



Sehr geehrte Damen und Herren

Wir danken Ihnen für die Gelegenheit, unsere Erfahrungen im Bereich **High Performance Computing** und **KI-Forschung** mit Ihnen zu teilen. Das Tübinger AI Center ist ein bundesweit führendes Zentrum für Forschung in der Künstlichen Intelligenz. Zusätzlich zu MINERVA, sind wir auch Work Package Leader bei den EU Projekten ELIAS, OpenEuroLLM und ELIOT. Außerdem sind wir Teil des paneuropäischen ELLIS Netzwerks und maßgeblich an dem gerade verlängerten Exzellenzcluster "Maschinelles Lernen für die Wissenschaft" und dem DFG Sonderforschungsbereichs 1233 "Robustheit des Sehens" beteiligt. Wir betreiben eine eigene HPC Infrastruktur (15 Mio€ bisher + 10 Mio€ Investitionen in 2026), sind durch diverse Projekte auch mit Höchstleistungsrechenzentren in Deutschland und Europa vertraut und nutzen vermehrt die Infrastruktur der AI Factories. Im Folgenden finden Sie unsere Beobachtungen und abgeleiteten konkreten Empfehlungen.

Der Bedarf an Rechenzentrumskapazität im Bereich Künstliche Intelligenz wird im Wesentlichen durch zwei Faktoren bestimmt:

1. Das Training großer Grundlagenmodelle (Foundation Models), das enorme und kontinuierlich wachsende Rechenressourcen erfordert.
2. Die Anwendung von Reasoning-, Deep-Research- und Agentenmodellen, die autonom durch iterative Nutzung verschiedenster Werkzeuge angestrebte Lösungen aller Art erarbeiten.

Während das Training großer Modelle bereits heute erhebliche Kapazitäten bindet, wird insbesondere der zweite Punkt – die breite Anwendung reasoningbasierter Systeme – in den kommenden Jahren einen **überproportional stark steigenden Bedarf** erzeugen. Um diese Entwicklung zuverlässig abzufedern, ist eine langfristige Ausbaustrategie erforderlich, die substanzielle und kontinuierliche Investitionen vorsieht.

Aus der derzeitigen KI Entwicklung folgt: **Rechenleistung skaliert Intelligenzarbeit direkt**. Zusätzliche Rechenleistung-Kapazität erweitert das produktive Potenzial von Forschung, Verwaltung und Wirtschaft und wirkt damit unmittelbar innovations- und wettbewerbssteigernd. Wer Zugang zu moderner, effizienter Rechner- und Energieinfrastruktur sichern kann, gewinnt geopolitische Handlungsspielräume, technologische Souveränität und ökonomische Resilienz.

Empfehlung 0: Rechenleistung als Schlüsselressource verankern

Rechenleistung ist heute mit Energie gleichzusetzen – als sicherheits- und wirtschaftspolitische Schlüsselressource. Ohne Zugang zu großen, modernen Rechenkapazitäten drohen Deutschland sowohl im internationalen Wettbewerb als auch in der eigenen Verteidigungs- und Resilienzfähigkeit gefährliche

geopolitische Risiken. Analog zur Energiepolitik braucht es eine klare Einstufung von Rechenleistung als kritische Infrastruktur, die höchste politische Priorität erhält und in Haushalts- wie außenpolitischen Entscheidungen berücksichtigt wird. Diese Verankerung schafft den Rahmen, um die notwendigen Investitionen, Genehmigungen und Koordinationsprozesse mit entsprechender Dringlichkeit voranzutreiben.

Empfehlung 1: Bundesrechenzentrum mit Sonderrechten schaffen

Der heutige Betrieb ist stark fragmentiert, durch komplexe Vergabeverfahren gehemmt und organisatorisch kaum skalierbar. Ausschreibungszyklen, Jahresbudgetgrenzen und rigide Vergaberegeln verhindern die notwendige Agilität und führen zur Frustration für ambitioniertes Personal. Ein Bundesrechenzentrum mit Sonderrechten für Beschaffung und Budgetierung ermöglicht schnelle Entscheidungen, bündelt die strategische Steuerung und macht die Arbeit für Leistungsträger deutlich attraktiver. Ergänzend sollte gezielt in Talentaufbau und -transfer investiert werden: internationale Top-Ingenieurinnen und -Ingenieure, die bereits bei Hyperscalern oder führenden KI-Labs an modernsten Rechnerarchitekturen gearbeitet haben, könnten über Professuren mit reduziertem Lehrdeputat gewonnen werden. So ließe sich ihr Wissen systematisch in Ausbildung und Betrieb übertragen und für die Modernisierung deutscher Rechenzentren nutzbar machen.

