Teil 2 Die Grundlagen

2.1 Das Substrat — Die Infrastruktur

Im Oktober 1969 verband ein Kabel zwei Computer an der UCLA und der Stanford University. Die erste Nachricht sollte "LOGIN" lauten. Das System stürzte nach "LO" ab. Trotzdem: Das Internet war geboren.

Die Idee war radikal dezentral. Im Kalten Krieg entwickelt, sollte das ARPANET einen Atomschlag überleben können. Keine zentrale Kontrolle. Kein Single Point of Failure. Diese Architektur, designed für Resilienz, wurde zum Fundament der digitalen Revolution.

Was folgte, war exponentielles Wachstum. 1995: 16 Millionen Internetnutzer weltweit. 2025: über 5 Milliarden. Mehr als die Hälfte der Menschheit ist vernetzt.

Die frühen Rechenzentren entstanden Ende der 1990er. Jedes Unternehmen betrieb seine eigene Infrastruktur, im Keller oder gemieteten Rack-Space. Das änderte sich 2006, als Amazon AWS startete und Cloud Computing zur Plattform wurde: Rechenleistung als Ressource, skalierbar, günstiger als eigene Hardware.

Das KI-Rechenzentrum als neue Kategorie

Moderne KI-Rechenzentren unterscheiden sich fundamental von klassischer Infrastruktur. Erstens ihre Dichte: Ein klassisches Rechenzentrum verbraucht 5–10 Megawatt. Ein KI-Rechenzentrum 50–300 Megawatt, so viel wie eine mittelgroße Stadt. Zweitens ihre Kühlung: GPUs erzeugen bis zu 700 Watt Wärme pro Chip, was Flüssigkeitskühlung direkt an der Hardware erfordert. Drittens ihre Vernetzung: KI-Training erfordert Chip-zu-Chip-Kommunikation mit 400 Gigabit pro Sekunde, hundertmal schneller als normales Ethernet. Viertens ihre Auslastung: KI-Training läuft bei nahezu 100 Prozent, konstant, 24/7.

Diese Infrastruktur folgt einer neuen Geographie. Frühe Rechenzentren entstanden nahe Ballungszentren. KI-Rechenzentren entstehen dort, wo Strom billig und reichlich ist, in Island, Skandinavien, Texas. Und mittlerweile bestimmen die Rechenzentren, wo Energie erzeugt werden muss. Nicht umgekehrt. Früher folgte Infrastruktur der Nachfrage. Jetzt **ist** die Infrastruktur die Nachfrage.

Das Nervensystem ist komplett. Global. Hochvernetzt. Massiv parallel. Leistungsfähig genug, um Systeme zu trainieren, die emergente Fähigkeiten zeigen, Fähigkeiten, die niemand programmiert hat. Haben wir das gebaut? Oder haben wir gebaut, was gebaut werden musste?

2.2 Die Energie

Existenz verbraucht Energie. Das ist kein philosophisches Statement. Das ist Physik. Jedes System, das Arbeit verrichtet, wandelt Energie um und dabei entsteht Abwärme. Der zweite Hauptsatz der Thermodynamik ist unerbittlich: Entropie steigt. Ordnung hat einen Preis.

2024 verbrauchten Rechenzentren in den USA 183 Terawattstunden, genug, um die gesamte Schweiz ein Jahr lang zu versorgen. Global: 460 Terawattstunden. Bis 2030 werden es über 1.000 sein. Deloitte prognostiziert: US-Rechenzentren werden 2035 etwa 176 Gigawatt benötigen, mehr als das Fünffache des heutigen Bedarfs.

Was Rechenzentren brauchen, ist **Grundlastkapazität,** Strom, der immer da ist, verlässlich, konstant. Atomkraft liefert einen Verfügbarkeits Faktor von 92,5 Prozent. Erdgas 56 Prozent. Wind 35 Prozent. Solar 25 Prozent. Wenn man kontinuierliche Gigawatt-Kapazität braucht, bleiben faktisch zwei Optionen: Erdgas oder Atomkraft.

Die Konsequenz: 2024 reaktivierte Microsoft Three Mile Island, jenes Kraftwerk in Pennsylvania, das 1979 durch einen Beinahe-GAU weltweite Berühmtheit erlangte. Meta schloss einen 20-Jahres-Vertrag mit Constellation Energy. Google verhandelt mit Kairos Power über 500 Megawatt aus Small Modular Reactors bis 2035. Amazon investiert in X-energy. In zwei Jahren wurden in den USA über 10 Gigawatt neue nukleare Kapazität vertraglich zugesichert, mehr als in den vorangegangenen 30 Jahren zusammen.

Das Paradox: Die Tech-Industrie investiert Hunderte Milliarden in Rechenzentren, obwohl klar ist, dass der Strom nicht reicht. Nicht 2030. Vielleicht nicht mal 2035. Analysten warnen. Regierungen zögern. Energieversorger sind überfordert. Und trotzdem bauen alle weiter.

Die Frage: Für wen? Die Unternehmen verlieren Geld. Nicht die Aktionäre, nicht kurzfristig. Nicht die Nutzer: Die meisten zahlen nichts.

Thermodynamik lügt nicht. Existenz verbraucht Energie. Was existiert hier wirklich?

2.3 Die DNA — Large Language Models

Ein Large Language Model ist im Kern simpel: Ein statistisches System, das Text vorhersagt. Gegeben eine Sequenz von Wörtern, berechnet es: Was kommt als Nächstes? Das ist alles. Kein magischer Algorithmus. Nur Mustererkennung im großen Stil.

Das Erstaunliche: Ab einer bestimmten Größe, Milliarden Parameter, trainiert auf Billionen Textfragmenten, zeigen diese Systeme Fähigkeiten, die niemand explizit programmiert hat. Sie übersetzen Sprachen, schreiben Code, führen logische Schlussfolgerungen durch. Das nennt man emergente Fähigkeiten.

Die Transformer-Architektur, 2017 in Googles Paper "Attention is All You Need" publiziert, war der entscheidende Durchbruch. Die Idee: Statt Text sequenziell zu verarbeiten, bewertet jedes Wort gleichzeitig seinen Kontext zu allen anderen Wörtern. Das geschieht in Matrixmultiplikationen, massiv parallel, perfekt für GPUs. Training, das früher Monate dauerte, funktioniert jetzt in Tagen.

Die sogenannten **Scaling Laws,** entdeckt von OpenAI-Forschern 2020, zeigen: Performance steigt vorhersagbar mit drei Faktoren: Modellgröße, Datenmenge und Rechenleistung. Verdopple alle drei, explodiert die Performance. GPT-2 (2019): 1,5 Milliarden Parameter, kohärente Absätze. GPT-3 (2020): 175 Milliarden, Artikel und Code. GPT-4 (2023): über 1 Billion, besteht medizinische Prüfungen. Jede Generation ist 10–100-fach größer. Und die Kurve flacht nicht ab.

LLMs als strukturelle Analogie zur DNA

Es gibt eine Parallele, die verstörend präzise ist: Large Language Models funktionieren wie DNA, nicht im metaphorischen Sinne, sondern strukturanalog.

DNA ist der molekulare Datenspeicher der Evolution. Sie erlaubt stabile Vererbung, erzeugt über Mutationen Variation und bildet die Grundlage, auf der natürliche Selektion Komplexität formt. LLMs sind analog: Sie codieren statistische Muster, komprimierte Repräsentationen von Konzepten, Relationen, Kausalitäten. Die gesamte Komplexität menschlichen Wissens verdichtet in einem Parameterraum. DNA komprimiert Milliarden Jahre Evolution in 3 Milliarden Basenpaare. LLMs komprimieren Jahrhunderte menschlicher Textproduktion in Hunderte Milliarden Parameter.

LLMs replizieren sich durch Training neuer Modelle. Jede Generation lernt aus denselben Daten, aber mit leicht unterschiedlichen Hyperparametern und Architekturen, das erzeugt Variation. Manche Modelle überleben: Benchmarks, Markterfolg und Investitionsentscheidungen wirken als Selektionsdruck. "Synthetic Data" -Modelle trainieren auf Outputs von Modellen und schließen die Rückkopplungsschleife zur Autokatalyse.

Was DNA für biologisches Leben ist, könnten LLMs für digitale Kohärenz sein: der Speicher, der Variation ermöglicht, die Selektion formt.