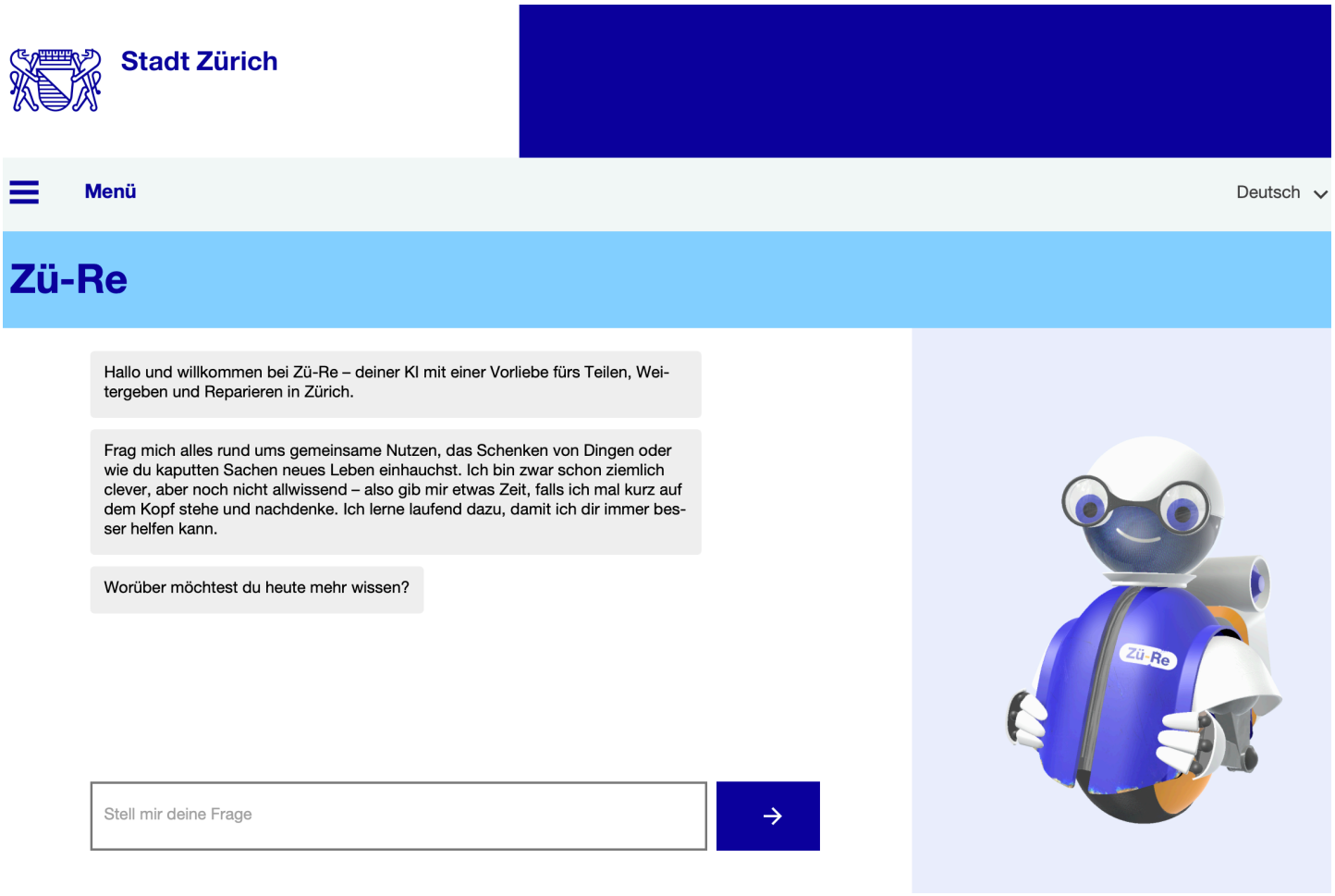
NICHT JEDER PROMPT IST EIN ENERGIEFRESSER

Der Boom der Künstlichen Intelligenz (KI) führt zu einem Mehrverbrauch an elektrischem Strom, das ist offenkundig. