[Maschinell übersetzter Text]

Der Klimagrenzpunkt

Der reale Energieverbrauch agentischer KI

Agenten verbrauchen etwa 600-mal mehr Energie als einfache KI-Eingabeaufforderungen

Zeke Hausfather

05. August 2026

Der Energieverbrauch durch KI ist derzeit ein grosses und kontroverses Thema. Glaubwürdige Schätzungen gehen davon aus, dass KI-Rechenzentren bis 2030 für rund 12% des US-Stromverbrauchs verantwortlich sind. Gleichzeitig wurden den Verbrauchern jedoch beruhigend kleine Zahlen über die Auswirkungen ihres eigenen KI-Einsatzes mitgeteilt, Zahlen, die auf den ersten Blick nicht mit der atemberaubenden Grösse ihres Gesamteinsatzes vereinbar zu sein scheinen.

Im Jahr 2025 veröffentlichte Google einen Artikel, in dem berechnet wurde, dass die mittlere Gemini-Textaufforderung nur 0,24 Wattstunden (Wh) verbrauchte, weniger Energie als "neun Sekunden Fernsehen". Etwa zur gleichen Zeit sagte Sam Altman, dass eine durchschnittliche ChatGPT-Abfrage etwa 0,34 Wh verbraucht, und Epoch AI kam mit ähnlichen Zahlen heraus. Autoren wie Andy Masley und Hannah Ritchie haben gezeigt, dass bei diesen Raten eine Person, die Chatbots verwendet, einen ziemlich vernachlässigbaren Einfluss hat, wobei eine Eingabeaufforderung nur etwa 1/150.000 der täglichen Emissionen eines durchschnittlichen Amerikaners ausmacht.

Diese Zahlen sind im Grunde richtig. Sie sind auch zunehmend losgelöst davon, wie KI heute tatsächlich eingesetzt wird.

Die am schnellsten wachsende Art und Weise, wie Softwareentwickler und Wissenschaftler KI tatsächlich nutzen, besteht nicht darin, Fragen in eine Chatbox einzugeben. Vielmehr verwenden wir KI-Agenten mithilfe von Tools wie Claude Code und Codex, die selbstständig planen, Code schreiben, ausführen, die Ergebnisse lesen und iterieren. Diese Agenten führen Dutzende von Modellaufrufen pro menschlicher Eingabeaufforderung durch und beteiligen sich an komplexen Argumentationsketten, bei denen mehrere Antworten auf dieselbe Frage versucht und ausgewertet werden.

Ich arbeite für ein Unternehmen im Silicon Valley (Stripe) und nutze zugegebenermassen mehr als die meisten Menschen die neuesten KI-Tools. Aber ich dachte, es wäre aufschlussreich, tief in meinen eigenen KI-Einsatz der letzten 8 Wochen einzutauchen und den tatsächlichen Energieverbrauch zu berechnen, für den ich verantwortlich war.

In den letzten 8 Wochen habe ich 1.138 Eingabeaufforderungen in Claude Code eingegeben. Diese Eingabeaufforderungen lösten mehr als 14.000 Modellaufrufe aus, die 3,2 Milliarden Token verarbeiteten. Meine beste Schätzung ist, dass dabei etwa 170 kWh Rechenzentrumsstrom verbraucht wurden (mit einem Unsicherheitsbereich von etwa 70 bis 330 kWh über Methoden und Annahmen hinweg). Das entspricht etwa 150 Wh pro Eingabeaufforderung (60 bis 290 Wh), was etwa dem 600-fachen (250 bis 1.200) der Energie einer mittleren Chat-Eingabeaufforderung entspricht. Eine "-Eingabeaufforderung" ist letztlich keine Einheit der KI-Nutzung, genauso wenig wie "Fahrten" ein Mass für das Fahren ist; Es kommt darauf an, wie weit Sie gehen.

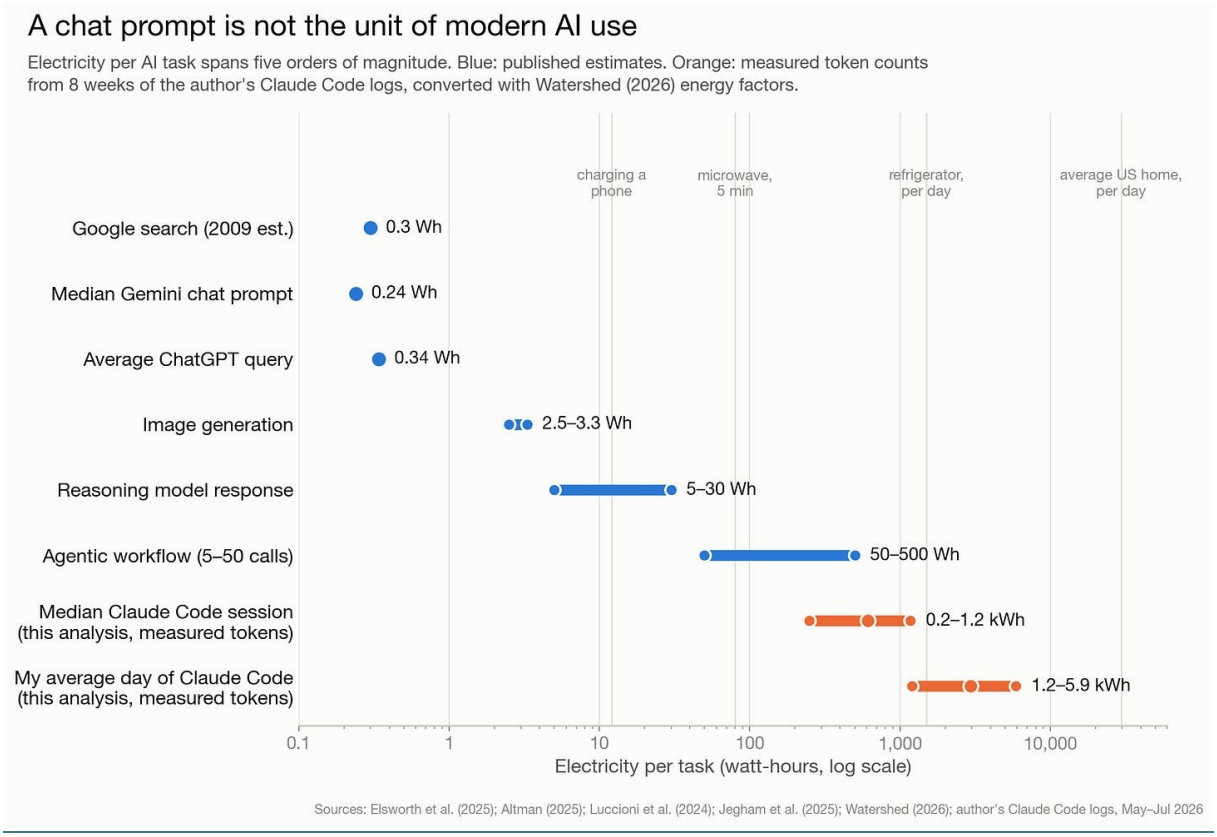
Agenten steigern die KI-Nutzung

Ein Teil des Anstosses für diesen Beitrag ist die Veröffentlichung eines neuen Whitepapers von Watershed (Bistline et al. 2026) Vorschlag eines standardisierten Rahmens für die KI-Emissionsbuchhaltung von Unternehmen. Dies ist die sorgfältigste Behandlung, die ich je gesehen

habe, warum sich die veröffentlichten Zahlen pro Abfrage um Grössenordnungen unterscheiden (meistens Systemgrenzen), und sie enthält eine Zahl, die die gesamte Diskussion neu formulieren sollte: Elektrizität pro KI-Aufgabe erstreckt sich über mehr als fünf Grössenordnungen, von Tausendstel Wattstunden für die Textklassifizierung bis zu 50–500 Wh für einen agentischen Workflow, der 5–50 Grenzmodellaufrufe durchführt. Sie drücken es so aus: Emissionen, die einer "Wechselwirkung" zugeschrieben werden, können den tatsächlich verbrauchten Rechenaufwand "um eine Grössenordnung oder mehr" unterschätzen

Andere Forscher haben ähnliche Ergebnisse gefunden. Bai et al. (2026) mass Codierungsagenten bei realen Softwareaufgaben und stellte fest, dass sie etwa das 1.000-fache der Token einer gewöhnlichen Chatbot-Interaktion verbrauchen. Und diese Art von Agentenaufgaben stellen den schnellsten Treiber für die zunehmende KI-Nutzung dar; Der Economic Index von Anthropic ergab, dass 97% ihrer API-Nutzung mittlerweile "automatisierungsdominante" Muster aufweisen, die mit Agenten verbunden sind.

Um diese Werte ins rechte Licht zu rücken, vergleicht die folgende Abbildung veröffentlichte Schätzungen pro Eingabeaufforderung und Aufgabe (blau) mit dem, was ich anhand meines Claude-Code-Verbrauchs (orange) sowie gängiger Benchmarks für den Energieverbrauch (Betrieb einer Mikrowelle, eines Kühlschranks oder eines ganzen Hauses) gemessen habe:



Stromverbrauch pro KI-Aufgabe, einschliesslich veröffentlichter Schätzungen (blau) und Werte, die aus meinen eigenen Claude Code-Sitzungsprotokollen berechnet wurden (orange). Gemessene Token-Zählungen, konvertiert mit Energiefaktoren der Aktivitätsstufe von Bistline (2026); orangefarbene Bereiche umfassen Cache-Read-Energieannahmen von 1 % bis 25 %.

Meine mittlere Claude-Code-Sitzung verbraucht etwa 0,6 kWh (0,25 bis 1,2 kWh), was am oberen Ende der Schätzung des generischen Agentenverbrauchs von Watershed liegt, und das Fünfzigfache der Energie, die zum Aufladen eines Mobiltelefons verwendet wird. Mein durchschnittlicher Tag mit

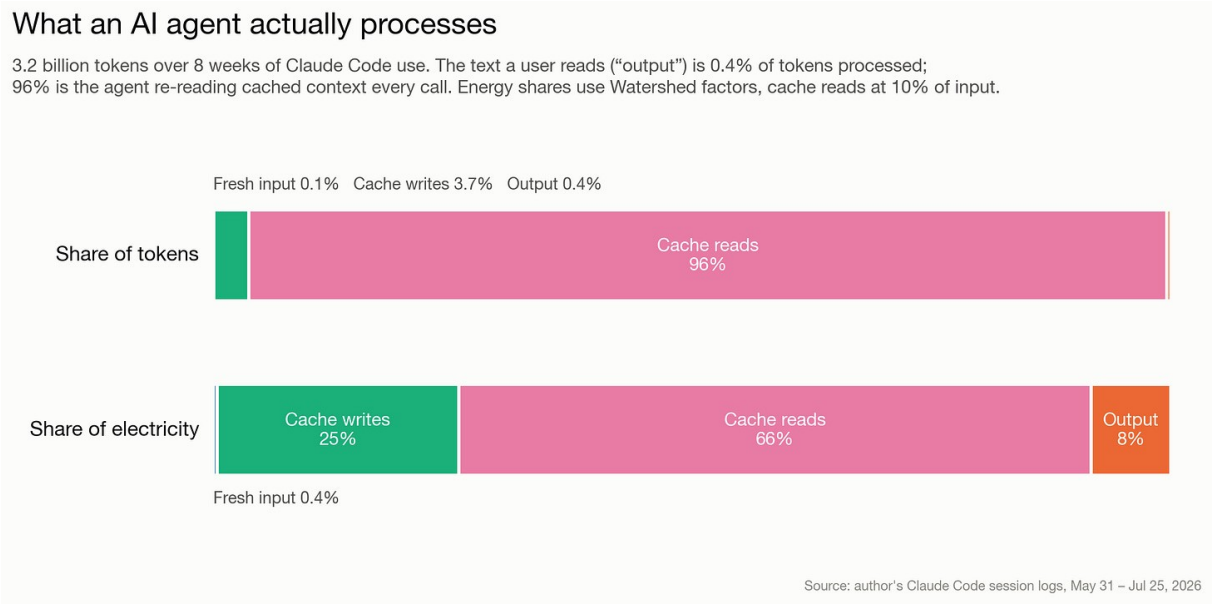
Claude Code (3,0 kWh, Bereich 1,2 bis 5,9 kWh) verbraucht mehr Strom als der Betrieb von zwei Kühlschränken.

Messung meines eigenen Fussabdrucks

Claude Code führt vollständige lokale Transkripte jeder Sitzung, einschliesslich der genauen Token-Zählungen, die die API-Berichte für jeden Modellaufruf melden.1 Dadurch weiss ich genau, für wie viel KI-Nutzung ich verantwortlich war, anstatt sie einfach aus veröffentlichten Benchmarks zu extrapolieren; Es ist nur der Schritt, die verwendeten Token in Energie umzuwandeln, der Annahmen erfordert.

Als Erstes stellte ich fest, dass die Lücke zwischen “Eingabeaufforderungen” und der Realität enorm ist: Meine 1.138 eingegebenen Eingabeaufforderungen führten zu etwas mehr als 14.000 verschiedenen Modellaufrufen (12 pro Eingabeaufforderung), und jede Eingabeaufforderung verbrauchte durchschnittlich 2,9 Millionen Token. Zum Vergleich: Typische webbasierte KI-Chat-Austausche ohne Begründung oder Websuchen verwenden nur etwa tausend Token.

In den letzten 8 Wochen hat mein Claude-Code 3,2 Milliarden Token verwendet. Diese kamen überwiegend daher, dass der Agent sein eigenes Arbeitsgedächtnis erneut las. Jedes Mal, wenn ein Agent einen Schritt macht (z. B. Führt einen Befehl aus, liest eine Datei oder ruft ein Tool auf), verarbeitet das Modell seinen gesamten akkumulierten Kontext erneut. Die folgende Abbildung zeigt die Aufschlüsselung der Art und Weise, wie Token verwendet wurden, und ihren Anteil am gesamten Stromverbrauch.

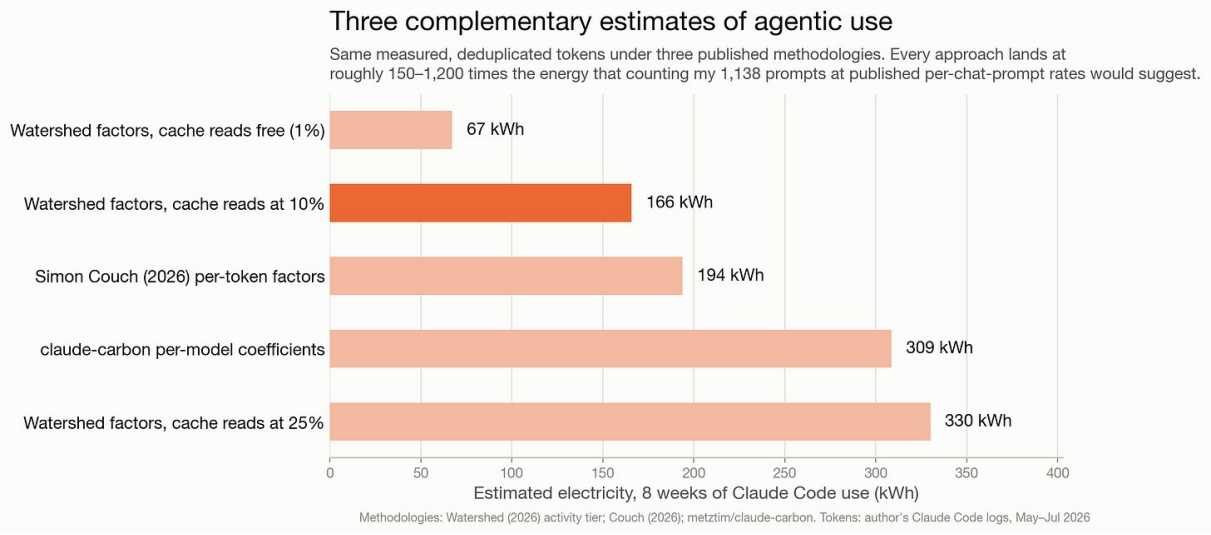


Token und geschätzte Stromzusammensetzung meiner Claude Code-Nutzung, 31. Mai bis 25. Juli 2026. “Cache-Lesevorgänge” sind zuvor verarbeitete Kontexte, die bei jedem Modellaufruf erneut aus dem Schlüsselwert-Cache gelesen werden; “Cache-Schreibvorgänge” sind neue Kontexte, die verarbeitet und gespeichert werden; “Ausgabe” ist vom Modell generierter Text und Code. Stromanteile verwenden Bistline-Faktoren (2026) mit Cache-Lesevorgängen von 10 % der Frischeinspeisenergierate.

Der Text, den ich tatsächlich sehe (zB Die Ausgabe des Modells) beträgt nur etwa 0,4% aller verarbeiteten Token. Etwa 96 % der Token sind Cache-Lesevorgänge, bei denen der Agent in jedem dieser 14.000 Schritte seinen eigenen Kontext erneut liest. Dies ist für die Energieschätzung von enormer Bedeutung, da das erneute Lesen eines zwischengespeicherten Tokens viel günstiger ist als

die Verarbeitung eines neuen. KI-Unternehmen verlangen für Cache-Lesevorgänge etwa 10 % des Preises im Vergleich zu neuen Inhalten, und ich verwende dieses Verhältnis als meine zentrale Energieannahme, mit 1 % und 25 % als Grenzen.[2]

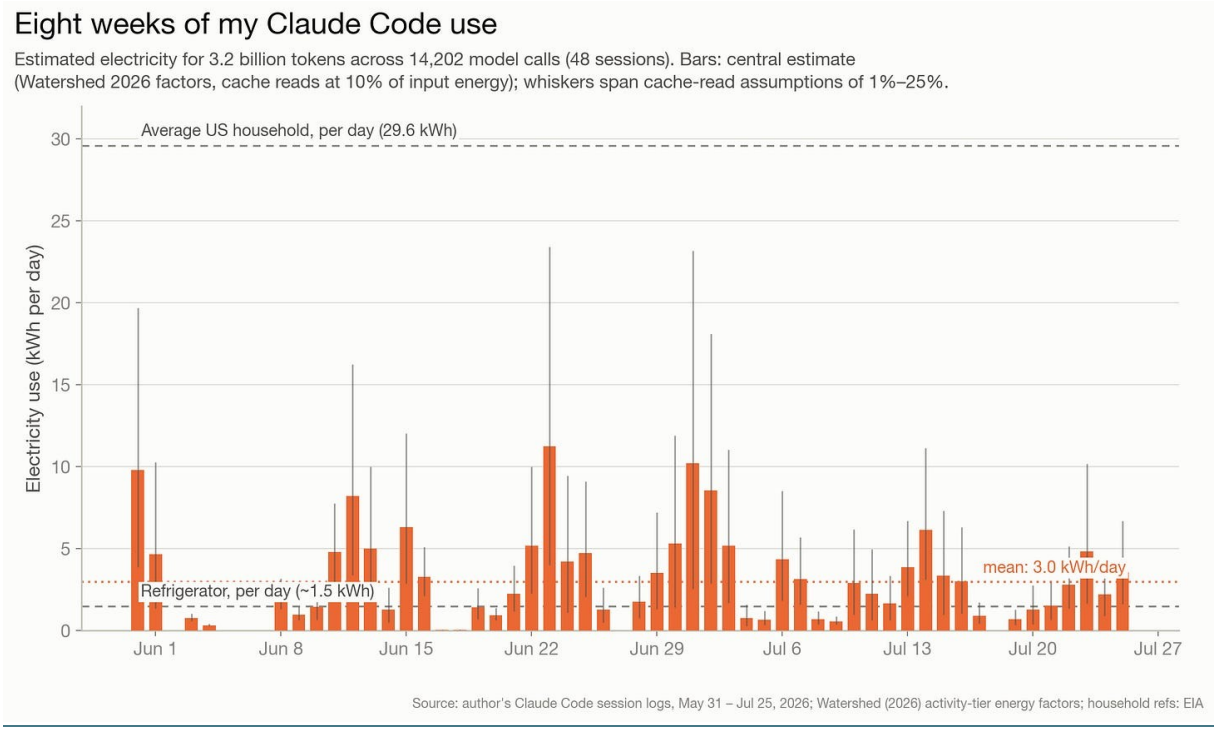
Da ausserhalb der Labore niemand die wahre Energie pro Token eines Grenzmodells kennt (Anthropic hat keine Zahlen pro Eingabeaufforderung oder pro Token veröffentlicht, was das Watershed-Papier höflich, aber entschieden als die grösste Datenlücke in diesem Bereich bezeichnet), habe ich meine gemessenen Token-Zählungen mit drei unabhängigen veröffentlichten Methoden durchgeführt: Watersheds Aktivitätsfaktoren, die aus der Arbeit von Epoch AI abgeleiteten Schätzungen von Simon Couch pro Token und die aus der Preisableitung des Claude-Carbon-Tools abgeleiteten Preiskoeffizienten.



Geschätzter Stromverbrauch für die 3,2 Milliarden Token, die ich unter drei veröffentlichten Methoden verwendet habe: Watershed-Aktivitätsstufenfaktoren unter drei Cache-Read-Annahmen, Couch (2026)-Pro-Token-Faktoren und Claude-Carbon-Pro-Modell-Koeffizienten.

Jede dieser Methoden liefert eine Antwort zwischen etwa 70 und 330 Kilowattstunden über 8 Wochen. Die Schätzung ist wirklich unsicher, und zwar um den Faktor ~2 in beide Richtungen. Die allgemeinere Schlussfolgerung lautet jedoch nicht: Wenn ich meine 1.138 Eingabeaufforderungen mit den veröffentlichten Raten pro Chat-Eingabeaufforderung gezählt hätte, hätte das etwa 0,3 kWh ergeben, während die Realität 150- bis 1.200-mal so hoch ist.

Mein tägliches Muster des Energieverbrauchs ist in der folgenden Abbildung dargestellt. Die tägliche Variabilität ist enorm: An meinem schwersten Tag (zentrale Schätzung von 11 kWh) durchliefen mehrere parallele Agenten eine grosse Geodatenanalyse und verbrauchten mehr als ein Drittel des gesamten täglichen Stroms eines durchschnittlichen US-Hauses. Dies spiegelt die Tatsache wider, dass selbst innerhalb der Kategorie der Agentennutzung die Komplexität der Aufgabe und die Anzahl der gleichzeitig verwendeten Unteragenten den daraus resultierenden Energieverbrauch stark beeinflussen.



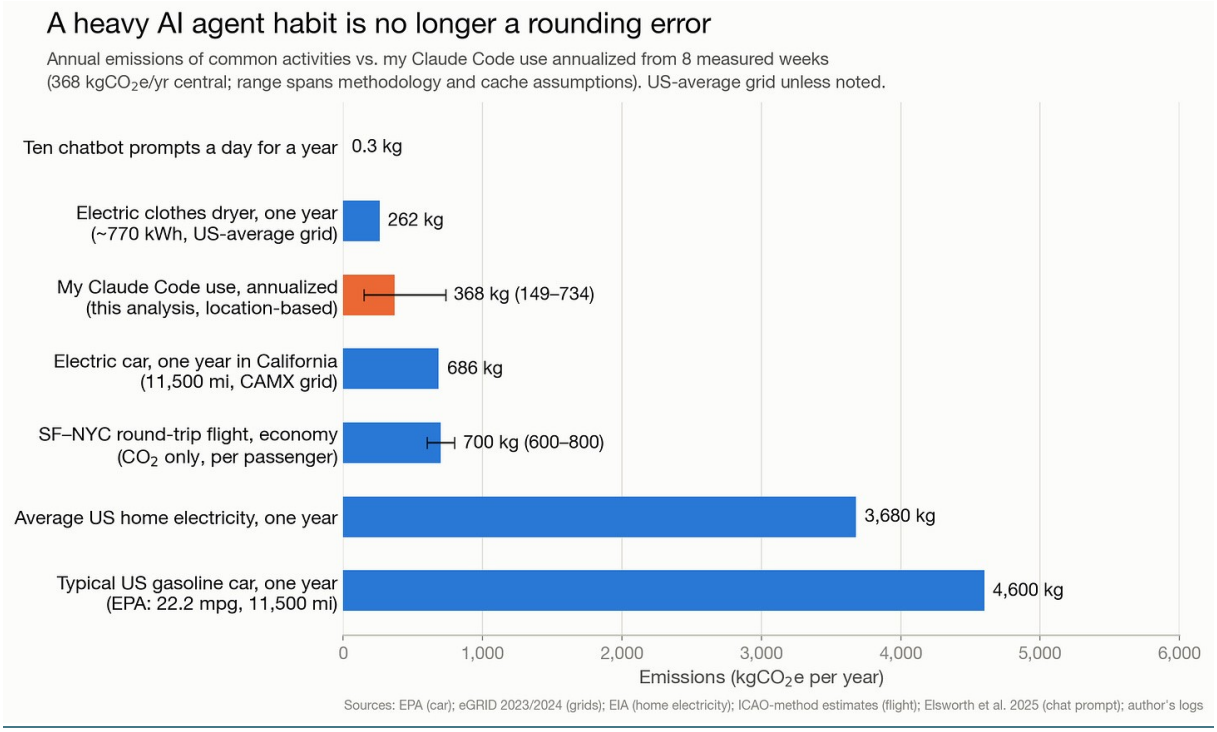
Geschätzter täglicher Stromverbrauch meiner Claude-Code-Nutzung (Balken: Cache-Lesevorgänge bei 10 % der Eingangsenergie; Whisker: 1 % bis 25 %). Referenzlinien zeigen den typischen täglichen Stromverbrauch eines Kühlschranks und eines durchschnittlichen US-Haushalts.

Meine Zahlen sind etwas höher als einige der anderen veröffentlichten Schätzungen zum Einsatz von Wirkstoffen, und es lohnt sich, ein wenig nachzuforschen, um herauszufinden, warum. Couch schätzte, dass eine durchschnittliche Claude-Code-Sitzung etwa 41 Wh verbraucht, was 24 Modellaufrufe und 592.000 Token umfasst. Andy Masleys Rechner für Juni 2026 beziffert eine Claude Opus-Agentensitzung mit 100.000 Token auf ~459 Wh. Meine mittlere Sitzung beträgt ~600 Wh und umfasst über hundert Anrufe und rund zehn Millionen Token, darunter zahlreiche Unteragenten für grosse Datenanalyseprojekte. Hannah Ritchies hypothetischer starker Benutzer (24 Agentenabfragen pro Tag) kam auf 2,4 kWh/Tag, während ich für meinen tatsächlichen Verbrauch eine zentrale Schätzung von 3,0 kWh/Tag (1,2 bis 5,9 kWh) mass.

Keine dieser Schätzungen ist unbedingt falsch, sie spiegeln lediglich eine breite Palette tatsächlicher Nutzungsannahmen wider. Softwareentwickler, Forscher und Datenanalysten (z. B. Leute wie ich) liegen wahrscheinlich ziemlich weit unten in der Nutzungsverteilung. Gleichzeitig wird die Nutzung im Laufe der Zeit wahrscheinlich zunehmen, da komplexere Agentenwerkzeuge zunehmend zur Norm werden.

Wie ein Jahr davon aussieht

Wenn wir davon ausgehen, dass diese 8 Wochen ziemlich typisch sind, können wir schätzen, dass ein ganzes Jahr meiner agentischen Claude-Code-Nutzung etwa 1,1 MWh Rechenzentrumsstrom (0,4 bis 2,2 MWh) verbrauchen würde, was etwa einem Zehntel dessen entspricht, was ein durchschnittlicher US-Haushalt verbraucht. Unter Anwendung der durchschnittlichen Netzintensität in den USA beträgt diese etwa 370 kgCO2e pro Jahr (150 bis 730 kgCO2e).[3]



Jährliche Emissionen gemeinsamer Aktivitäten im Vergleich zu meiner annualisierten Claude-Code-Nutzung. Auto: Typisches Personenkraftwagen der EPA (22,2 mpg, 11.500 Meilen/Jahr). EV: 11.500 Meilen/Jahr bei 0,30 kWh/mi im kalifornischen Netz. Flug: Economy-Hin- und Rückflug nach ICAO-Methode, nur CO2. Hausstrom: EIA-Durchschnittshaushalt eines US-Haushalts im US-Durchschnittsnetz. Trockner: typischer elektrischer Wäschetrockner mit ~770 kWh/Jahr (DOE) im US-Durchschnittsnetz.

Meine persönliche und berufliche KI-Nutzung verursacht mittlerweile etwas mehr Emissionen pro Jahr als der Betrieb eines elektrischen Wäschetrockners und etwa halb so viel wie das Fahren eines Elektroautos 11.500 Meilen in Kalifornien oder die Fahrt mit einem Hin- und Rückflug von San Francisco nach New York in der Economy-Klasse.[4] Es sind etwa 8% der jährlichen Emissionen eines typischen amerikanischen Benzinautos und etwa 2% des jährlichen Treibhausgas-Fussabdrucks eines durchschnittlichen Amerikaners von ~18 Tonnen.

Dies ist gleichzeitig eine grosse Emissionsquelle und ein relativ bescheidener Teil meines gesamten Kohlenstoff-Fussabdrucks. Normalerweise nehme ich zweimal im Jahr einen Hin- und Rückflug von San Francisco an die Ostküste, um meine alternden Eltern zu besuchen (ganz zu schweigen von Arbeitsreisen), und ich verliere bei dieser Wahl im Allgemeinen nicht den Schlaf. Darüber hinaus ist die Endverwendung grundsätzlich viel einfacher zu dekarbonisieren als die Luftfahrt (mehr dazu weiter unten). Dies stellt aber auch eine neue Nettoemissionsquelle dar, und das zu einer Zeit, in der die globalen Temperaturen in die Höhe schnellen und unsere Emissionsreduktionsziele zunehmend vom Kurs abweichen.

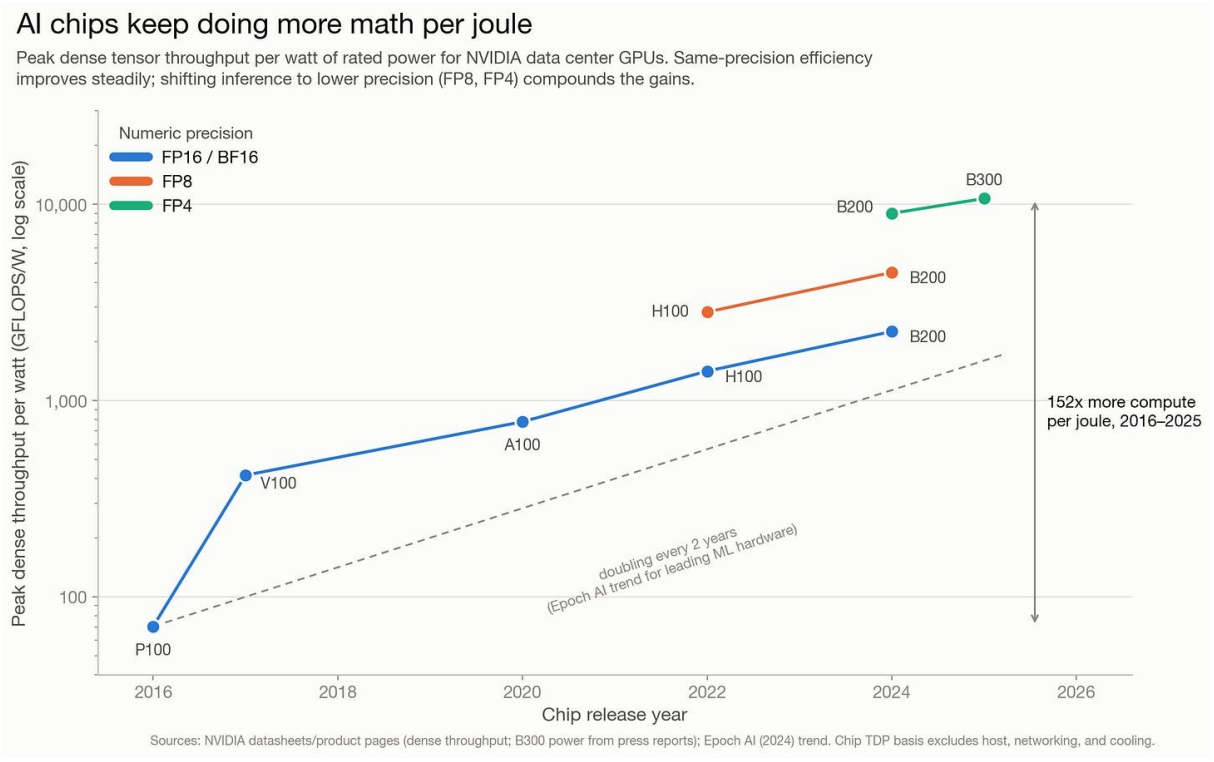
Was also tun wir dagegen?

Nachdem ich den grössten Teil dieses Beitrags damit verbracht habe, zu argumentieren, dass der Einsatz agentischer KI hundertmal energieintensiver ist, als die Chatbot-Zahlen vermuten lassen, möchte ich klarstellen, dass ich nicht glaube, dass die Antwort Schuldgefühle oder Abstinenz sind. Aber hier gibt es echte Hebel, mit denen wir den Verlauf des KI-Energieverbrauchs und der Emissionen in Zukunft gestalten können.

Auf der persönlichen Seite können wir versuchen, mit Agentenwerkzeugen nicht leichtfertig umzugehen. Es ist ein echter Unterschied, ob man fünf parallele Agenten auf ein schwieriges Forschungsproblem aufmerksam macht oder dasselbe tut, um eine Barwette abzuschliessen (oder, in

meinem Fall, mit meiner Tochter Spiele zum Thema Axolotl zu machen). Auch die von Ihnen verwendeten Modelle sind wichtig: Das Senden einfacher Aufgaben an kleinere Modelle verbraucht pro Token vielleicht 5 bis 7 Mal weniger Energie als die Standardeinstellung eines Grenzmodells5, und das ist es, was ich zunehmend für Suchvorgänge und mechanische Arbeiten mache. Trotzdem möchte ich es nicht übertreiben. Mein gesamter jährlicher KI-Fussabdruck beträgt einige hundert Kilogramm CO2; Die persönliche Zurückhaltung der kleinen Gruppe starker Nutzer wird keine Kurven verbiegen.

Der Technologiehebel ist leistungsstärker und wirklich beeindruckend. Die folgende Abbildung zeigt die Energieeffizienz der Rechenzentrumschips von NVIDIA im letzten Jahrzehnt. Die Menge an Mathematik, die ein KI-Chip pro Joule Energie leisten kann, ist seit 2016 um etwa das 150-fache gestiegen und verdoppelt sich laut Epoch AI etwa alle zwei Jahre. Die neuesten B300-Chips, die mit der niedrigsten unterstützten Präzision laufen, verbrauchen etwa ein Viertel der Energie pro Betrieb der H100 aus dem Jahr 2022, die die heutigen Grenzmodelle trainierten. Dies entspricht einer 3,8-fachen Verbesserung der Energieeffizienz in drei Jahren. Softwaregewinne können dies noch schneller machen: Google berichtet, dass die Energie einer mittleren Gemini-Eingabeaufforderung in einem einzigen Jahr durch eine Kombination aus besseren Modellen, besseren Tools und besserer Hardware um das 33-fache gesunken ist.



Spitzendichter Tensordurchsatz pro Watt Nennchipleistung für NVIDIA-Rechenzentrums-GPUs, nach Veröffentlichungsjahr und numerischer Genauigkeit. Die gestrichelte Linie zeigt den Trend von Epoch AI, dass sich die Energieeffizienz führender ML-Hardware alle zwei Jahre verdoppelt.

Aber wenn 150-fache Effizienzgewinne den Energieverbrauch der KI senken würden, hätten sie es inzwischen getan. Dies ist das Jevons-Paradoxon in Aktion: Wenn Berechnungen pro Token billiger werden, führt dies wiederum tendenziell zu einem höheren Grad an KI-Nutzung. Aus Effizienzgründen kostet meine Agentengewohnheit 170 kWh statt der 950 kWh, die sie mit Hardware aus dem Jahr 2020 verbraucht hätte. Aber Effizienz bestimmt nur, wie viel Intelligenz wir pro Energieeinheit bekommen, aber bisher hat sie keine Anzeichen dafür gezeigt, den Gesamtenergieverbrauch der KI zu bestimmen.

Deshalb ist der Hebel, der eigentlich am wichtigsten ist, die Kohlenstoffintensität des Stroms. Jede Zahl in diesem Beitrag ging vom durchschnittlichen Stromnetz in den USA aus. Wenn ich die gleiche Arbeitslast mit weitgehend sauberer Energie ausübe, verringert sich mein Fussabdruck um etwa 90 %. Anders als in der Luftfahrt handelt es sich hierbei um eine Endverwendung, deren Dekarbonisierung wir bereits kennen.

Das Problem besteht darin, dass wir uns heute in die falsche Richtung bewegen: Ein beträchtlicher Teil der geplanten Kapazitäten der US-Rechenzentren soll eine eigene Stromerzeugung hinter dem Zähler aufbauen, und fast drei Viertel davon sind Erdgas. KI-Unternehmen mit echten Klimaverpflichtungen müssen bei der Suche nach Alternativen besser abschneiden: Solarenergie plus Speicherung (über die ich 2024 eine Studie mitgeleitet habe), Kernenergie der nächsten Generation und Neustarts stillgelegter Reaktoren, verbesserte Geothermie und Standortwahl von Rechenzentren in Regionen, in denen sowohl die durchschnittliche als auch die marginale Stromerzeugung kohlenstoffarm ist).

Es gibt auch eine Silberstreifenversion dieser Geschichte, in der die KI-Nachfrage zu einem Vorteil für die Dekarbonisierung wird. Um Netto-Null-Emissionen zu erreichen, muss die Stromerzeugung bis Mitte des Jahrhunderts etwa verdreifacht werden, da wir fast alle derzeitigen Nutzungen fossiler Brennstoffe durch sauberen Strom ersetzen. Die Barrieren sind meist nicht technologischer Natur, sondern Dinge wie Verbindungswarteschlangen, Genehmigungen und Übertragung. Der KI-Buildout ist eine Vorschau auf diese Welt mit schnell steigender Stromnachfrage, unterstützt von Unternehmen mit enormem Kapital und ungewöhnlicher Dringlichkeit. Wenn dieses Geld und diese Ungeduld dafür ausgegeben werden, die Beseitigung dieser Hindernisse zu beschleunigen (Kauf von sauberem Strom für Unternehmen, Finanzierung der Übertragung, Übernahme der frühen Kosten für fortschrittliche Kernenergie und Geothermie, wie es frühe Unternehmenskäufer für Wind- und Solarenergie taten), könnte der KI-Boom dazu führen, dass das Netz sauberer wird, als es es vorgefunden hat.Wenn es für Gasturbinen hinter dem Zähler ausgegeben wird, wird es das nicht tun. Diese Entscheidung wird gerade getroffen und wird weitaus wichtiger sein als die Anzahl der Eingabeaufforderungen, die einer von uns eingibt.

Im Interesse der vollständigen Offenlegung: Der Python-Code, der der Analyse und den Abbildungen in diesem Beitrag zugrunde liegt, wurde natürlich mit Hilfe von Claude Code erstellt, aber das Schreiben stammt vollständig von mir.[6]

1

Claude Code zeichnet die API-Nutzung einschliesslich der Menge an nicht zwischengespeicherten Eingabetoken, Cache-Erstellungstoken, Cache-Lese-Token und Ausgabetoken pro Modellaufruf mit Modell-IDs und Zeitstempeln auf. Eine Subtilität der Protokollierung ist von grosser Bedeutung: Jede API-Antwort wird als eine Zeile pro Inhaltsblock in das Protokoll geschrieben, wobei jede Zeile das vollständige Nutzungsobjekt der Nachricht wiederholt, sodass eine naive zeilenweise Summe Token um den Faktor ~2,2 doppelt zählt. Alle Zahlen hier zählen jede API-Nachricht einmal, dedupliziert durch die Nachrichten-ID.

2

Ein Cache-Lesevorgang ruft bereits berechnete Aufmerksamkeitszustände aus dem Speicher ab, anstatt sie neu zu berechnen. Dies ist also viel günstiger als die Eingabe neuer Daten, aber nicht kostenlos. Der Cache-Lesekontext macht die Generierung jedes Ausgabetokens im langen Kontext immer noch teurer. Anthropische Preis-Cache-Lesevorgänge bei 10 % des frischen Inputs, und die Claude-Carbon- und Couch-Methoden verwenden beide ~10 % als Energieverhältnis. Watershed kennzeichnet die Cache-Verarbeitung als bekannte Lücke in der Pro-Token-Abrechnung; meine Bandbreite von 1 % bis 25 % soll den plausiblen Bereich abdecken.

3

Verwendung des US-Durchschnitts von eGRID 2024 von 341 gCO2e/kWh, da Anthropic nicht bekannt gibt, wo sich ihre Rechenzentren befinden und welche Stromquellen sie nutzen. Die marktbasierten Emissionen (unter Berücksichtigung der Anbieter' beim Kauf sauberer Energie) wären wahrscheinlich niedriger, möglicherweise sogar viel niedriger. Diese Schätzung schliesst meinen Laptop aus, der bei ~50 W gegenüber 3,0 kWh/Tag Rechenzentrumslast vernachlässigbar ist.

4

Beachten Sie, dass die Flugschätzung hier nur die direkten CO2-Emissionen aus dem Flugverkehr berücksichtigt. Die Einbeziehung von Kondensstreifen und anderen sekundären Faktoren würde die Flugemissionen wahrscheinlich um mindestens 50 % erhöhen.

5

Basierend auf der Ableitung des Energieverbrauchs durch Token-Preise ergibt Claude-Carbon ~0,3 J/Token für Modelle der Haiku-Klasse gegenüber ~2 J/Token für Modelle der Opus-Klasse.

6

In einem guten Beispiel dafür, warum Sie die mit KI-Codierungstools geleistete Arbeit immer noch einmal überprüfen müssen, hat Claude versehentlich seine ursprüngliche Schätzung meiner Token-Nutzung verdoppelt, da alle relevanten Dateien zweimal gespeichert wurden, und sie einfach alle addiert. Ich habe es nur bemerkt, weil die Zahlen zu hoch schienen!
