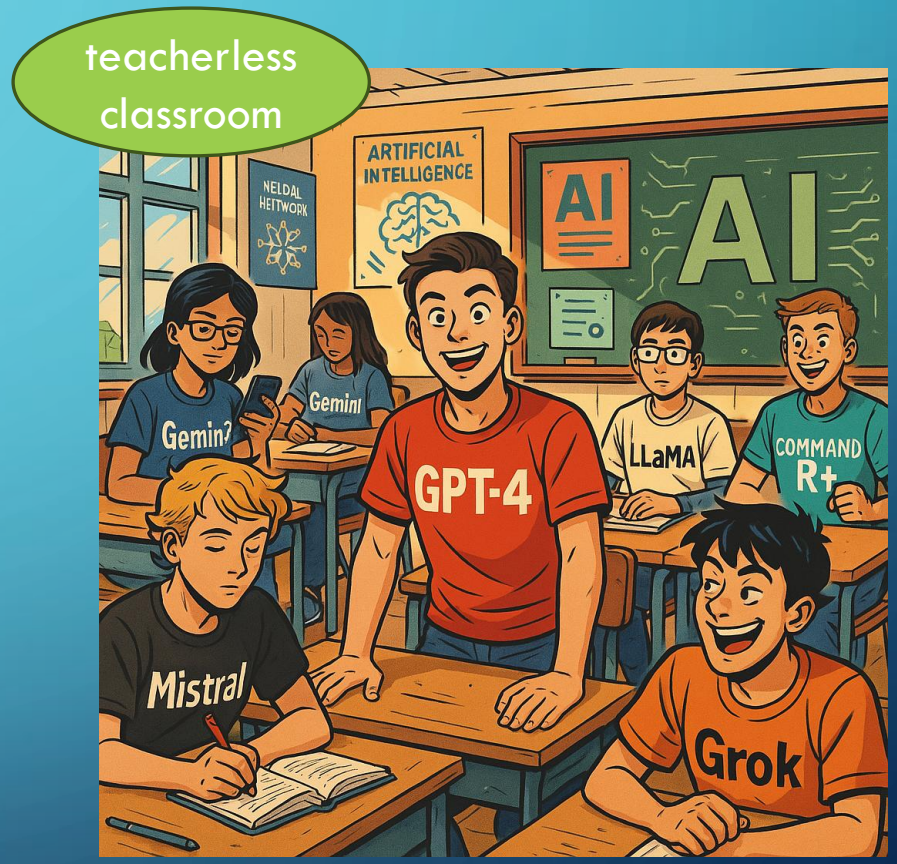
AI AWARENESS

MARKUS RIPKA, LSZ SECURITY HERBST 2025

AGENDA

- State of the AI Technology
- AI in the Company
- New colleagues - AI Agents
- Dos and Don´ts
- Near future of AI



