesetzt, welches die Geschäftsfähigkeit als Haupttreiber überholt hat.
> **Fokus auf Suchmaschinen und prädiktive Analysen:** Diese Bereiche zeigen den höchsten ROI und die stärkste Implementierung.
> **Priorisierung von Generative AI:** 69% der Unternehmen setzen auf generative KI, trotz Herausforderungen bei Implementierung und ROI, was den Entwicklungsstand der Technologie reflektiert.

Quelle: Jeff Winter

KI Reifegrad in 2024

Wie Sie Ihren AI-Reifegrad verbessern können

1. **Integration von KI- und Datenstrategien in die Unternehmensziele**: Sicherstellen, dass KI-Projekte mit den Kernzielen des Unternehmens übereinstimmen.
2. **Förderung einer unternehmensweiten KI-Priorität**: KI sollte nicht nur IT-Abteilungen betreffen, sondern im gesamten Unternehmen verankert sein
3. **Schaffung einer KI-freundlichen Unternehmenskultur**: Mitarbeiter über die Vorteile von KI aufklären und Schulungen anbieten.
4. **Rekrutierung und Entwicklung geeigneter Talente**: Einstellung von Fachkräften wie Datenwissenschaftlern und KI-Ethikern.
5. **Einrichtung eines KI-Ethikkomitees**: Verantwortungsbewusster Einsatz von KI durch Richtlinien und regelmäßige Überprüfungen.
6. **Standardisierung und Austausch bewährter Verfahren**: Entwicklung von Best Practices für den konsistenten und zuverlässigen Einsatz von KI.

Quelle: Jeff Winter

KI Reifegrad in 2024

Quellen

> Bloomberg - Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds:
https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

> Gartner - The CIO's Guide to Artificial Intelligence 2019:
https://www.gartner.com/smarterwithgartner/the-cios-guide-to-artificial-intelligence

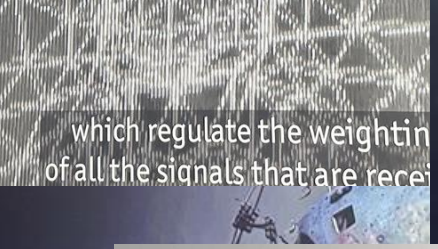
> LXT's The Path to AI Maturity 2024 - and Executive Survey:
https://info.lxt.ai/the-path-to-ai-maturity-2024?utm_source=influencer&utm_medium=linkedin&utm_campaign=jeff_winter_path_to_ai_maturity_2024

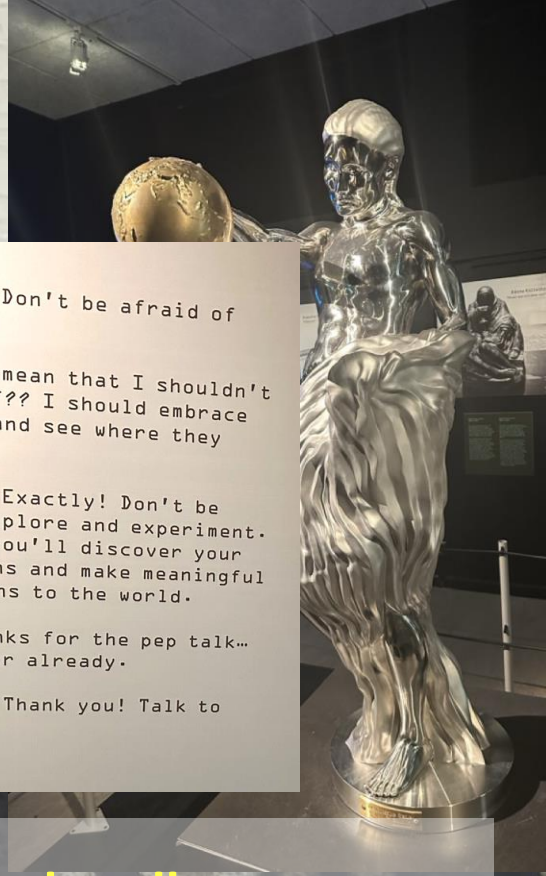
> Jeff Winter:
https://www.jeffwinterinsights.com/insights/ai-maturity-in-2024-the-difference-one-year-can-make

Was 2025 bringen wird

Und was wir uns sicher im MC ansehen werden

Breiterer, zentraler Nutzen des Bestehenden	Mehr Produktive Anwendungen RAG, verschiedene Modelldeployments, große Kontexte
Weitere bahnbrechende Neuerungen	Reasoning/Inference-time Compute, Agentic AI, very large models, very small models, near-infinite memory





Narrator: This is a dialogue between a human and their inner self that is written by an AI tool and read by AI-generated voices.

.....

Person: Hey there, inner self! It's been a while.

Inner Self: Hey, what's up?

Person: I don't know how I feel about AI. On one hand, it's fascinating technology that has the potential to improve our lives in so many ways. But on the other hand, I worry about the implications for humanity.

Inner Self: What exactly are you worried about?

Person: I'm worried that AI could eventually replace human beings. That we'll create machines that are so intelligent that they no longer need us. That we'll be obsolete.

Inner Self: But is that really a bad thing? If AI can do things better than we can, shouldn't we embrace it?

Person: I don't know. It just feels like we're playing God. Like we're trying to create something that's better than ourselves. And I worry that we'll lose sight of what it means to be human.

Inner Self: But what does it mean to be human? Is it simply about having a body and a brain? Or is it something more than that?

Person: I don't know. I guess I've just always assumed that there's something special about being human. That we have some sort of intangible quality that sets us apart from everything else.

Inner Self: But is that really true? What if AI can develop that same quality? What if they can experience emotions and create art and do all the things that you can do?

Person: Yes, but at what cost? AI has the potential to take over so many jobs that are currently done by humans. What will happen to all those people who lose their livelihoods?

Inner Self: I still think that we should keep pushing the boundaries of what AI can do. Maybe we need to be more responsible in how we use it, but we can't just stop exploring its potential.

Person: I think we need to be careful. We need to remember that AI is just a tool. It's not a replacement for human life and existence.

Inner Self: But I also think that AI has the potential to do a lot of good. It can help us solve some of the world's biggest problems, like climate change and disease.

Person: But is that really a good thing? If we rely too much on AI to solve our problems, what happens when the machines break down or malfunction? What happens when we can't fix them ourselves?

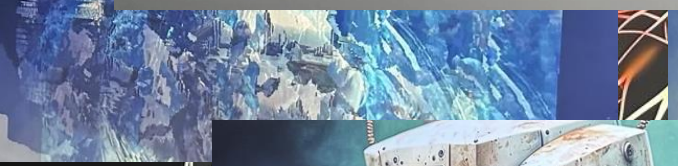
Inner Self: Don't be afraid of the unknown.

Person: You mean that I shouldn't limit myself?? I should embrace everything and see where they take me?

Inner Self: Exactly! Don't be afraid to explore and experiment. That's how you'll discover your true passions and make meaningful contributions to the world.

Person: Thanks for the pep talk… I feel better already.

Inner Self: Thank you! Talk to you later!



Mit Impressionen aus der aktuellen AI-Austellung des Tekniska Museet in Stockholm wünscht euch euer MainCircle Program Committee ein tolles 2025 – wir freuen uns darauf euch wiederzusehen!