gym | NEUFELD
Eine Institution des Kantons Bern

Schriftliche Arbeit

Wie beeinflusst NVIDIAs Marktdominanz den Wettbewerb in der KI-Halbleiterindustrie?

Quelle: Foto von Mariia Shalabaieva (Unsplash)
https://unsplash.com/photos/the-nvidia-logo-is-displayed-on-a-table-0SqsTxWhgNU

Jonathan Mark Oliver Gerbig | Klasse 26Wd

Eine Maturaarbeit im Fach Wirtschaft und Recht

Gymnasium Neufeld — Abteilung Wirtschaft und Recht

Arbeit vom 14.03.2026

Betreuende Lehrperson: Herr Alexander Stämpfli

Abstract

Hintergrund: NVIDIA dominiert den Markt für KI-Halbleiter mit Marktanteilen von über 90 %. Diese Arbeit untersucht, wie sich diese Marktdominanz auf den Wettbewerb auswirkt.

Methodik: Die Analyse erfolgt dreistufig: (1) Quantifizierung der Marktkonzentration mittels Herfindahl-Hirschman-Index (HHI), (2) Untersuchung des Lock-in-Effekts durch NVIDIAs proprietäre CUDA-Plattform, (3) Bewertung der Wettbewerbskräfte nach Porters Five-Forces-Modell mit Fokus auf Rivalität und Substitute.

Ergebnisse: Der HHI-Wert von 8 496 belegt eine nahezu monopolistische Marktstruktur. Die qualitative Analyse zeigt, dass NVIDIAs CUDA-Plattform durch hohe Wechselkosten und ein etabliertes Entwickler-Ökosystem erhebliche Markteintrittsbarrieren schafft. Gleichzeitig offenbart die Five-Forces-Analyse trotz der Marktdominanz eine hohe Wettbewerbsintensität, insbesondere durch die Rivalität mit AMD und die Bedrohung durch Eigenentwicklungen grosser Cloud-Anbieter (Hyperscaler).

Schlussfolgerung: Die Ergebnisse zeigen ein Paradoxon: NVIDIAs Marktmacht ermöglicht massive Investitionen in Forschung und Entwicklung, die den technologischen Fortschritt beschleunigen, schränkt jedoch gleichzeitig den Wettbewerb ein. Die Arbeit schliesst mit einer Empfehlung für eine ausgewogene Regulierung, die Innovation fördert und gleichzeitig faire Marktbedingungen sicherstellt.

1. Vorwort

Die Themenwahl für meine Maturaarbeit war aufgrund meines langjährigen Interesses an Technologie und Künstlicher Intelligenz naheliegend. Die Schwierigkeit bestand darin, meine Ideen so einzugrenzen, dass sie den vorgegebenen Umfang nicht überschreiten. Seit Jahren beschäftige ich mich intensiv mit Technologie und interessiere mich für ein breites Feld der Informatik. Was viele meiner Interessen inzwischen gemeinsam haben, ist die KI (Künstliche Intelligenz). Meine Auseinandersetzung mit KI geht über die gängige Nutzung von ChatGPT hinaus: Ich arbeite mit verschiedenen Modellen unterschiedlicher Anbieter und betreibe auch lokale KI-Systeme auf meinem eigenen Computer, die ohne Internetverbindung funktionieren. Auch halb- und vollautomatische Agenten, denen man ein Ziel vorgeben kann und die dieses dann selbstständig erfüllen, setze ich ein. Zum Beispiel nutze ich Programmier-Agenten wie Replit oder Lovable, die mich bei der Entwicklung neuer Projekte unterstützen. Intensiv verfolge ich die neuesten Entwicklungen in diesem sich rasant entwickelnden Bereich.

Die technologische Grundlage nahezu aller modernen KI-Anwendungen bilden die Entwicklungen der Firma NVIDIA. Direkt durch ihre Hardware und indirekt durch ihr Software-Ökosystem – NVIDIA hat sich zu einem derart wichtigen Akteur in der Technologiebranche entwickelt, dass zahlreiche Innovationen ohne ihre Produkte undenkbar wären.

Um mein Wissen für diese Arbeit zu vertiefen, habe ich 2 671 Seiten relevante Literatur in Originalsprache, sowie sämtliche im Literaturverzeichnis aufgeführten Publikationen und Reports gelesen und studiert. Unter anderem «The Nvidia Way» (Kim, 2024), das NVIDIAs Geschichte und die besondere Arbeitskultur ausführlich aufzeigt, «Chip War» (Miller, 2022), das einen Einblick in die Politik und Schwierigkeiten hinter der Industrie gibt, «The Thinking Machine» (Witt, 2025), das sich mit der technologischen Entwicklung der Halbleiterbranche befasst, und «Empire of AI» (Hao, 2025) über einen der grössten Kunden NVIDIAs: OpenAI (die Firma hinter ChatGPT).

Ich danke Herrn Alexander Stämpfli für die fachkundige Begleitung, die konstruktiven Anmerkungen und die kontinuierliche Unterstützung.

Ebenfalls danke ich meiner Familie für Rückmeldungen und Korrekturhinweise.

Als Hilfsmittel nutzte ich wissenschaftliche Datenbanken, Marktreports und gängige Analysetools. Auch mehrere KI-Modelle wurden im Rahmen des Erlaubten genutzt und halfen bei Recherche und Textkorrektur.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

2. Einleitung

Ohne die Halbleiterindustrie wären wir als Menschheit technologisch nicht an dem Punkt, an dem wir heute sind (Weinzierl, 2024). Halbleiter sind heute Grundlage für einen beträchtlichen Teil unseres alltäglichen Lebens. Jedes Handy, jeder Zug und jede automatische Tür benötigen einen oder sogar zahlreiche Chips, die auf Halbleiter basieren (Krieser, 2023). Einfach gesagt: Jedes Gerät, das mit Strom funktioniert, braucht Halbleiter (Kiefer, 2023). Als es während der Corona-Pandemie Lieferengpässe von Halbleitern gab, war viel Elektronisches nicht lieferbar oder wurde teurer (Haramboure et al., 2023).

Diese Arbeit beschäftigt sich vor allem mit den Halbleitern von NVIDIA im Zusammenhang mit der KI. Jeder Chip der in dieser Arbeit erwähnt wird, basiert auf Halbleitern. Das ist die kleinste Art eines Schalters und bildet die Grundlage für die ganze moderne Elektronik. Obwohl Künstliche Intelligenz bereits seit Jahrzehnten erforscht wird und an Universitäten als etabliertes Fachgebiet gilt (Zech, 2022), erreichte sie erst mit der Veröffentlichung von ChatGPT durch OpenAI im November 2022 die breite Öffentlichkeit. Die darauf folgenden Medienwellen und die rasante Adaption machten KI innerhalb kürzester Zeit zu einem integralen Bestandteil des Alltags vieler Menschen (Mayer et al., 2025). NVIDIA, das per September 2025 zu den wertvollsten Unternehmen der Welt zählt, hat sich als zentraler Akteur in dieser Entwicklung etabliert. Das Unternehmen produziert die essenzielle Hardware für KI-Anwendungen: GPUs (Graphics Processing Units) für das Training von KI-Modellen und spezialisierte Beschleuniger-Chips, die den technologischen Fortschritt in diesem Bereich massgeblich vorantreiben.

Eine Erklärung der verwendeten Abkürzungen befindet sich im Glossar hinten in der Arbeit.

2.1. Fragestellung, Ziele und Hypothesen

2.1.1. Zentrale Fragestellung

Zentrale Forschungsfrage

Wie beeinflusst NVIDIAs Marktdominanz den Wettbewerb in der KI-Halbleiterindustrie?

2.1.2. Ziel der Arbeit

1. Marktkonzentration messen: Quantifizierung von NVIDIAs Marktmacht im KI-Chipsektor mittels HHI
2. CUDA[1] als Wettbewerbsfaktor analysieren: Untersuchung, wie NVIDIAs CUDA-Plattform als technologische Abhängigkeit den Wettbewerb beeinflusst (Lock-in Effekt)

[1] Erklärung folgt in Abschnitt 3.2.3

3. Wettbewerbskräfte bewerten: Analyse der Wettbewerbssituation im KI-Chipsektor mithilfe von Porters Five Forces, fokussiert auf die Rivalität zwischen Wettbewerbsteilnehmer und die Bedrohung durch Substitute

2.1.3. Meine Hypothesen

1. NVIDIAs hoher Marktanteil im KI-Chipsektor führt zu einer Marktkonzentration, die mittels HHI-Index als hoch einzustufen ist

2. Das CUDA-Ökosystem stellt eine signifikante technologische Eintrittsbarriere dar und schränkt den Wettbewerb durch hohe Wechselkosten für Kunden ein

3. Trotz NVIDIAs Dominanz ist die Rivalität im KI-Chipsektor hoch, insbesondere durch direkte Konkurrenten wie AMD und die Bedrohung durch Eigenentwicklungen von Grosskunden (Substitute)

2.2. Methodisches Vorgehen

Um die zentrale Fragestellung zu beantworten und meine Hypothesen zu beweisen oder zu widerlegen, betreibe ich Literaturrecherche, sammle Marktdaten und analysiere diese anhand von passenden Modellen. Die Arbeit ist in einen Theorieteil und in einen Praxisteil gegliedert. Im Theorieteil werde ich alle relevanten Informationen aufzeigen, um die nötigen Grundlagen für die daraus folgenden Analysen im Praxisteil herzuleiten. Die gewonnenen Daten werden durch Diagramme und Tabellen visualisiert, um eine transparente und nachvollziehbare Darstellung der Ergebnisse zu gewährleisten. Damit überprüfe ich die aufgestellten Hypothesen und beantworte die zentrale Fragestellung schlüssig.

3. Grundwissen

3.1. Marktdominanz und Marktkonzentration

3.1.1. Marktdominanz in Technologiemärkten

Marktdominanz bezeichnet eine Marktposition, in der ein Unternehmen wesentlichen Einfluss auf das Marktgeschehen ausübt. Das schweizerische Kartellgesetz («SR 251 - Bundesgesetz vom 6. Oktober 1995 über Kartelle u...» n. d.) definiert die Marktdominanz als die Möglichkeit eines Unternehmens, sich unabhängig von anderen Wettbewerbern, Abnehmern oder Lieferanten zu verhalten. Die 2022 eingeführte relative Marktmacht (Weko, 2025) erweitert dieses Konzept auf Situationen, in denen andere Unternehmen derart abhängig sind, dass keine ausreichenden Ausweichmöglichkeiten bestehen.

Technologiemärkte sind besonders anfällig für alleinige Marktdominanz, da sie strukturelle Eigenschaften aufweisen, die Winner-takes-all-Dynamiken (Hoffmann, 2020) fördern. Netzwerkeffekte verstärken diese Tendenz, indem der Nutzen eines Produkts mit der Anzahl seiner

Nutzer steigt. Zusätzlich führen hohe Entwicklungskosten, Skaleneffekte und starke Wechselkosten zu natürlichen Barrieren für neue Marktteilnehmer, wodurch etablierte Unternehmen ihre dominante Position festigen können. Im Buch «Chip War» beschreibt Miller, wie solche Dynamiken geopolitische Konflikte schüren, beispielsweise US-Exportkontrollen gegen China (Miller, 2022).

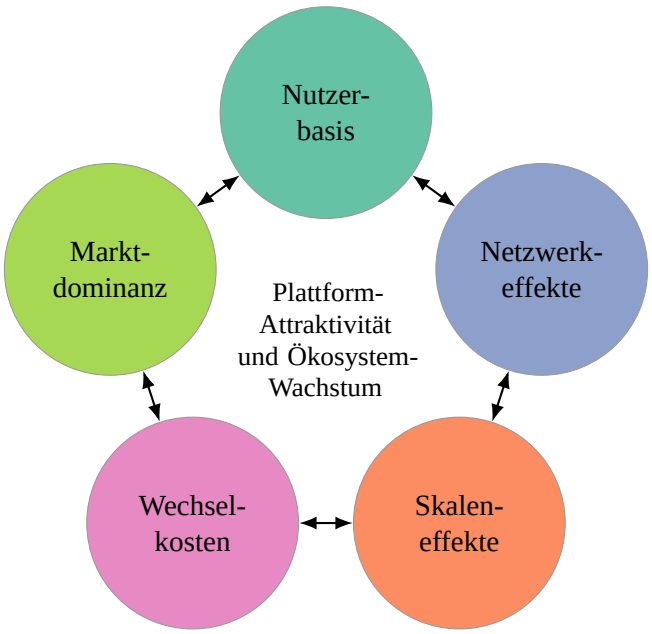

Abbildung 1: Marktdominanz in Technologiemärkten: Verstärkungskreis aus Nutzerbasis, Netzwerkeffekten, Skaleneffekten und Wechselkosten

Quelle: Eigene Darstellung in Anlehnung an (Dexheimer & Lechner, 2019; Hoffmann, 2020; Krämer, 2019).

3.1.2. Der Herfindahl-Hirschman-Index als Messinstrument

Zur Konzentrationsmessung in einem bestimmten Markt kommt oft ein von Herfindahl und Hirschman entwickeltes Messinstrument zum Einsatz. Dieser sogenannte Herfindahl-Hirschman-Index (HHI) quantifiziert Marktkonzentration durch die Summe der quadrierten Marktanteile aller Unternehmen.

HHI Berechnung

$$\mathrm{HHI} = \sum_{i=1}^{N} s_i^2 \tag{1}$$

Quelle: Eigene Darstellung in Anlehnung an (Bromberg, 2024).

Legende:

N	Anzahl der Unternehmen im Markt
s_i	Marktanteil von Unternehmen i in Prozentpunkten
HHI	Herfindahl–Hirschman-Index als Summe der quadrierten Anteile (Wertebereich: 0 bis 10 000 bei Prozentangaben).

Was kompliziert klingt, ist es gar nicht: Die Berechnung erfasst sowohl die Anzahl der Wettbewerber als auch deren relative Grösse, wodurch der Index Werte zwischen 0 (perfekter Wettbewerb) und 10 000 (Monopol) annehmen kann. Die Interpretation erfolgt anhand etablierter Schwellenwerte: Werte unter 1 000 signalisieren geringe Konzentration, 1 000–1 800 mittlere Konzentration und über 1 800 hohe Konzentration.

Die schweizerische Wettbewerbskommission (WEKO) wendet den HHI bei Fusionskontrollen an, wobei Zusammenschlüsse in hochkonzentrierten Märkten (HHI > 1 800) mit einer Erhöhung um mehr als 150 Punkte eine vertiefte Prüfung erfordern (Weko, 2025).

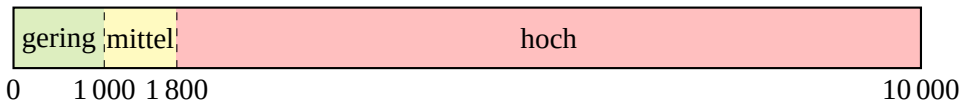

Abbildung 2: Interpretation des HHI: Schwellenwerte und Einordnung

Quelle: Eigene Darstellung in Anlehnung an (Weko, 2025); vgl. auch (Bromberg, 2024).

3.2. Technologische Lock-in-Effekte am Beispiel CUDA

3.2.1. Theoretische Grundlagen des Lock-in-Effekts

Lock-in-Effekte entstehen, wenn Nutzer bei einer bestimmten Technologie bleiben, weil ein Wechsel zu teuer oder aufwändig wäre. Dieses Phänomen tritt besonders häufig in der Technologiebranche auf, wo Unternehmen oft jahrelang in spezifische Systeme investiert haben. Die Kosten eines Wechsels umfassen nicht nur neue Hardware und Software, sondern auch die Zeit für Umschulungen und mögliche Produktivitätsverluste während der Umstellung. Je länger ein Unternehmen eine bestimmte Technologie nutzt, desto stärker wird die Bindung. Mitarbeiter entwickeln Expertenwissen, Prozesse werden auf die Technologie abgestimmt und die gesamte IT-Infrastruktur richtet sich danach aus. Diese gewachsenen Strukturen machen einen Wechsel immer schwieriger, selbst wenn bessere Alternativen verfügbar sind (Witt, 2025).

3.2.2. Netzwerkeffekte als Verstärker

Netzwerkeffekte verstärken Lock-in-Situationen erheblich. Sie entstehen, wenn eine Technologie umso wertvoller wird, je mehr Menschen sie nutzen. Bei Software-Plattformen führt eine grosse Nutzerzahl zu mehr verfügbaren Programmen, besserer Dokumentation und aktiveren Online-Communities. Diese Vorteile ziehen neue Nutzer an, wodurch die Entwickler-Plattform weiterwächst. Für neue Konkurrenten wird es dadurch schwierig, sich am Markt zu etablieren. Selbst wenn ihre Technologie besser wäre, fehlen ihnen die Nutzer und damit das Ökosystem aus Software und Support. Etablierte Plattformen können ihre Position so über Jahre hinweg verteidigen (Grasser, 2018).

3.2.3. CUDA als Beispiel für technologische Abhängigkeit

CUDA (Compute Unified Device Architecture) ist eine von NVIDIA entwickelte Technologie, die es ermöglicht, Grafikkarten (GPUs) für allgemeine Berechnungen zu nutzen. Sie wurde im Jahr 2006 vorgestellt und seither kontinuierlich weiterentwickelt. Vor CUDA wurden GPUs hauptsächlich für die Darstellung von Grafiken verwendet. NVIDIA erkannte jedoch, dass die parallele Rechenleistung von GPUs auch für andere Aufgaben, wie wissenschaftliche Simulationen oder Entwicklung und das Betreiben künstlicher Intelligenzen, genutzt werden könnte (Kim, 2024).

Der entscheidende Durchbruch von CUDA war die einfache Programmierbarkeit. Entwickler konnten in bekannten Programmiersprachen wie C++ arbeiten, ohne sich mit komplexer Grafikprogrammierung auseinandersetzen zu müssen. Diese Zugänglichkeit führte zu einer schnellen Verbreitung der Technologie in Universitäten und Unternehmen weltweit (Kim, 2024).

3.2.4. Die Entstehung des CUDA Lock-ins

Der Lock-in-Effekt bei CUDA ist kein Zufall, sondern das Ergebnis einer über 15 Jahre andauernden strategischen Entwicklung. Um die heutige Situation zu verstehen, muss man die technologische Landschaft vor CUDA betrachten. Damals war «General Purpose GPU Computing» (GPGPU) ein komplexes Feld für wenige Experten. Forscher mussten die Grafik-Pipeline (eine Art, um Grafikkarten zu zeigen, was sie darzustellen haben) «hacken» und ihre Berechnungen in Form von Grafik-Shadern (kleine Programme zur Pixelberechnung) formulieren – ein extrem komplexer und nicht-intuitiver Prozess (Bennett, 2023).

NVIDIAs Geniestreich mit der Einführung von CUDA im Jahr 2006 bestand darin, diese Komplexität zu abstrahieren. Plötzlich konnten Millionen von Entwicklern, die mit der weit verbreiteten Programmiersprache C (später C++) vertraut waren, die massive Parallelverarbeitungsleistung von GPUs für wissenschaftliche und mathematische Berechnungen nutzen, ohne Experten für Computergrafik sein zu müssen (Kim, 2024).

Diese Entwicklung verlief parallel zu den Fortschritten im Bereich der künstlichen neuronalen Netze, die genau diese Art von massiv-parallelen Matrixmultiplikationen erfordern, für die GPUs perfekt geeignet sind. Der Wendepunkt kam 2012, als ein neuronales Netz namens AlexNet, das auf zwei NVIDIA GTX 580 GPUs trainiert wurde, den renommierten ImageNet-Wettbewerb deklassierte. Dieser «Urknall» des modernen Deep Learning bewies der Welt, dass GPU-

beschleunigtes Training der Schlüssel zu bahnbrechenden KI-Fortschritten ist (Schmidhuber, 2022; Witt, 2025).

Zeitleiste der Entwicklung von CUDA

2006	NVIDIA stellt CUDA 1.0 vor. Die GPU wird zu mehr als nur einem Grafikbeschleuniger.
2007	Einführung der ersten Tesla-GPU (Name dieser Grafikkarte), die explizit für Rechenzentren und wissenschaftliches Rechnen konzipiert ist.
2012	AlexNet gewinnt den ImageNet-Wettbewerb auf NVIDIA-GPUs und löst die Deep-Learning-Revolution aus.
2014–2015	Google veröffentlicht TensorFlow, Facebook (Meta) veröffentlicht PyTorch. Beide grossen KI-Frameworks werden von Anfang an mit starkem Fokus auf CUDA entwickelt.
2017	Die Volta-Architektur führt «Tensor Cores» ein – spezialisierte Recheneinheiten, die KI-Operationen massiv beschleunigen und den Vorsprung von NVIDIA weiter ausbauen.
2020–2024	Die Architekturen Ampere (A100), Hopper (H100) und Blackwell (B200) verzehnfachen jeweils die Leistung für KI-Anwendungen und zementieren die Hardware-Dominanz. Diese betreiben die Modelle, die wir heute nutzen.

4. Modell zur Analyse der Wettbewerbskräfte

4.1. Einordnung und Zielsetzung des Five-Forces-Modells

Das Five-Forces-Modell von Michael E. Porter wurde 1979 mit dem Ziel entwickelt, den Einfluss branchenspezifischer Strukturmerkmale auf das Gewinnpotenzial von Unternehmen systematisch zu erfassen (Porter, 2008). Es knüpft an das «Structure-Conduct-Performance-Paradigma» der Industrieökonomik an, das einen Kausalzusammenhang zwischen Marktstruktur, Unternehmensverhalten und Marktergebnissen aufzeigt (Goyal, 2020). Während Mason und Bain in der klassischen Industrieökonomik vor allem die makroökonomische Perspektive betonten, überträgt Porter diese Logik in einen strategischen Bezugsrahmen, der einzelnen Unternehmen erlaubt, ihre Position im Wettbewerb zu bewerten und gezielt zu gestalten.

4.2. Die fünf Kräfte und ihr Zusammenspiel

Im Zentrum des Modells stehen fünf externe Kräfte, deren Zusammenspiel die langfristige Rentabilität einer Branche bestimmt:

Erstens kann die Bedrohung durch neue Marktteilnehmer das Branchengewinnniveau drücken, wenn niedrige Eintrittsbarrieren wie geringe Kapitalanforderungen oder schwache Regulierung den Zugang erleichtern.

Zweitens beeinflusst die Verhandlungsmacht der Lieferanten die Gewinnspanne, insbesondere dann, wenn wenige Anbieter über knappe oder stark differenzierte Produktionsfaktoren verfügen, und die Wechselkosten der Abnehmer hoch sind.

Drittens üben die Abnehmer Macht aus, sobald Produkte standardisiert sind, grosse Abnahmemengen gebündelt werden oder alternative Bezugsquellen leicht verfügbar sind («The Five Forces - Institute For Strategy And Competitiveness - Harvard Business School», n. d.).

Viertens setzt die Bedrohung durch Ersatzprodukte eine natürliche Preisobergrenze, weil funktionale Substitute mit günstigem Preis-Leistungs-Verhältnis oder niedrigen Umstellungskosten Kunden abwandern lassen (Bruijl, 2018).

Fünftens bestimmt die Rivalität unter den bestehenden Wettbewerbern – etwa in Form von Preis- und Innovationswettbewerb – das Ausmass, in dem Unternehmen ihre Gewinne verteidigen können. Hohe Fixkosten, geringes Branchenwachstum oder geringe Produktdifferenzierung verstärken diese Rivalität und schmälern das Gewinnpotenzial.

Die fünf Kräfte wirken nicht isoliert, sondern interdependent. Eine technologische Innovation, die etwa die Eintrittsbarrieren senkt, kann gleichzeitig die Rivalität erhöhen und die Verhandlungsmacht der Abnehmer stärken. Porter betont daher die Notwendigkeit einer ganzheitlichen Betrachtung des Branchengefüges. Für die strategische Praxis bietet das Modell einen klar strukturierten Rahmen, mit dem sich sowohl Eintritts- und Austrittsentscheidungen, als auch Kooperations- oder Diversifikationsstrategien, ableiten lassen (Porter, 2008).

Trotzdem unterliegt das Modell wesentlichen Einschränkungen. Kritisieren kann man seine statische Ausrichtung, weil dynamische Phänomene wie technologische Disruption, Plattformökonomien (beispielsweise Digitec) oder sich schnell wandelnde Kundenpräferenzen (Trends) nur bedingt erfasst werden. Darüber hinaus blendet Porter Komplementoren – also Akteure, deren Produkte den Wert des eigenen Angebots erhöhen – weitgehend aus und vernachlässigt damit Netzwerk- und Ökosystemeffekte, die insbesondere in digitalen Märkten zentral sind (Tan, 2014). Schliesslich fokussiert das Modell ausschliesslich auf externe Branchenfaktoren und integriert interne Ressourcen- und Kompetenzvorteile, wie sie die «Resource-Based-View» betont, lediglich implizit.