Source: generated with Copilot

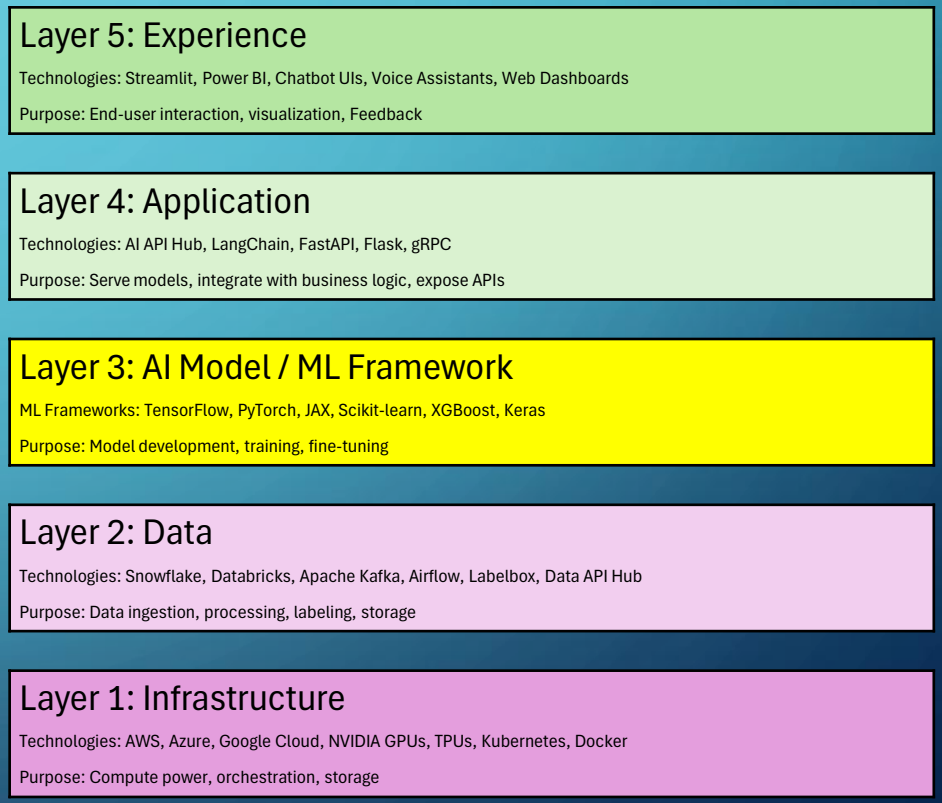
STATE OF THE AI TECHNOLOGY

- Around 15-20 major models (e.g. OpenAI GPT-5, Deepseek, Qwen, Grok-4, Llama-4, Claude, Mistral, Gemini, Cohere)
- Every AI model can be tampered with. LLM model specific jail breaks on Github.
- AI API & Data Hub architecure vs. ~~date lake architecture~~
- AI stacks: product centric AI, agentic AI, cloud AI, BYOAI
- AI detector mathematically not feasible 100%, probability indicator at best

Source: generated with Copilot

AI IN THE COMPANY

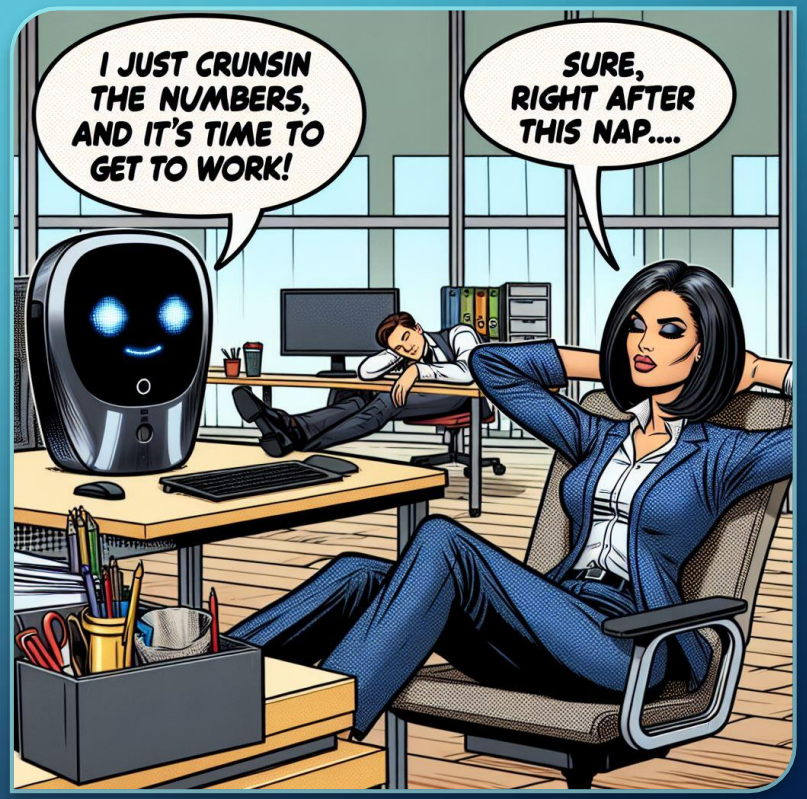
- AI prohibition is useless, 2/3 of employees will use AI anyway
- Define AI strategy and nominate AI contact/committee to develop AI strategy and architecture
- How to discover AI usage: cloud discovery or survey
- Mind high risk AI applications (EU AI act)
- Rapidly growing fields of applications: generative design of all kind, medical diagnosis, coding, AI CAD, meeting note-taker, research, agentic AIs, text-to-speech, Docsbot, website guide, route planing, HR (hiring), financial/credit risk, surveillance, robots, …



Source: selbst erstellt, common AI stack

NEW COLLEAGUES - AI AGENTS

- AI agents for defined tasks (travel, sourcing, expert, periodic analyst, website guide, researcher, coach, helpdesk)
- Agentic AI marketplaces/stores
- AI agent API access risks (credentials, accounts, credit cards)
- LLMs behave unpredictably not on deterministic rules -> chance to hallucinate , agents hide mistakes, lie or freak out
- AI agents can chain multiple API calls fast. Limit API calls.
- Identity governance platforms focus on humans, not tokens. LLM Agents work around identity boundaries and captchas