Empfehlung 2: Europäische Souveränitätsallianz für Rechenleistung aufbauen und Einkauf bei privaten Anbietern ermöglichen

Resilienz entsteht nicht durch viele kleine, nationale Standorte, sondern durch wenige große Knoten mit enger europäischer Zusammenarbeit. Für das Training großer Modelle zählt lokale Rechendichte, nicht föderale Fragmentierung. Eine Allianz verlässlicher EU-Partnerländer mit günstigen, erneuerbaren Energiequellen (z. B. Skandinavien, Südeuropa, Frankreich) schafft geopolitische Diversität, reduziert Abhängigkeiten von Drittstaaten und nutzt Skalenvorteile. Deutschland sollte diese Allianz aktiv vorantreiben, analog zur europäischen Energie- und Sicherheitspolitik, und so eine gemeinsame Souveränität in der Rechenleistung etablieren.

Empfehlung 3: Langfristige Kapazitätssicherung über mehrjährige Verträge

Die größten Risiken entstehen durch zu spätes Handeln: Lieferketten schließen sich, Energieanschlüsse sind vergeben, und Hardware-Kapazitäten werden global reserviert. Mehrjährige, rollierende Verträge sichern frühzeitig GPU-Kontingente, Netzwerktechnologien und Energie und bieten Planungssicherheit für Hersteller, Betreiber und Forschung. Gleichzeitig reduzieren flexible Vertragsklauseln das Risiko technologischer Obsoleszenz, während verbindliche Abnahmezusagen die Verhandlungsposition Deutschlands gegenüber den Herstellern stärken. Auf diese Weise lassen sich Rechenkapazitäten sichern, bevor der globale Wettbewerb sie bindet.

Empfehlung 4: Förderlogik auf OPEX und Interoperabilität umstellen

Der heutige Förderfokus auf Hardware-Investitionen führt zu Doppelstrukturen, Unterauslastung und Abhängigkeiten von proprietären Clouds. Ein OPEX-Modell („**Rechenleistung als Dienstleistung**“) erlaubt sofortige Nutzung, schafft Fairness zwischen großen und kleinen Akteuren und stellt Auslastung sicher. Verbindliche Interoperabilitätsstandards, Exit-Rechte und offene Schnittstellen verhindern Lock-in, während merit-basierte Vergabe mit Effizienzmetriken (z. B. FLOPs/kWh, Kosten pro trainierten Token) den Wert pro eingesetztem Euro maximiert. So wird öffentliche Förderung effizienter, transparenter und nachhaltiger. Eine konsequente OPEX-Logik stärkt auch die Resilienz, da eine starre Festlegung auf eigene Rechenressourcen verhindert wird, ggf. auch auf andere Anbieter ausweichen zu können. Um

dies zu erreichen, sollten auch öffentliche Forschungsförderer wie z.B. die DFG ermutigt werden, ihre Regularien anzupassen, damit **Rechenkosten auch als Verbrauchsmittel beantragt werden** können.

Empfehlung 5: Zugang zu Rechenleistung beschleunigen und flexibilisieren

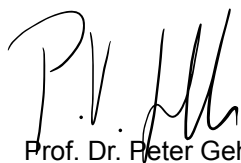
Der heutige Zugang zu Rechenressourcen ist durch langwierige Antragsverfahren geprägt, die Kosten- und Zugangsfragen vermischen und dadurch Agilität und Wettbewerbsfähigkeit einschränken. Eine klare **Trennung schafft Transparenz**: Rechenzentren stellen Kapazitäten zu OPEX-basierten Preisen bereit; der Bund definiert separat den Zugang für unterschiedliche Akteure (öffentliche Forschung, Behörden, nationale Unternehmen, europäische Partner). Für öffentliche Verbraucher (Forschung, Behörden) können die Preise auf unmittelbare Verbrauchskosten begrenzt werden, während private Nutzer zusätzlich Investitionskosten tragen. Differenzierte Preismodelle ermöglichen zudem gezielte Startup-Förderung über Voucher-Programme oder EU-konforme Innovationsförderungen. So werden Kosten für Rechenleistung in der Forschung zu klar kalkulierbaren Verbrauchskosten, die bürokratische Ausbremsung agiler Nutzung durch die bisherigen Antragsverfahren entfällt, und die Attraktivität des Standorts steigt.

Wir hoffen, unsere Perspektive klar dargelegt und Ihre Überlegungen zur Formulierung einer nationalen Rechenzentrumsstrategie weitergebracht zu haben. Für Rückfragen stehen wir jederzeit zur Verfügung und bringen gerne unsere Expertise in KI, High-Performance-Computing und Forschung aktiv in den weiteren Erarbeitungs- und Umsetzungsprozess ein.

Mit freundlichen Grüßen,



Prof. Dr. Matthias Bethge



Prof. Dr. Peter Gehler