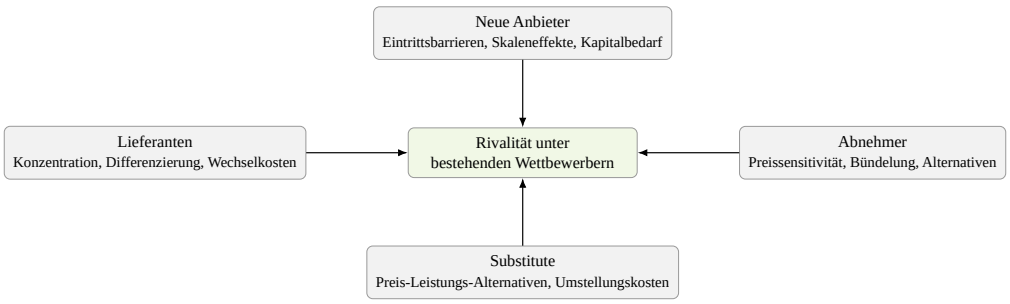

Abbildung 3: Porters Five Forces: Theoretischer Bezugsrahmen

Quelle: Eigene Darstellung in Anlehnung an (Bruijl, 2018; Porter, 2008).

4.3. Grenzen und Weiterentwicklung des Ansatzes

Trotz dieser Limitierungen behält das Five-Forces-Modell seine fundamentale Bedeutung als Analyseinstrument. Forscher wie E. Dobbs (E. Dobbs, 2014) betonen, dass die Einfachheit und Klarheit des Frameworks seine breite Anwendbarkeit sichert. Es ist eine Art gemeinsame Sprache für strategische Diskussionen und erleichtert die Diskussionen zwischen verschiedenen Organisationsebenen.

5. Anwendung und Analyse: NVIDIAs Position im KI-Ökosystem

5.1. Methodenbeschreibung

Um die Forschungsfrage zu beantworten und meine Hypothesen zu prüfen, nutze ich in dieser Arbeit verschiedene Methoden. Ich vertiefe mich in aktueller Fachliteratur und werte die zusammengetragenen Daten zum KI-Halbleitermarkt und zu den relevanten Unternehmen aus.

Erstens berechne ich, wie stark sich der Markt für KI-Chips auf wenige Firmen konzentriert. Dafür verwende ich den Herfindahl-Hirschman-Index (HHI) (Bromberg, 2024). Als Grundlage dienen mir aktuelle Zahlen (Fernandez, 2025; «NVIDIA Announces Financial Results for Fourth Quarter and Fiscal 2025», 2025) zu den Marktanteilen von NVIDIA und seinen wichtigsten Konkurrenten. Mit dieser Berechnung überprüfe ich Hypothese 1, die aussagt, dass der Markt stark konzentriert ist.

Zweitens untersuche ich den sogenannten «Lock-in-Effekt» (Hase, 2019) durch NVIDIAs Software-Plattform CUDA («CUDA – Thomas-Krenn-Wiki», n. d.) genauer. Ich analysiere Texte und Berichte, um herauszufinden, wie sehr Kunden an CUDA gebunden sind und wie schwierig oder teuer ein Wechsel zu anderen Systemen wäre (Center for Security and Emerging Technology et al., 2020). Damit prüfe ich Hypothese 2, wonach CUDA es anderen Firmen schwer macht, in den Markt einzusteigen oder zu konkurrieren.

Drittens schaue ich mir den Wettbewerb in der Branche mit dem bekannten Fünf-Kräfte-Modell von (Porter, 2008) an. Damit die Arbeit nicht zu umfangreich wird, konzentriere ich mich auf die zwei wichtigsten Kräfte für diese Untersuchung: die Rivalität zwischen den bestehenden Firmen (also wie stark NVIDIA, AMD und Intel miteinander konkurrieren) und auf die Bedrohung durch Ersatzprodukte (zum Beispiel, wenn grosse Kunden wie Google oder Amazon eigene Chips entwickeln). Damit überprüfe ich Hypothese 3 zur Stärke des Wettbewerbs.

5.2. Resultate

In diesem Kapitel werden die in der Methodik beschriebenen Analysen durchgeführt, um die aufgestellten Hypothesen zu überprüfen.

5.2.1. Hypothese 1: Hohe Marktkonzentration gemessen am HHI

In der ersten Hypothese behaupte ich, dass NVIDIAs hoher Marktanteil zu einer Marktkonzentration führt, die mittels HHI-Index als hoch einzustufen ist. Für die Berechnung werden die Marktanteile im Markt für KI-Chips in Rechenzentren herangezogen, da diese Chips die primäre Hardware für das Training von KI-Modellen darstellen.

Gemäss Analystenberichten von Ende 2023 und Anfang 2024 kontrollierte NVIDIA rund 92 % dieses Marktes. Der Hauptkonkurrent AMD kam auf etwa 4 %, während Intel und andere Hersteller die restlichen 4 % unter sich aufteilten (Fernandez, 2025).

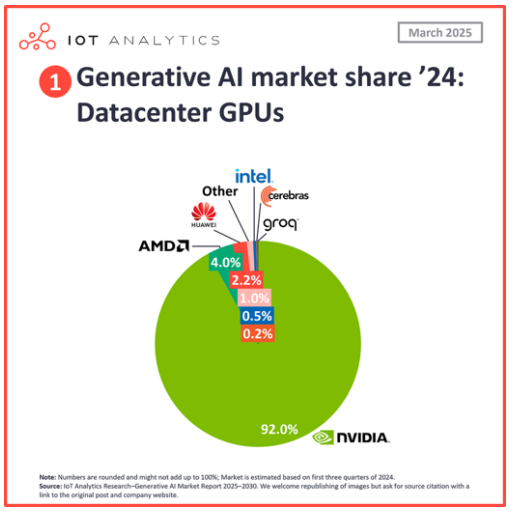

Abbildung 4: Marktanteile im Markt für KI-Chips in Rechenzentren (IoT Analytics)

Quelle: IoT Analytics, *Generative AI market share '24: Datacenter GPUs* (Fernandez, 2025). Bild: https://iot-analytics.com/wp-content/uploads/2025/03/GenAI-market-share-2024-Datacenter-GPUs-vw.png. Artikel: https://iot-analytics.com/leading-generative-ai-companies/. *Übernommen.*

In der folgenden Grafik sieht man den Berechnungsweg des HHI im KI-Chipmarkt anhand der Daten von IOT Analytics:

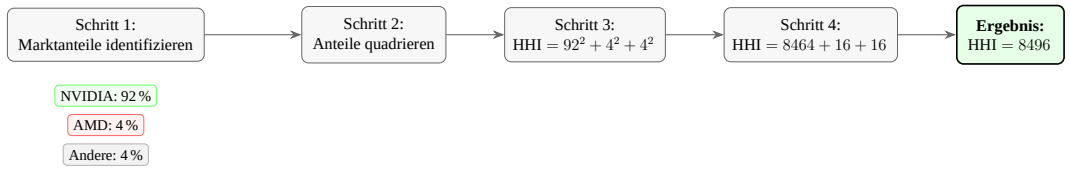

Abbildung 5: Berechnung des Herfindahl–Hirschman-Index (HHI) für KI-Chips in Rechenzentren (2024)

Hinweis: HHI $= \sum s_i^2$ mit Marktanteilen in Prozentpunkten. **Quelle:** Eigene Darstellung nach (Fernandez, 2025). Details siehe Tab. 5 im Anhang.

Zur transparenten Nachvollziehbarkeit zeigt Tabelle 5 im Anhang die einzelnen Anteile und die resultierende Summe der quadrierten Marktanteile (HHI).

Ein HHI-Wert von 8 496 liegt weit über dem Schwellenwert von 1 800, den die schweizerische Wettbewerbskommission (Weko, 2025) und andere Kartellbehörden zur Definition eines hochkonzentrierten Marktes verwenden. Ein solcher Wert nähert sich bereits dem theoretischen Maximum eines Monopols (10 000) an.

Zusätzlich habe ich den HHI für den Zeitraum 2022–2025 berechnet. Die Werte bleiben durchgehend im hochkonzentrierten Bereich (2022: 9 038, 2023: 9 606, 2024: 8 496, 2025: 8 528); Details stehen in Tabelle 6 im Anhang.

Um ein besseres Verständnis zu erlangen, wie riesig dieses Resultat ist, sind hier noch die Berechnungen anderer Branchen mit starker Marktführung:

Tabelle 1: Vergleichende HHI-Werte ausgewählter Märkte (2023–2025)

Branche/Markt	HHI	Anmerkung/Quelle
KI-Chips in Rechenzentren	8496	Vgl. (Fernandez, 2025).
Erfrischungsgetränke (Schweiz)	2250	Coca-Cola (≈ 25%), Pepsi (≈ 10%), Rivella (≈ 15%), Private Labels (≈ 20%), Rest (≈ 30%). Quelle: Research & Markets https://www.researchandmarkets.com/reports/5917846/switzerland-carbonated-soft-drinks-market; Näherung (2024–2025).
Bürosoftware -Suites (weltweit)	4550	Google Workspace 50%, Microsoft 365 45%, andere 5%. Quelle: Exploding Topics https://explodingtopics.com/blog/google-workspace-stats (2025).
Webbrowser (weltweit, 2025-08)	5076	Chrome 69.23%, Safari 14.98%, Edge 5.03%, Firefox 2.26%, Samsung 1.97%, Opera 1.85%, Rest 4.68%. Quelle: Statcounter https://gs.statcounter.com/browser-market-share.
Lebensmitteleinzelhandel (Schweiz, 2020)	2526	Coop 35.1%, Migros (≈ 30.0%), Denner 9.1%, Aldi (≈ 5.0%), Lidl (≈ 4.5%), Rest (≈ 16.3%). Quellen: Statista https://www.statista.com/statistics/787298/switzerland-market-share-of-food-retailers, ESM Magazine https://www.esmmagazine.com/retail/top-10-supermarket-retail-chains-in-switzerland-238497.

Hinweis: HHI = $\sum s_i^2$ (Marktanteile in Prozentpunkten). Anteile gerundet; wo als *Näherung* gekennzeichnet, wurden seriöse Sekundärquellen und Residualbildung verwendet. Zeitbezug je Zeile angegeben.

Kurz gesagt zeigen die Vergleichswerte eine klare Einordnung: Browser und Bürosoftware sind mit HHI-Werten über 4’500 stark konzentriert, Softdrinks in der Schweiz liegen mit rund 2’250 nur moderat. Der KI-Chipmarkt übertrifft alle Vergleichsmärkte deutlich (HHI 8’496) und verdeutlicht eine sehr hohe Marktkonzentration mit einem klaren Marktführer.

Das Ergebnis bestätigt Hypothese 1 uneingeschränkt. Der Markt für KI-Chips weist eine extrem hohe Konzentration auf, die fast ausschliesslich auf die dominante Stellung von NVIDIA zurückzuführen ist. Diese quantitative Messung untermauert die Wahrnehmung einer monopolistischen Marktstruktur.

5.2.2. Hypothese 2: Technologischer Lock-in durch das CUDA-Ökosystem

Die zweite Hypothese lautet, dass das CUDA-Ökosystem eine signifikante technologische Eintrittsbarriere darstellt und den Wettbewerb durch hohe Wechselkosten einschränkt. Wie im Theorieteil beschrieben, besteht dieser «Lock-in» aus der engen Kopplung von NVIDIAs Hardware mit ihrer proprietären Software-Plattform CUDA.

Die Analyse der Praxis bestätigt dies eindrücklich. Praktisch alle relevanten KI-Frameworks, wie Googles «TensorFlow» und Metas «PyTorch», sind primär für die Ausführung auf CUDA optimiert. Ein riesiges Ökosystem aus spezialisierten Bibliotheken (Bausteine für Compu-

terprogramme), Entwicklungswerkzeugen und einer globalen Gemeinschaft von Millionen Entwicklern ist über mehr als ein Jahrzehnt gewachsen (Corporation, 2024). Für Unternehmen und Forschungseinrichtungen bedeutet dies:

- **Investitionsschutz:** Unternehmen und Forschungslabore haben über mehr als ein Jahrzehnt Millionen Arbeitsstunden und Milliarden von Dollar in die Entwicklung von CUDA-basiertem Code investiert. Ein Wechsel auf eine alternative Architektur wie AMDs ROCm wäre nicht nur ein einfaches Einkaufen der neuen Chips. Es würde eine komplette Neuentwicklung von Kernalgorithmen, eine aufwändige Verifizierung und das Risiko von Leistungseinbussen und neuen Fehlern bedeuten. Ein Unternehmen mit einer über Jahre optimierten CUDA-basierten Simulationssoftware (z.B. in der Medikamentenforschung oder im autonomen Fahren) stünde vor Kosten in Millionenhöhe und jahrelangen Verzögerungen bei einem Wechsel (Center for Security and Emerging Technology et al., 2020; Hase, 2019).