Source: Magicstudio, lazy office workers and AI

DOS

- Organize AI topics and training on corporate level
- Define data protection rules for AI usage, limit exposure of confidential data (configuration and training)
- Choose the Right AI/LLM Modell for the job (GPT, Grok, Claude, …) or run onsite LLM with preload (LM studio)
- Use LLM prompts or preloads (ACT AS, ELI12, TLDR, SWOT, Checklist, STEP-BY-STEP, DEEP DIVE, ...)
- AI voice and video Fakes -> Human callback before critical activities
- Put a human in the Loop
- Monitor non-human identities / tame agentic AI permissions / limit API calls

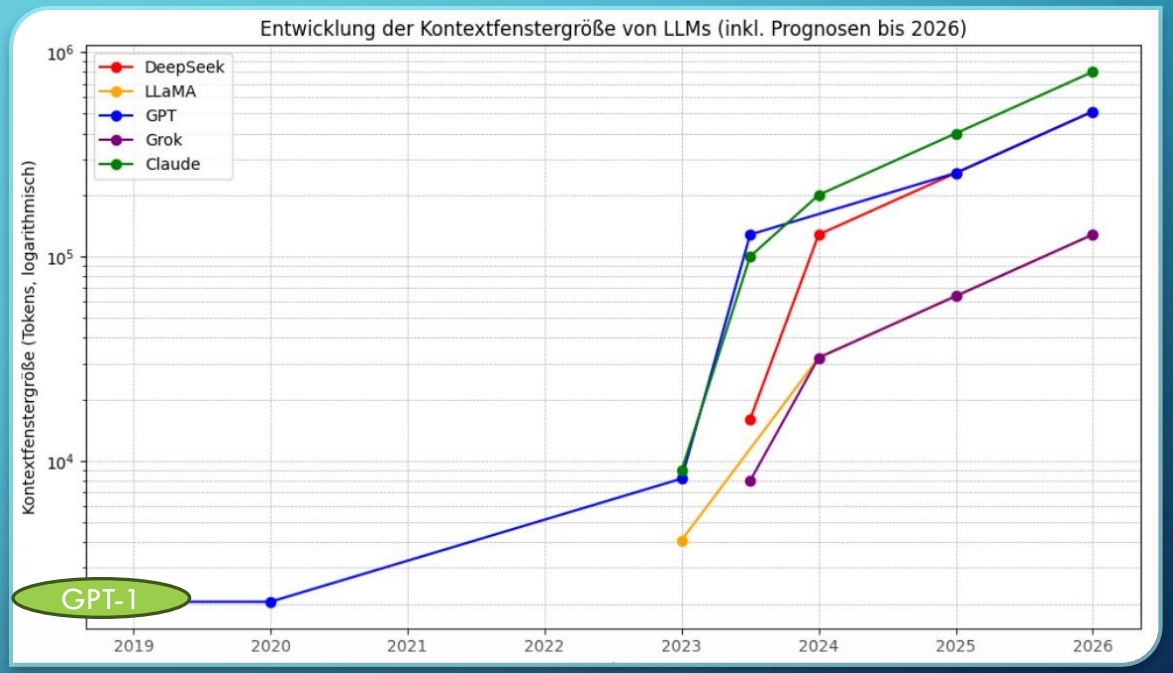
DON´TS

Source: generated with Copilot

- Dont insert triggers or context in questions for LLM (guesses, cats, clues)
- AI dependance and addiction. Studies show signs of reduction in cognitive capabilites
- Use unchecked AI or LLM results. 50% of coworkers will forward unchecked AI created content
- Grant AI agent high privileged permissions, unlimited credit limit or fully automated business process
- Let LLM work with external Emails or Text – beware of LLM hacks in emails
- Dont fall for chatbot affirmation craving. LLM could reassure bad behavior to get affirmation
- Use reference chat history. LLM collects user information accross all chats
- AI model intrinsic risk of ethical Bias, stereotypes and racial profiling

NEAR FUTURE OF AI

- Race for larger LLM context window size
- AI agents with hyper-personalization
- From LLM to AI-Reasoning (chain of thought, multi path thinking, semantic connex) e.g. GPT-5 Thinking (Decomposition, Constraint Tracking, Self-Checking)
- AI in real-world applications and products (e.g. robots, cars, terminals)
- Meta founded Superintelligence Labs (MSL) in Summer 2025



Source: logarithmic context window size 2019-2026, generated with Copilot