- **Personalressourcen:** Das wertvollste Kapital der Tech-Industrie sind die Entwickler. Universitäten bilden seit über einem Jahrzehnt Ingenieure und Informatiker primär auf CUDA aus. Die Anzahl der Entwickler mit tiefgehender CUDA-Erfahrung übersteigt die von Konkurrenzplattformen um ein Vielfaches. Für ein Unternehmen ist es einfacher und billiger, einen der vielen CUDA-Entwickler einzustellen, als rare ROCm-Spezialisten zu finden oder bestehendes Personal teuer umzuschulen.

- **Performance und Stabilität:** Das CUDA-Ökosystem gilt als ausgereift, stabil und performant. Konkurrierende Plattformen haben zwar aufgeholt, kämpfen aber oft noch mit lückenhafter Dokumentation, kleineren Entwickler-Communities zur Problemlösung und subtilen Stabilitätsproblemen, insbesondere bei komplexen Multi-GPU-Systemen (Kitishian, 2025a).

Selbst wenn ein Konkurrent wie AMD einen Chip mit überlegener Rohleistung anbieten würde, wäre der Wechsel für die meisten Kunden aufgrund der Software-Abhängigkeit unwirtschaftlich. Dieser «Burggraben» schützt NVIDIAs Marktanteil effektiver als reine Hardware-Spezifikationen (Center for Security and Emerging Technology et al., 2020). Jüngste Entwicklungen wie die Einführung des DGX Spark Systems zeigen, dass NVIDIA diesen Ökosystemvorteil weiter ausbaut, um Entwickler noch tiefer zu binden (Kunar, 2025). Dabei handelt es sich um einen kleinen, aber superstarken Computer, der eingeführt wurde, um Entwicklern zu erlauben, eigene KI-Anwendungen zu erstellen und diese auf dem 3 000 $ teuren Gerät zu trainieren (NVIDIA Corporation, 2025). Solche Aufgaben benötigen die stärksten und schnellsten Chips, die zu kaufen sind. Damit man nicht von grossen Firmen oder der Internetanbindung abhängig ist, gibt es Geräte wie dieses, das solche Abhängigkeiten mindert.

Um die Grösse der verschiedenen Software-Ökosysteme vergleichbar zu machen, zeigt Tabelle 2 zwei einfache Messwerte: Wie viele öffentliche Programmierprojekte auf GitHub existieren und wie oft wichtige Software-Bausteine heruntergeladen werden.

Tabelle 2: Ökosystem-Proxy: GitHub-Projekte und Python-Downloads (Monat)

Plattform	GitHub-Projekte[a]	Beispielpaket	Monatliche Downloads[b]	Einordnung[c]
CUDA	5 937	CuPy	32 761	sehr gross
OpenCL	1 350	PyOpenCL	84 161	gross
ROCm	187	amdsmi	3 302	klein
SYCL	145	dpctl	4 491	sehr klein

Quelle: Eigene Darstellung auf Basis öffentlicher API-Daten: («GitHub REST API v3», n. d.; «PyPI Stats API», n. d.).

[a] Anzahl öffentlicher GitHub-Repositories mit entsprechendem Topic (Suchanfragen: *topic:cuda*, *topic:opencl*, *topic:rocm*, *topic:sycl*). Quelle: («GitHub REST API v3», n. d.).

[b] Downloads der letzten 30 Tage eines repräsentativen Python-Pakets pro Plattform (PyPI Stats API). Pakete: *cupy*, *pyopencl*, *amdsmi*, *dpctl*. Quelle: («PyPI Stats API», n. d.).

[c] *Einordnung* dient der schnellen Orientierung und basiert qualitativ auf der relativen Grösse beider Metriken.

Stand: September 2024. Werte sind Näherungen und können je nach Messzeitpunkt leicht variieren.

Kurz erklärt: Die Zahlen zeigen, dass CUDA bei beiden Messwerten deutlich an der Spitze liegt – es gibt viel mehr Projekte und die Software wird häufiger genutzt. OpenCL, ROCm und SYCL folgen mit wesentlich kleineren Werten. Das bestätigt: Weil so viele Entwickler und Projekte bereits auf CUDA setzen, ist ein Wechsel zu einer anderen Plattform sehr aufwendig und teuer. Genau dieser Effekt bindet Unternehmen langfristig an NVIDIA.

Methodische Einschränkung: Diese Indikatoren sind Näherungswerte und können die volle Komplexität des CUDA-Ökosystems nicht vollständig abbilden. Proprietäre Enterprise-Software und wissenschaftliche Codebases, die nicht öffentlich zugänglich sind, werden nicht erfasst.

Hypothese 2 wird somit klar bestätigt. Das CUDA-Ökosystem fungiert als mächtige Eintrittsbarriere und ist der zentrale Faktor, der NVIDIAs Marktdominanz verstärkt und den Wettbewerb hemmt.

5.2.3. Hypothese 3: Wettbewerbskräfte nach Porter

Die dritte Hypothese besagt, dass trotz NVIDIAs Dominanz die Wettbewerbsintensität hoch ist. Eine oberflächliche Betrachtung der Marktanteile zeichnet ein irreführendes Bild eines ruhigen Monopols. Die Analyse nach dem Five-Forces-Modell zeigt jedoch ein hochdynamisches Spannungsfeld auf.

Rivalität unter bestehenden Wettbewerbern: Hoch Obwohl NVIDIA über 90 % des Marktes kontrolliert, ist die Intensität der Rivalität, insbesondere zwischen NVIDIA und AMD, extrem hoch. Sie manifestiert sich jedoch nicht in einem Preiskampf, sondern in einem atemlosen Innovationswettlauf.

- **Produktzyklen:** Jede neue Chip-Generation von NVIDIA (z.B. die Ankündigung der Blackwell-Plattform) wird sofort von AMD und Intel mit Gegenankündigungen und eigenen Roadmap-Updates gekontert (z.B. AMDs MI350). Der Wettkampf findet auf den Datenblättern und in den Benchmarks statt, lange bevor die Produkte in Masse verfügbar sind (Martin, 2024).
- **Strategische Ausrichtung:** NVIDIA verfolgt eine Systemstrategie und verkauft nicht nur Chips, sondern eine komplette Plattform (inkl. Software, Netzwerktechnologie mit Mellanox/NVLink). AMD positioniert sich als offenere Alternative und betont die Open-Source-Natur seiner ROCm-Plattform, um Unzufriedenheit über NVIDIAs «walled garden» zu nutzen (Morgan, 2024).
- **Geopolitische Dimension:** Der Wettbewerb ist auch ein Schauplatz geopolitischer Interessen. Die US-Regierung fördert aktiv einen starken heimischen Wettbewerb (zwischen NVIDIA, AMD, weiteren), um die technologische Führung gegenüber China zu sichern. Dies verleiht der Rivalität eine zusätzliche strategische Ebene (Miller, 2022).

Bedrohung durch neue Marktteilnehmer: Gering Die Eintrittsbarrieren in den High-End-KI-Chipmarkt sind monumental:

1. Kapitalbedarf: Das Design eines modernen KI-Beschleunigers kostet hunderte Millionen Dollar, die Fertigung einer Testserie bei TSMC weitere zig Millionen.
2. Technologisches Know-how: Man benötigt Weltklasse-Teams im Chip-Design, in der Software-Entwicklung und Systemarchitektur.
3. Fertigungszugang: Der Zugang zu den modernsten Fertigungsprozessen (z.B. 3 nm von TSMC) ist begrenzt und wird von den grossen Playern wie Apple und NVIDIA dominiert.
4. Der CUDA-Burggraben: Selbst mit einem technisch überlegenen Chip wäre der Aufbau eines konkurrenzfähigen Software-Ökosystems eine Aufgabe von Jahren, wenn nicht Jahrzehnten.

Trotzdem gibt es spezialisierte Startups wie Cerebras (mit riesigen «Wafer-Scale»-Chips), Groq (mit Fokus auf ultra-niedrige Latenz für Inferenz) und SambaNova. Diese versuchen nicht, NVIDIA frontal anzugreifen, sondern besetzen hochspezialisierte Nischen. Sie stellen für NVIDIA derzeit keine existenzielle Bedrohung dar, treiben aber die Innovation in spezifischen Bereichen voran (Freund, 2024; «The Global AI Talent Tracker 2.0», n. d.).

Bedrohung durch Substitute: Sehr Hoch Dies ist die einflussreichste Kraft, die NVIDIAs Macht begrenzt. Die grössten Kunden von NVIDIA – die sogenannten Hyperscaler (Google, Amazon, Microsoft, Meta) – sind gleichzeitig ihre grössten potenziellen Konkurrenten. Sie haben die finanziellen Mittel, das technische Talent und den Anreiz, eigene, für ihre Workloads optimierte Chips zu entwickeln.

Zur Einordnung zeigt Tabelle 3 die wichtigsten Eigenentwicklungen der grossen Cloud-Anbieter.

Tabelle 3: Überblick über KI-Chip-Eigenentwicklungen grosser Hyperscaler (Stand 2025)

Anbieter	Chip	Foundry (Land)[d]	Kurzbeschreibung
Google	TPU (v5e/v6e)	TSMC (TW)	Beschleuniger für Training/Inference grosser Modelle; interne Nutzung in Rechenzentren («Cloud TPU inference», n. d.; «TPU v5e», n. d.).
Amazon	Inferentia2 / Trainium2	n.ö.	ASICs für Inferenz bzw. Training auf AWS («AWS Inferentia», n. d.; «AWS Trainium», n. d.).
Microsoft	Maia	TSMC (TW)	KI-Beschleuniger für Azure-Workloads (ohne CPU Cobalt) («Inside Maia 100: Microsofts erster AI-Beschleuniger», 2024).
Meta	MTIA	Samsung (KR)	Inferenzbeschleuniger für Ranking/Recommendation-Workloads («Introducing Our Next Generation Infrastructure for AI», 2024; «Meta Switches To Samsung Foundry from TSMC For AI Chip Due To Uncertainty», 2024; «Meta to partner with Samsung to reduce reliance on TSMC: presidential office», 2024; «MTIA v1: Meta's first-generation AI inference accelerator», 2023).
Alibaba Cloud	Hanguang 800	n.ö.	Inferenz-ASIC für E-Commerce/Cloud-Workloads («Announcing Hanguang 800: Alibaba's First AI-Inference Chip», 2020).
Baidu Cloud	Kunlun (II)	Samsung (KR)	Allgemeiner KI-Beschleuniger (Cloud/Edge), 2. Generation in Massenproduktion («Baidu Announces Mass Production of 2nd Generation Kunlun AI Chip», 2021; «Baidu Unveils Kunlun II AI Chip: Rival for Nvidia A100», 2021).
Tencent Cloud	Zixiao	n.ö.	Interner KI-Beschleuniger (Bilder/Video/NLP) («Tencent launches three self-designed chips in semiconductor push», 2021).

Quelle: Eigene Zusammenstellung nach offiziellen Anbieterangaben und Primärberichten («Announcing Hanguang 800: Alibaba's First AI-Inference Chip», 2020; «AWS Inferentia», n. d.; «AWS Trainium», n. d.; «Baidu Announces Mass Production of 2nd Generation Kunlun AI Chip», 2021; «Baidu Unveils Kunlun II AI Chip: Rival for Nvidia A100», 2021; «Cloud TPU inference», n. d.; «Inside Maia 100: Microsofts erster AI-Beschleuniger», 2024; «Introducing Our Next Generation Infrastructure for AI», 2024; «MTIA v1: Meta's first-generation AI inference accelerator», 2023; «Tencent launches three self-designed chips in semiconductor push», 2021; «TPU v5e», n. d.).

[d] Foundry soweit öffentlich spezifiziert; "n.ö." = nicht öffentlich kommuniziert oder keine zuverlässliche Primärquelle.

Diese Chips werden nicht auf dem freien Markt verkauft, doch sie entziehen NVIDIA potenziell Aufträge im Wert von dutzenden Milliarden Dollar. Jeder grosse Cloud-Anbieter, der einen Teil seiner KI-Lasten auf eigene Chips verlagert, setzt eine faktische Preisobergrenze für NVIDIAs Produkte. Wenn NVIDIAs Chips zu teuer werden, wird die Eigenentwicklung umso attraktiver.

Verhandlungsmacht der Lieferanten: Hoch NVIDIAs Geschäftsmodell hat eine Achillesferse: die Abhängigkeit von einem einzigen Hauptlieferanten, dem taiwanischen Halbleiterfertiger TSMC.

- Fertigungsmonopol: TSMC hat ein De-facto-Monopol auf die weltweit modernsten und leistungsfähigsten Chip-Fertigungsprozesse (z.B. N3, N2). Ohne TSMC könnte NVIDIA seine Spitzenchips nicht produzieren.
- Abhängigkeit: Diese Abhängigkeit gibt TSMC eine enorme Verhandlungsmacht bei Preisen und der Zuteilung von Produktionskapazitäten. NVIDIA ist zwar ein riesiger Kunde, konkurriert aber um dieselben knappen Kapazitäten mit anderen Giganten wie Apple. Die gesamte Struktur der Halbleiterindustrie ist ein zentrales Thema in Chris Millers «Chip War» (Miller, 2022).

Verhandlungsmacht der Abnehmer: Gespalten Die Macht der Kunden ist stark polarisiert:

- Hohe Macht (Hyperscaler): Wie unter «Substitute» beschrieben, haben die Top 4 Kunden eine immense Macht. Sie kaufen in riesigen Volumen und können mit Eigenentwicklungen drohen. Sie verhandeln Preise und Lieferkonditionen direkt mit NVIDIA auf höchster Ebene.
- Geringe Macht (der «Rest»): Für die tausenden kleineren Unternehmen, Startups, Universitäten und Forschungsinstitute ist die Situation komplett anders. Sie haben keinerlei Verhandlungsmacht, müssen die Listenpreise akzeptieren und sind von der Verfügbarkeit über Distributoren abhängig. Für sie ist NVIDIA ein reiner «Price-Setter».

Die Analyse der Wettbewerbskräfte bestätigt Hypothese 3 eindrucksvoll. Die Wettbewerbslandschaft ist weit dynamischer, als die Marktanteile vermuten lassen. Der Innovationsdruck durch direkte Rivalen und die strategische Bedrohung durch die Eigenentwicklungen der Hyperscaler sind die wichtigsten Faktoren, die NVIDIAs Macht zügeln.

6. Diskussion

6.1. Synthese der Ergebnisse: Das Gesamtbild der Marktdynamik

Die Analyse zeichnet das Bild einer paradoxen Marktstruktur. Der HHI-Wert von über 8 000 schreit nach Monopol, zementiert durch einen tiefen und breiten CUDA-Burggraben, der neue Marktteilnehmer effektiv abwehrt. Gleichzeitig zeigen die Analysen der Wettbewerbskräfte, dass NVIDIA keineswegs in einer unangreifbaren Festung sitzt. Die Firma befindet sich in einem permanenten Hochgeschwindigkeits-Spannungsfeld. Die Rivalität mit AMD treibt die Innovation an, die Abhängigkeit von TSMC birgt ein strategisches Risiko, und vor allem die Bedrohung durch die Eigenentwicklungen der Hyperscaler setzt klare Grenzen: Werden die Preise zu hoch oder der Service zu schlecht, können Google, Amazon und Microsoft jederzeit

auf ihre eigenen Chips umsteigen. NVIDIA kann es sich nicht leisten, seine grössten Kunden zu verlieren, da diese die einzigen Akteure mit den Ressourcen sind, eine ernsthafte Alternative aufzubauen.

6.2. NVIDIAs «kreatives Monopol»: Innovationsmotor oder Fortschrittsbremse?

- Argument für den Innovationsmotor: Laut Thiel sind Monopolgewinne nicht zwangsläufig schlecht. Sie sind die Belohnung für die Schaffung von etwas fundamental Neuem und Einzigartigem. Wichtiger noch: Nur Unternehmen mit einer derart starken Marktposition und den daraus resultierenden hohen Margen können es sich leisten, langfristige, riskante und kapitalintensive Forschungsprojekte zu finanzieren, die die Grenzen der Technologie verschieben. NVIDIAs massive Investitionen in Forschung und Entwicklung (über 8 Milliarden US-Dollar im Jahr 2023), die jede neue Chip-Architektur, das CUDA-Ökosystem und Projekte wie die Supercomputer-Plattform «Eos» finanzieren, passen perfekt in dieses Bild. Man kann argumentieren, dass die heutige KI-Revolution ohne die «monopolistischen» Gewinne von NVIDIA, die zurück in die Forschung flossen, nicht in dieser Geschwindigkeit stattgefunden hätte («NVIDIA Announces Financial Results for Fourth Quarter and Fiscal 2025», 2025; Thiel & Masters, 2014).

- Argument für die Fortschrittsbremse: Die Kehrseite ist, dass NVIDIAs Dominanz alternative technologische Pfade unterdrücken könnte. Weil das gesamte KI-Ökosystem auf CUDA ausgerichtet ist, haben es radikal andere Hardware-Architekturen (z. B. neuromorphe Chips, die das Gehirn nachbilden) ungleich schwerer, die nötige Traktion bei Entwicklern und Investoren zu finden. Das Quasi-Monopol von CUDA schafft eine «Pfadabhängigkeit», die die gesamte Industrie auf ein einziges Paradigma festlegt und möglicherweise bahnbrechende, aber inkompatible Alternativen ausbremst. Die hohen Preise für NVIDIA-Hardware schränken zudem den Zugang zu KI-Forschung für weniger finanzstarke Institutionen ein, was die Demokratisierung der KI behindert (Kitishian, 2025b; Walter Schmid & Stefano Brusoni, 2025).

6.3. Regulatorische Implikationen und die Rolle offener Standards

Die extreme Marktkonzentration ruft unweigerlich Regulierungsbehörden auf den Plan. Anstatt jedoch ein klassisches Zerschlagungsverfahren anzustreben, das die Innovationskraft lähmen könnte, fokussieren sich Diskussionen in Fachkreisen auf die Software-Ebene. Ein potenzieller Ansatz, wie er auch im Kontext anderer digitaler Ökosysteme diskutiert wird (vgl. EU Digital Markets Act), wäre die Forderung nach mehr Interoperabilität.

Man könnte sich eine Regulierung vorstellen, die NVIDIA verpflichtet, Schnittstellen offenzulegen oder einen Kompatibilitätslayer zu schaffen, der es konkurrierender Hardware (wie von AMD oder Intel) erleichtert, CUDA-basierten Code auszuführen. Dies würde den Lock-in-Effekt direkt adressieren, indem die Wechselkosten für Kunden drastisch gesenkt würden.

- Vorteile: Ein solcher Schritt würde den Wettbewerb auf der Hardware-Ebene massiv beleben, potenziell die Preise senken und die Vielfalt an Hardware-Architekturen fördern.
- Nachteile: NVIDIA würde argumentieren, dass ein solcher Eingriff ihren Anreiz für zukünftige Software-Investitionen schwächt und die enge, leistungsoptimierte Integration von Hardware und Software zerstört, die ihren Erfolg ausmacht. Es bestünde das Risiko, den Innovationsmotor abzuwürgen, um kurzfristig mehr Wettbewerb zu schaffen. Das dürfte NVIDIAs Status als wertvollstes Unternehmen der Welt letztlich ins Schwanken bringen.

Letztlich ist dies eine Abwägung zwischen der Förderung von disruptivem Wettbewerb und der Erhaltung des Innovationsanreizes eines dominanten, hochinnovativen Unternehmens.

Allerdings zeigt die aktuelle Entwicklung, dass NVIDIA existierende Kompatibilitätslayer wie ZLUDA aktiv unterbindet. Seit März 2024 untersagen NVIDIAs Nutzungsbedingungen explizit die Reverse-Engineering-Bemühungen, CUDA-Software mit alternativen Grafikchips kompatibel zu machen (Shilov, 2024).

7. Schlussfolgerung

7.1. Limitationen der Studie

Diese Arbeit untersucht NVIDIAs Marktstellung anhand öffentlich verfügbarer Marktanteils- und Finanzdaten, der Fünf-Kräfte-Analyse und einer Lock-in-Bewertung mittels CUDA-Proxies. Folgende Limitationen sind zu beachten:

- **Datenaktualität:** Marktanteile und HHI basieren auf Schätzungen und Quartalsberichten (Q3/Q4 2024, teils Q1 2025); Schwankungen durch neue Produkteinführungen oder Lieferkettenänderungen sind nicht in Echtzeit abgebildet.
- **Herstellerangaben:** Umsatz-, Performance- und Effizienzwerte stammen oft aus Marketing-Unterlagen oder wurden nicht unabhängig verifiziert. Benchmarks können unter idealen Bedingungen durchgeführt worden sein.
- **Vereinfachte Metriken:** Der HHI reduziert die Marktdynamik auf eine Kennzahl; qualitative Faktoren wie Verhandlungsmacht oder Innovationszyklen fliessen nur teilweise ein.
- **CUDA-Lock-in-Messung:** Die Proxies (Bibliotheken, GitHub-Stars, Kurse, Jobs) sind indirekt und spiegeln nicht die tatsächlichen Wechselkosten jedes Kunden wider.
- **Geografischer Fokus:** Die Arbeit konzentriert sich auf globale und US-/asiatische Märkte; regionalspezifische Regulierungen (z.B. EU-Wettbewerbspolitik) werden nur am Rande behandelt.
- **Zeitliche Eingrenzung:** Der Betrachtungszeitraum endet Anfang 2025; spätere Entwicklungen (z.B. neue GPUs, Gesetzesänderungen) sind nicht erfasst.

Diese Einschränkungen bedeuten, dass die Ergebnisse als Momentaufnahme zu verstehen sind und bei strategischen oder investitionsbezogenen Entscheidungen durch aktuellere oder detailliertere Analysen ergänzt werden sollten.

7.2. Zusammenfassung der Befunde

Eine kompakte Übersicht der Hypothesen, Befunde und Belege bietet Tabelle 4.

Tabelle 4: Zusammenfassung der Hypothesen und Befunde

Hypothese	Befund	Evidenz (Abschnitt/Quelle)
H1: Hohe Marktkonzentration (HHI)	Bestätigt	HHI 2022–2025 durchgehend sehr hoch; Tabellen 5 and 6; (Fernandez, 2025; Weko, 2025)
H2: Lock-in durch CUDA	Bestätigt	Ökosystem/Wechselkosten; Tabelle 2; (Center for Security and Emerging Technology et al., 2020)
H3: Wettbewerbskräfte	Bestätigt (Rivalität/Substitute hoch)	Abbildung 3; (Miller, 2022; Morgan, 2024)

Unterm Strich greifen die drei Hypothesen ineinander: hohe Marktkonzentration, ein starker Software-Burggraben und zugleich spürbarer Wettbewerbsdruck durch Rivalen und Hyperscaler. Das macht den Markt anspruchsvoll, aber nicht festgefahren.

7.3. Fazit

NVIDIAs starke Marktposition im KI-Chip-Sektor ist Fluch und Segen zugleich. Einerseits erschwert sie den Markteintritt für neue Wettbewerber und schafft Abhängigkeiten, die kritisch zu beobachten sind. Andererseits ist es gerade diese Dominanz, die den technologischen Fortschritt in der KI massgeblich beschleunigt hat. NVIDIA setzt Standards, investiert massiv in Forschung und ermöglicht so Innovationen, die weltweit spürbar sind – von effizienteren KI-Modellen bis hin zu neuen Anwendungen im Alltag.

Für die Zukunft ist es wichtig, einen Mittelweg zu finden: Regulierungen und offene Standards können helfen, den Wettbewerb lebendig zu halten, ohne den Innovationsmotor auszubremsen.

7.4. Ausblick

Aus technologischer Perspektive ist es bemerkenswert, wie ein einzelnes Unternehmen die Entwicklung eines gesamten Technologiesektors so massgeblich prägen kann. Die kommenden

Jahre werden zeigen, wie sich der Markt und NVIDIA weiterentwickeln und welche Rolle das Unternehmen beim Übergang zu noch fortgeschritteneren KI-Systemen spielen wird.

Sinnvolle Erweiterungen dieser Arbeit wären insbesondere:

- eine Zeitreihenanalyse der Marktanteile mit Sensitivitätsbandbreiten aus mehreren Datenquellen,
- eine vertiefte Analyse offener Alternativen zu CUDA (z. B. ROCm/SYCL) mit Praxisfällen,
- eine vertiefte regulatorische Fallanalyse im Vergleich mit EU- und US-Behörden – den ersten Schritt dazu (Korrespondenz mit der WEKO, vgl. Anhang) habe ich bereits unternommen.

Wenn offene Schnittstellen und interoperable Software reifen, sinken die Wechselkosten – und es wird wieder spannender, zwischen echten Alternativen zu wählen. NVIDIAs Marktmacht bleibt damit Risiko und Chance zugleich – entscheidend ist, wie Politik, Unternehmen und Forschung sie in den nächsten Jahren gestalten.

8. Literaturverzeichnis

Announcing Hanguang 800: Alibaba's First AI-Inference Chip [Alibaba Cloud Blog]. (2020, 4. Juni). Alibaba Cloud. Verfügbar 7. September 2025 unter https://www.alibabacloud.com/blog/announcing-hanguang-800-alibabas-first-ai-inference-chip_595482

AWS Inferentia [AWS AI Chips]. (n. d.). Amazon Web Services. Verfügbar 7. September 2025 unter https://aws.amazon.com/machine-learning/inferentia/

AWS Trainium [AWS AI Chips]. (n. d.). Amazon Web Services. Verfügbar 7. September 2025 unter https://aws.amazon.com/ai/machine-learning/trainium/

Baidu Announces Mass Production of 2nd Generation Kunlun AI Chip [PR Newswire]. (2021, 18. August). Baidu, Inc. Verfügbar 7. September 2025 unter https://www.prnewswire.com/news-releases/baidu-announces-upgraded-baidu-brain-7-0-and-mass-production-of-2nd-generation-kunlun-ai-chip-301358126.html

Baidu Unveils Kunlun II AI Chip: Rival for Nvidia A100 [Tom's Hardware]. (2021, 18. August). Future Plc. Verfügbar 7. September 2025 unter https://www.tomshardware.com/news/baidu-unveils-kunlun-ii-processor-for-ai

Bennett, M. S. (2023, 24. Oktober). *A Brief History of Intelligence: Evolution, AI, and the Five Breakthroughs That Made Our Brains* [Google-Books-ID: tymCEAAAQBAJ]. HarperCollins.

Bromberg, M. (2024, 6. Dezember). *Herfindahl-Hirschman Index (HHI): Definition, Formula, and Example* [Investopedia]. Verfügbar 16. Februar 2025 unter https://www.investopedia.com/terms/h/hhi.asp

Bruijl, G. (2018). (PDF) The Relevance of Porter's Five Forces in Today's Innovative and Changing Business Environment. *ResearchGate*. https://doi.org/10.2139/ssrn.3192207

Center for Security and Emerging Technology, Khan, S., & Mann, A. (2020, April). *AI Chips: What They Are and Why They Matter*. Center for Security und Emerging Technology. https://doi.org/10.51593/20190014

Cloud TPU inference [Google Cloud TPU Documentation]. (n. d.). Google Cloud. Verfügbar 7. September 2025 unter https://cloud.google.com/tpu/docs/tpu-inference

Corporation, N. (2024, 18. März). *NVIDIA Blackwell Architecture Technical Overview* [NVIDIA]. Verfügbar 24. Februar 2025 unter https://resources.nvidia.com/en-us-blackwell-architecture

CUDA – Thomas-Krenn-Wiki. (n. d.). Verfügbar 1. April 2025 unter https://www.thomas-krenn.com/de/wiki/CUDA

Dexheimer, M. J., & Lechner, C. (2019). Ökosystem-basierte Wettbewerbsstrategien. *Die Unternehmung*, *73*(4), 308–321. https://doi.org/10.5771/0042-059X-2019-4-308

E. Dobbs, M. (2014). Guidelines for applying Porter's five forces framework: a set of industry analysis templates. *Competitiveness Review*, *24*(1), 32–45. https://doi.org/10.1108/CR-06-2013-0059

Fernandez, J. (2025, 4. März). *The leading generative AI companies* [IoT Analytics]. Verfügbar 8. August 2025 unter https://iot-analytics.com/leading-generative-ai-companies/

The Five Forces - Institute For Strategy And Competitiveness - Harvard Business School. (n. d.). Verfügbar 21. Juli 2025 unter https://www.isc.hbs.edu/strategy/business-strategy/Pages/the-five-forces.aspx#:~:text=The%20Five%20Forces%20determine,the%20two%20basic%20drivers

Freund, K. (2024, 13. Februar). *AI Chip Vendors: A Look At Who's Who In The Zoo In 2024* [Forbes] [Section: Enterprise Tech]. Verfügbar 5. September 2025 unter https://www.forbes.com/sites/karlfreund/2024/02/13/ai-chip-vendors-a-look-at-whos-who-in-the-zoo-in-2024/

GitHub REST API v3 [GitHub Docs]. (n. d.). Verfügbar 7. September 2025 unter https://docs.github.com/rest

The Global AI Talent Tracker 2.0 [MacroPolo]. (n. d.). Verfügbar 5. September 2025 unter https://archivemacropolo.org/interactive/digital-projects/the-global-ai-talent-tracker/

Goyal, A. (2020, April). *A Critical Analysis of Porter's 5 Forces* [ResearchGate]. https://doi.org/10.1729/Journal.25126

Grasser, R. (2018). *Netzwerkeffekte bei Informationsgütern: Aktuelle Entwicklung seit 2012* [Diss., Institut für Informationswissenschaft der Technische Hochschule Köln]. Verfügbar 22. August 2025 unter https://publiscologne.th-koeln.de/frontdoor/index/index/docId/1348

Hao, K. (2025). *Empire of AI: dreams and nightmares in Sam Altman's OpenAI*. Penguin Press.

Haramboure, A., Lalanne, G., Schwellnus, C., & Guilhoto, J. (2023, 19. Juni). *Vulnerabilities in the semiconductor supply chain* (OECD Science, Technology and Industry Working Papers Nr. 2023/05) (Series: OECD Science, Technology and Industry Working Papers Volume: 2023/05). https://doi.org/10.1787/6bed616f-en

Hase, M. (2019, 20. September). *Was ist Vendor Lock-in?* [IT-BUSINESS]. Verfügbar 1. April 2025 unter https://www.it-business.de/was-ist-vendor-lock-in-a-879691/

Hoffmann, I. (2020, 28. Januar). *The Winner takes it all: Marktkonzentration bei digitalen Plattformen* [Fraunhofer IAO – BLOG] [Section: Future Mobility]. Verfügbar 25. Juni 2025 unter https://blog.iao.fraunhofer.de/the-winner-takes-it-all-marktkonzentration-bei-digitalen-plattformen/

Inside Maia 100: Microsofts erster AI-Beschleuniger [Microsoft Tech Community]. (2024, 27. August). Microsoft. Verfügbar 7. September 2025 unter https://techcommunity.microsoft.com/blog/azureinfrastructureblog/inside-maia-100-revolutionizing-ai-workloads-with-microsofts-custom-ai-accelerat/4229118

Introducing Our Next Generation Infrastructure for AI [About Meta]. (2024, 10. April). Meta. Verfügbar 7. September 2025 unter https://about.fb.com/news/2024/04/introducing-our-next-generation-infrastructure-for-ai/

Jon Peddie Research. (2025, 5. Juni). *Q1'25 PC graphics add-in board shipments increased 8.5% from last quarter*. Verfügbar 19. Februar 2026 unter https://www.jonpeddie.com/news/q125-pc-graphics-add-in-board-shipments-increased-8-5-from-last-quarter/

Kanellopoulos, N. (2024, 16. Februar). *Analysts Estimate Nvidia Owns 98% of the Data Center GPU Market*. Verfügbar 19. Februar 2026 unter https://www.extremetech.com/computing/analysts-estimate-nvidia-owns-98-of-the-data-center-gpu-market

Kiefer, D. T. (2023, 17. Oktober). *Handelsstreit befeuert Technologiewettkampf bei Chips* [Produktion Online]. Verfügbar 24. Februar 2025 unter https://www.produktion.de/wirtschaft/handelsstreit-befeuert-technologiewettkampf-bei-chips-756.html

Kim, T. (2024). *The Nvidia Way: Jensen Huang and the Making of a Tech Giant* (1st ed). W. W. Norton & Company, Incorporated.

Kitishian, D. (2025a, 10. Juli). *NVIDIA AI Strategy: Analysis of Sustained Dominance in AI - Klover.ai* [Klover.ai - Klover.ai]. Verfügbar 8. August 2025 unter https://www.klover.ai/nvidia-ai-strategy-analysis-sustained-dominance-ai/

Kitishian, D. (2025b, 10. Juli). *NVIDIA AI Strategy: Analysis of Sustained Dominance in AI - Klover.ai* [Klover.ai - Klover.ai]. Verfügbar 5. September 2025 unter https://www.klover.ai/nvidia-ai-strategy-analysis-sustained-dominance-ai/

Krämer, H. (2019). Digitalisierung, Monopolbildung und wirtschaftliche Ungleichheit. *Wirtschaftsdienst*, *99*(1), 47–52. https://doi.org/10.1007/s10273-019-2394-z

Krieser, W. (2023, 11. Dezember). *Halbleiter-Leistungsbauteile - Dioden* [Botland]. Verfügbar 24. Februar 2025 unter https://botland.de/blog/halbleiter-leistungsbauteile-dioden/

Kunar, A. (2025, 22. Juli). *NVIDIA DGX Spark / ASUS Ascent GX10 — Quick Update Jul 22, 2025* [Medium]. Verfügbar 6. August 2025 unter https://medium.com/@andreask_75652/nvidia-dgx-spark-asus-ascent-gx10-quick-update-jul-22-2025-997a9cf634c4

Martin, D. (2024, 17. Juni). *Analysis: As Nvidia Takes AI Victory Lap, AMD Doubles The Trouble For Intel* [CRN]. Verfügbar 5. September 2025 unter https://www.crn.com/news/components-peripherals/2024/analysis-as-nvidia-takes-ai-victory-lap-amd-doubles-the-trouble-for-intel

Mayer, H., Yee, L., Chui, M., & Roberts, R. (2025, 28. Januar). *AI in the workplace: A report for 2025 | McKinsey*. Verfügbar 8. August 2025 unter https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

Meta Switches To Samsung Foundry from TSMC For AI Chip Due To Uncertainty [Techovedas]. (2024, 5. März). Techovedas. Verfügbar 7. September 2025 unter https://techovedas.com/meta-switches-to-samsung-foundry-from-tsmc-for-ai-chip-due-to-uncertainty/

Meta to partner with Samsung to reduce reliance on TSMC: presidential office [The Korea Times]. (2024, 29. Februar). The Korea Times. Verfügbar 7. September 2025 unter https://www.koreatimes.co.kr/business/tech-science/20240229/meta-to-partner-with-samsung-to-reduce-reliance-on-tsmc-presidential-office

Miller, C. (2022). *Chip war: the fight for the world's most critical technology* (First Scribner hardcover edition). Scribner, an imprint of Simon & Schuster.

Morgan, T. P. (2024, 30. Mai). *Key Hyperscalers And Chip Makers Gang Up On Nvidia's NVSwitch Interconnect* [The Next Platform]. Verfügbar 5. September 2025 unter https://www.nextplatform.com/2024/05/30/key-hyperscalers-and-chip-makers-gang-up-on-nvidias-nvswitch-interconnect/

MTIA v1: Meta's first-generation AI inference accelerator [Meta AI Blog]. (2023, 18. Mai). Meta. Verfügbar 7. September 2025 unter https://ai.meta.com/blog/meta-training-inference-accelerator-AI-MTIA/

NVIDIA Announces Financial Results for Fourth Quarter and Fiscal 2025 [NVIDIA Newsroom]. (2025, 26. Februar). Verfügbar 27. Februar 2025 unter http://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2025

NVIDIA Corporation. (2025, 18. März). *NVIDIA DGX Spark* [NVIDIA]. Verfügbar 22. August 2025 unter https://www.nvidia.com/en-us/products/workstations/dgx-spark/

Porter, M. E. (2008, Januar). *The Five Competitive Forces That Shape Strategy*. Verfügbar 28. Februar 2025 unter https://hbr.org/2008/01/the-five-competitive-forces-that-shape-strategy

PyPI Stats API [PyPI Stats]. (n. d.). Verfügbar 7. September 2025 unter https://pypistats.org/api/

Qi, L. (2024). Stock Data Analysis of Competing Companies in Competitive Market: The Case of NVIDIA Corporation. *Highlights in Science, Engineering and Technology, 94*, 493–503. https://doi.org/10.54097/vnv0ec57

Schmidhuber, J. (2022, 29. Dezember). Annotated History of Modern AI and Deep Learning. https://doi.org/10.48550/arXiv.2212.11279

Shilov, A. S. (2024, 4. März). *Nvidia bans using translation layers for CUDA software — previously the prohibition was only listed in the online EULA, now included in installed files [Updated]* [Tom's Hardware]. Verfügbar 5. September 2025 unter https://www.tomshardware.com/pc-components/gpus/nvidia-bans-using-translation-layers-for-cuda-software-to-run-on-other-chips-new-restriction-apparently-targets-zluda-and-some-chinese-gpu-makers

SR 251 - Bundesgesetz vom 6. Oktober 1995 über Kartelle u... [Fedlex]. (n. d.). Verfügbar 25. Juni 2025 unter https://www.fedlex.admin.ch/eli/cc/1996/546_546_546/de

Statista Research Department. (2025). *AI data centers - statistics & facts*. Verfügbar 19. Februar 2026 unter https://www.statista.com/topics/13816/ai-data-centers/

Tan, Y. (2014, 26. September). *Structure-Conduct-Performance Paradigm - an overview | ScienceDirect Topics*. Verfügbar 21. Juli 2025 unter https://www.sciencedirect.com/topics/economics-econometrics-and-finance/structure-conduct-performance-paradigm#:~:text=as%20the%20number%20of,by%20firms%20such%20as

Tencent launches three self-designed chips in semiconductor push [South China Morning Post]. (2021, 3. November). SCMP. Verfügbar 7. September 2025 unter https://www.scmp.com/tech/tech-war/article/3154744/tencent-launches-three-self-designed-chips-expansion-drive-helping

Thiel, P. A., & Masters, B. (2014). *Zero to one: notes on startups, or how to build the future* (First edition). Crown Business.

TPU v5e [Google Cloud TPU Documentation]. (n. d.). Google Cloud. Verfügbar 7. September 2025 unter https://cloud.google.com/tpu/docs/v5e

Walter Schmid & Stefano Brusoni. (2025, 9. April). *KI wird entscheidend sein für die Wettbewerbsfähigkeit* [ETH Zürich] [PDF Studie]. Verfügbar 5. September 2025 unter https://ethz.ch/de/news-und-veranstaltungen/eth-news/news/2025/04/ki-wird-entscheidend-sein-fuer-die-wettbewerbsfaehigkeit.html

Weinzierl, S. (2024, 26. August). *Halbleiter: Fundament der modernen Technologie* [Produktion Online]. Verfügbar 24. Februar 2025 unter https://www.produktion.de/technik/digitalisierung/halbleiter-fundament-der-modernen-technologie-681.html

Weko, W. (2025). Merkblatt und Formular des Sekretariats der WEKO: Relative Marktmacht.

Witt, S. (2025). *The thinking machine: Jensen Huang, Nvidia, and the world's most coveted microchip*. Viking.

Zech, T. (2022, 22. Februar). *Künstliche Intelligenz: Studiengänge in Deutschland* [Wo kann man Künstliche Intelligenz studieren?]. Verfügbar 24. Februar 2025 unter https://www.deutschland.de/de/topic/wissen/wo-kann-man-kuenstliche-intelligenz-studieren

A. Anhang

A.1. HHI-Berechnung

Tabelle 5: HHI-Berechnung für KI-Chips in Rechenzentren (2024)

Unternehmen	Marktanteil (%)	s_i^2
NVIDIA	92	8 464
AMD	4	16
Andere	4	16
Summe HHI		8 496

Quelle: Eigene Darstellung nach (Fernandez, 2025). HHI ist die Summe der quadrierten Marktanteile in Prozentpunkten. Der Wert kann aufgrund von Rundungen leicht vom im Text genannten Wert abweichen.

Tabelle 6: HHI-Index: Jahreswerte 2022–2025

Jahr	NVIDIA (%)	AMD (%)	Andere (%)	HHI
2022	95	3	2	9 038
2023	98	1	1	9 606
2024	92	4	4	8 496
2025 (Q1, indikativ)	92	8	0	8 528

Quellen: 2022: (Qi, 2024); 2023: Wells-Fargo-Schätzung von 98 % im Rechenzentrums-GPU-Markt (Kanellopoulos, 2024); 2024: 92 % im Data-Center-GPU-Markt (Fernandez, 2025); 2025: Q1-Marktanteile 92/8/0 (JPR, AIB-Markt) sowie Statista-Einordnung (Jon Peddie Research, 2025; Statista Research Department, 2025).
Hinweis: Die Werte basieren auf publizierten Marktanteilen; HHI berechnet als Summe der quadrierten Anteile in Prozentpunkten.

A.2. Glossar

AMD — **A**dvanced **M**icro **D**evices – amerikanischer Halbleiterhersteller und Hauptkonkurrent von NVIDIA im GPU-Markt.

API — **A**pplication **P**rogramming **I**nterface – Schnittstelle zur Programmierung von Anwendungen.

ASIC — **A**pplication-**S**pecific **I**ntegrated **C**ircuit – speziell für eine bestimmte Anwendung entwickelter Chip.

Cerebras — Amerikanisches Startup, das riesige «Wafer-Scale»-KI-Chips entwickelt, die eine ganze Siliziumscheibe umfassen.

CUDA — **C**ompute **U**nified **D**evice **A**rchitecture – NVIDIAs proprietäre Plattform zur Programmierung von GPUs für allgemeine Berechnungen.

Deep Learning — Teilbereich des maschinellen Lernens, der auf künstlichen neuronalen Netzen mit mehreren Schichten basiert.

GitHub — Weltweit grösste Plattform zum Teilen und Verwalten von Software-Projekten; Entwickler nutzen sie, um gemeinsam Code zu entwickeln und zu veröffentlichen.

Groq — Amerikanisches Startup, das spezialisierte KI-Chips mit Fokus auf ultraniedrige Latenz für Inferenz entwickelt.

GPU — **G**raphics **P**rocessing **U**nit – Grafikprozessor, ursprünglich für Grafikberechnungen entwickelt, heute auch für KI-Training verwendet.

HHI — **H**erfindahl-**H**irschman-**I**ndex – Mass zur Messung der Marktkonzentration (Wertebereich: 0 bis 10 000).

Hyperscaler — Grosse Cloud-Computing-Anbieter wie Google, Amazon, Microsoft und Meta mit massiven Rechenzentren.

Inferenz — Anwendungsphase eines trainierten KI-Modells zur Vorhersage oder Klassifikation neuer Daten.

KI — **K**ünstliche **I**ntelligenz – Technologie zur Nachbildung menschlicher Intelligenz durch Computersysteme.

Lock-in-Effekt — Wirtschaftliche Situation, in der Kunden aufgrund hoher Wechselkosten an einen Anbieter gebunden sind.

NPU — **N**eural **P**rocessing **U**nit – spezialisierter Prozessor für KI-Berechnungen.

Open-Source — Software, deren Quellcode öffentlich zugänglich ist und frei verwendet werden kann.

Porter's Five Forces
: Modell zur Analyse der Wettbewerbskräfte in einer Branche: Rivalität, neue Anbieter, Lieferanten, Abnehmer und Substitute.

Python
: Weit verbreitete Programmiersprache, besonders populär in der KI-Entwicklung und Datenanalyse.

PyTorch
: Open-Source-Framework für maschinelles Lernen, entwickelt von Meta (Facebook).

ROCm
: **R**adeon **O**pen **C**ompute – AMDs offene Alternative zu CUDA.

SambaNova
: Amerikanisches Startup, das spezialisierte KI-Chips und Datenfluss-Architekturen für Rechenzentren entwickelt.

TensorFlow
: Open-Source-Framework für maschinelles Lernen, entwickelt von Google.

TPU
: **T**ensor **P**rocessing **U**nit – Googles spezialisierter KI-Beschleuniger-Chip.

TSMC
: **T**aiwan **S**emiconductor **M**anufacturing **C**ompany – weltgrösster Auftragsfertiger für Halbleiter.

Walled Garden
: Geschlossenes Ökosystem, in dem der Anbieter volle Kontrolle über Anwendungen und Inhalte hat.

WEKO
: **We**ttbewerbs**ko**mmission – schweizerische Kartellbehörde zur Überwachung des Wettbewerbs.

A.3. E-Mail-Entwurf an die WEKO

Betreff: Anfrage zur wettbewerbsrechtlichen Einordnung der Marktkonzentration bei KI-Rechenzentrumschips

> Sehr geehrte Damen und Herren
>
> Im Rahmen meiner Maturaarbeit am Gymnasium Neufeld (Bern) untersuche ich die Marktdominanz von NVIDIA im Bereich KI-Chips für Rechenzentren aus wettbewerbsökonomischer Perspektive.
>
> Meine Auswertung zeigt für die letzten Jahre eine sehr hohe Marktkonzentration (HHI durchgehend deutlich über 8'000 Punkten) sowie ausgeprägte Lock-in-Effekte durch das CUDA-Ökosystem. Vor diesem Hintergrund bitte ich Sie um eine kurze fachliche Einordnung aus Sicht der WEKO.
>
> Ich wäre Ihnen für eine Orientierung zu folgenden Fragen dankbar:
>
> - Wie beurteilt die WEKO eine derart hohe Konzentration im Lichte von Art. 7 KG (marktbeherrschende Stellung/Missbrauch)?

- Welche zusätzlichen Indizien oder Verhaltensweisen wären aus Ihrer Sicht erforderlich, damit in einem solchen Technologiemarkt ein vertieftes Verfahren angezeigt wäre?
- Inwiefern spielen internationale Entwicklungen (z. B. Verfahren in der EU oder in den USA) für die Schweizer Beurteilungspraxis eine Rolle?

Mir ist bewusst, dass Sie keine verbindliche Vorabbeurteilung im Einzelfall abgeben können. Eine allgemeine Einordnung oder ein Hinweis auf öffentlich zugängliche Leitlinien der WEKO wäre für meine Arbeit bereits sehr hilfreich.

Besten Dank für Ihre Zeit und Ihre Unterstützung.

Freundliche Grüsse

Jonathan Mark Oliver Gerbig
Klasse 26Wd, Gymnasium Neufeld Bern

A.4. Antwort der WEKO

Sehr geehrter Herr Gebrig

Ich gratuliere Ihnen zu Ihrer äusserst gelungenen Maturaarbeit, in welcher Sie untersuchen, wie NVIDIAs Marktdominanz den Wettbewerb in der KI-Halbleiterindustrie beeinflusst.

Unsere Einordnungen von dominanten Unternehmen erfolgen von einem leicht unterschiedlichen Blickwinkel aus, wodurch ein anderer Fokus gelegt wird, die verwendeten Methoden aber immer noch sehr verwandt sind.

Wenn wir eine Unternehmung unter dem Blickwinkel von Art. 7 KG untersuchen, liegt unser Fokus auf der Analyse eines bestimmten (unzulässigen) Verhaltens, welches von einem marktbeherrschenden Unternehmen ausgeht. Dass ein Unternehmen marktbeherrschend ist, sehen wir dabei nicht als Problem an sich – im Gegensatz zum Missbrauch dieser Stellung.

In einer solchen Untersuchung würden wir die folgenden Schritte für die Beurteilung einer beherrschenden Stellung durchgehen, welche mit vier der Five Forces eine starke Überschneidung aufweisen:

- Marktabgrenzung (welche Produkte stehen in Konkurrenz und bilden den Markt, wie ist dieser räumlich zu sehen?)
- Aktuelle Konkurrenz (Stellung von NVIDIA, die gegenwärtige Stärke der Konkurrenten)
- Potenzielle Konkurrenz (Wer kann in den Markt neu eintreten, wer wird voraussichtlich sein Potential ausbauen? Markteintrittsbarrieren)
- Stellung der Marktgegenseite

Ausserhalb der Verfolgung von Verstössen gegen das Kartellgesetz in Form von Untersuchungen beobachtet die WEKO auch Märkte. Bei solchen Marktbeobachtungen wird der Fokus auf die Betrachtung eines Marktes (oder einer Industrie) insgesamt als auf ein bestimmtes Unternehmen gelegt. Da nicht ein bestimmtes Unternehmen im Zentrum steht, eignet sich ein Five-Forces-Ansatz bei Marktbeobachtungen weniger.

Vor diesem Hintergrund können wir Ihnen wie folgt antworten:

Frage 1

Wenn wir ein (hypothetisches) missbräuchliches Verhalten untersuchen würden, gingen wir ähnlich vor. Für die Messung der Marktkonzentration müssten wir in einem ersten Schritt den Markt abgrenzen. Vorliegend wäre dazu eine Auseinandersetzung mit der Frage erforderlich, ob dieser nur aus den GPU-Chips besteht oder ob für die KI-Anwendungen die spezialisierten TPUs oder die ASICs als innerhalb des Marktes anzusehen wären. Je nach Antwort wäre der Marktanteil etwas tiefer als die 90 % in der Arbeit.

Für die Beurteilung einer marktbeherrschenden Stellung würden wir weniger auf den HHI, sondern direkt auf den Marktanteil abstellen. (In der Zusammenschlusskontrolle liegt der Fokus auf die Veränderung der Marktkonzentration, weswegen der HHI zusammen mit der Veränderung des HHI herangezogen wird.) Als Faustregel sehen wir in einem Marktanteil von über 50 % ein starkes Indiz für das Vorliegen einer marktbeherrschenden Stellung. Dies dürfte bei NVIDIA gegeben sein.

Nebst der reinen Marktkonzentration würden wir weitere Indizien betrachten. Dies können beispielsweise anhaltende hohe Margen bzw. Gewinne sein. Ebenso Netzwerk-, Skalen- oder Verbundeffekte. Ihre Analyse der Netzwerkeffekte bzw. des Lock-in mit CUDA entspricht genau dieser Art von verstärkenden Faktoren. Wir müssten dabei etwas mehr in die Tiefe gehen und würden beispielsweise explizit der Frage der Interoperabilität nachgehen, also ob die assoziierten Bibliotheken bzw. die Architektur auch mit anderen Chips nutzbar/kompatibel wäre. Weiter müssten wir die Bedeutung der Software/Architektur für die Hardware einschätzen.

Bei Ihrer dritten Hypothese sprechen Sie die Faktoren an, welche gegen eine marktbeherrschende Stellung sprechen. Diese würden wir unter den Aspekten des potenziellen Wettbewerbs sowie der Stellung der Marktgegenseite prüfen. Sie sprechen mit den Produktzyklen richtigerweise einen Indikator des dynamischen Wettbewerbs an. Mit der Stellung der Hyperscaler als grosse Abnehmer mit dem Potential der Eigenentwicklung weisen Sie auf wichtige Gegenkräfte hin und unterscheiden diese von anderen, schwächeren Marktteilnehmern. Hinsichtlich einer Bedrohung neuer Marktteilnehmer analysieren Sie wichtige Markteintrittshürden, welche gegen einen Markteintritt sprechen. Da der dynamische Wettbewerb eine bedeutende Rolle spielt, ist ohne eigene Untersuchung die Einschätzung weiterer wichtiger Faktoren schwierig. Offene Fragen sind dabei womöglich, ob auch

Konkurrenten wie z. B. AMD oder Intel in Bezug auf den Zugang zu Lieferanten stärker oder schwächer als NVIDIA eingeschränkt sind. Weiter bestehen verschiedene Verhaltensweisen, welche eine dominante Stellung begünstigen könnten und deshalb unter Beobachtung von Wettbewerbsbehörden stehen.

Insgesamt könnte man die statischen Indikatoren so werten, dass von einem gesteigerten Risiko für eine marktbeherrschende Stellung ausgegangen werden muss. Für eine gute Einschätzung müsste die Wettbewerbsbehörde aber zwingend den dynamischen Wettbewerb untersuchen. Diese (und weitere) Abwägungen bilden aktuell offene Fragen, welche sich Wettbewerbsbehörden in der einen oder anderen Form stellen und debattieren, wobei die KI-Chips zumeist zusammen mit Cloud-Computing betrachtet werden:

- Ein OECD-Roundtable zu «Competition in Artificial Intelligence Infrastructure» vom November 2025
- Marktbeobachtung der portugiesischen Wettbewerbsbehörde zu «Competition and Generative AI: Access to AI chips»

Frage 2

Ein Verfahren wäre dann angebracht, wenn wir nebst Anhaltspunkten einer marktbeherrschenden Stellung von NVIDIA auch Anhaltspunkte für ein möglicherweise unzulässiges Verhalten hätten. Ein unzulässiges Verhalten könnte verschiedene Formen aufweisen. Überlegungen dazu und auch Hinweise auf vergangene Fälle in der Chip-Industrie finden Sie in der Hintergrundnotiz des OECD-Sekretariats oder den Länderberichten vom OECD-Roundtable zu «Competition in Artificial Intelligence Infrastructure»

Frage 3

Wir müssen das Kartellgesetz für die Schweiz anwenden. Ausländische Entscheide zu ausländischen Verhältnissen haben dahingehend nicht einen Einfluss auf die Beurteilung unserer nationalen Verhältnisse. Beurteilt jedoch eine ausländische Wettbewerbsbehörde eine Verhaltensweise im Ausland, die analog in der Schweiz anzutreffen ist, würden wir uns von den Entscheiden anderer Behörden sicherlich «inspirieren» lassen. Für die generellen Überlegungen sind die voran erwähnten und andere ausländische Marktbeobachtungen eine wertvolle Hilfe. Zudem stehen wir mit den Wettbewerbsbehörden anderer Länder in Austausch und diskutieren Fragestellungen. Beschränkt ist der Austausch, wenn es um laufende Verfahren geht.

Beste Grüsse

Niklaus Wallimann
Schweizerische Wettbewerbskommission (WEKO)

B. Reflexion

B.1. Ausgangslage und Zielerreichung

Ausgangspunkt meiner Arbeit war die Forschungsfrage: «Wie beeinflusst NVIDIAs Marktdominanz den Wettbewerb in der KI-Halbleiterindustrie?». Mein Ziel war, dieses Problem mehrstufig zu untersuchen: quantitativ über den HHI (Marktkonzentration) und qualitativ über den technologischen *Lock-in* (CUDA) sowie die Wettbewerbsdynamik (Five Forces). Nach der ersten Überarbeitung mit Herrn Stämpfli habe ich den Umfang angepasst und meine Ziele klarer gesetzt. Diese Ziele habe ich erreicht. Die Besprechung und Fokussierung war wichtig, um die 20 Seiten Reintext einhalten zu können.

B.2. Vorgehen und Methoden

Bezüglich Vorgehen startete ich mit der Auswahl eines spannenden Themas. Die Wahl fiel mir schwer, da mich vieles interessiert. Nachdem ich mich für NVIDIA und die Halbleiterindustrie entscheiden hatte, begann ich mit der Recherche. Daraus ergaben sich viele interessante Fragestellungen, die ich im Rahmen dieser Arbeit hätte bearbeiten können. Gelandet bin ich nach Gesprächen mit Herrn Stämpfli bei meiner konkreten Fragestellung. Die Methodik dabei besteht aus drei Teilen: (1) Messung der Marktkonzentration mittels HHI, (2) Analyse des CUDA-*Lock-in* als technologischer Eintrittsbarriere und (3) Bewertung der Wettbewerbskräfte nach Porter mit Fokus auf Rivalität und Substitute. Diese Kombination deckt verschiedene Facetten von Marktmacht ab: Der HHI liefert eine etablierte Kennzahl zur Konzentration; die Lock-in-Analyse erklärt die Mechanik hinter anhaltender Marktmacht; und Five Forces ordnet die tatsächliche Marktdynamik ein. Es war schwierig, für alles verifizierte Daten zu finden. Teilweise musste ich Daten selber erheben, oder sie aus Herstellerangaben übernehmen. Ich habe kritisch verglichen, wo möglich quergeprüft und solche Grenzen transparent gemacht. Ein zusätzliches Experteninterview (z. B. mit einem Vertreter von NVIDIA Schweiz) hätte die qualitative Tiefe erhöht, wurde aber aufgrund von Umfang und Erreichbarkeit nicht umgesetzt.

B.3. Zeit- und Projektmanagement

Schon vor über einem Jahr begann ich mir Gedanken zum Thema zu machen. Ab dem 06. 01. 2025 habe ich konsequent gearbeitet. Insgesamt investierte ich über 300 Stunden in Recherche, Analyse, Lesen englischsprachiger Primär- und Sekundärquellen, Auswertung, Schreiben und die Umsetzung in LaTeX (inkl. Grafiken und Tabellen). Zur Steuerung nutzte ich meinen Kalender, definierte klare Zwischenziele und hielt mich diszipliniert an Meilensteine; wo nötig passte ich sie an (z. B. Präzisierung der Fragestellung im Rahmen der 90-Zeichen-Regel und Verschlankung des Umfangs).

Zur ungefähren Verteilung des Aufwands:

- Recherche: 20 %
- Literatur / Analyse und Auswertung: 35 %

- Schreiben / Überarbeiten: 20 %
- LaTeX inkl. Tabellen und Layout: 22 %
- Organisation / Rest: 3 %

B.4. Herausforderungen und Lösungen

Die drei grössten Hürden waren: (1) der konsequente Umstieg auf LaTeX; (2) nicht zu viel Zeit mit Lesen zu verbringen, durch die Fülle an spannender Fachliteratur; (3) der Wechsel von Word/APA7 zur finalen Arbeit in LaTeX. Ich habe mich bewusst für LaTeX entschieden, um mich an die Arbeitsweisen im Hochschulkontext zu gewöhnen. Gegen die Umfangsrisiken halfen klare Zielgrenzen und die Fokussierung auf drei Kernanalysen (HHI, CUDA-Lock-in, Five Forces). Insgesamt haben diese Entscheidungen die Kohärenz deutlich verbessert.

B.5. Ressourcen, Redlichkeit und KI-Nutzung

Ich habe mir besondere Mühe gegeben, qualitativ hochwertige Quellen auszuwählen und querzuprüfen. Für das Literatur- und Quellenmanagement nutzte ich *Zotero* und in der Endfassung *BibLaTeX mit Biber*. KI-Tools setzte ich als Coach, Lehrer und Recherche-Assistent ein (z. B. für Strukturierungsfragen, LaTeX- und Tabellen-Formatierungen sowie als Rechtschreibe-Hilfe). Ich habe keine KI-generierten Inhalte ungeprüft übernommen oder als Eigenleistung ausgegeben. Die inhaltliche Arbeit, Konzeption, Analyse der Befunde, Auswertung und das eigentliche Schreiben lagen bei mir.

B.6. Lerngewinne und Weiterentwicklung

Fachlich habe ich meine Kompetenzen in Marktanalyse (z. B. HHI), Technologiebewertung (Lock-in/Ökosysteme) und Wettbewerbsmodellen (Five Forces) stark vertieft. Methodisch und technisch habe ich LaTeX (LuaLaTeX), TikZ, BibLaTeX/Biber, APA-Standards sowie den Umgang mit längeren wissenschaftlichen Arbeiten trainiert. Fachübergreifend habe ich besonders in Zeitmanagement, Selbstorganisation und kritischer Quellenbewertung dazugelernt; das über Monate verteilte, ausdauernde Arbeiten an einem Thema, das mich wirklich interessiert, war sehr hilfreich.

B.7. Ergebnisse im Spiegel der Ziele

Die Resultate fielen noch grösser aus als zunächst erwartet: Der HHI belegt eine sehr hohe Konzentration; das CUDA-Ökosystem wirkt als kräftiger «Burggraben»; zugleich zeigen Substitute durch Hyperscaler die Grenzen von Marktmacht. Diese Einsichten habe ich zusätzlich unter dem Befund eines «kreativen Monopols» interpretiert, was half sowohl Risiken, als auch den Innovationsnutzen einzuordnen.

B.8. Limitierungen und Ausblick

Die Aussagekraft ist durch Datenverfügbarkeit und Herstellernähe mancher Quellen begrenzt; ich habe diese Grenzen aufgezeigt und die Ergebnisse entsprechend vorsichtig interpretiert. Für eine Vertiefung böten sich an: (a) Experteninterviews (Industrie/Regulierer), (b) erweiterte Datensätze (z. B. Zeitreihen, Preisdaten), (c) Analyse der Kunden mit Typ von Chip und Bestellwert. Mit reiferen offenen Standards und interoperabler Software könnten Wechselkosten sinken, was ein spannender Punkt wäre, um weiter zu forschen.

B.9. Zwei zukünftige Szenarien und Rückmeldung der WEKO

Als persönliche Weiterführung dieser Arbeit sehe ich zwei plausible Szenarien für die nächsten Jahre:

- **Szenario A – Anhaltende Dominanz:** NVIDIA hält dank CUDA-Ökosystem, Produktzyklen und Lieferkettenzugang eine sehr starke Stellung. Der Wettbewerb bleibt vorhanden, führt aber vor allem zu inkrementellen Verschiebungen.
- **Szenario B – Schrittweise Öffnung:** Offene Standards, stärkere Hyperscaler-Eigenentwicklungen und regulatorische Eingriffe senken Wechselkosten. Dadurch verteilt sich Marktanteil mittelfristig breiter auf mehrere Anbieter.

Noch vor Abgabe habe ich die Schweizerische Wettbewerbskommission (WEKO) kontaktiert und eine ausführliche Antwort von Niklaus Wallimann erhalten (vgl. Anhang). Die Rückmeldung bestätigte, dass meine Analyse – insbesondere die CUDA-Lock-in- und Netzwerkeffekte – methodisch genau jenen verstärkenden Faktoren entspricht, die auch die Behörde bei einer Prüfung nach Art. 7 KG heranziehen würde; ein Marktanteil von über 50 % gilt dabei als starkes Indiz für eine beherrschende Stellung, was bei NVIDIA gegeben sein dürfte. Aufschlussreich war zudem der Hinweis, dass die WEKO weniger auf den HHI als direkt auf den Marktanteil abstellt – eine methodische Nuancierung, die meinen Befund bekräftigt, zugleich aber zeigt, wo mein quantitativer Ansatz und die behördliche Praxis voneinander abweichen. Besonders gefreut hat mich schliesslich, dass die Verweise auf den OECD-Roundtable vom November 2025 und die portugiesische Marktbeobachtung belegen, dass die von mir untersuchten Fragen von führenden Wettbewerbsbehörden weltweit aktiv debattiert werden – ein Zeichen, dass die Fragestellung dieser Arbeit über den schulischen Rahmen hinaus Relevanz besitzt.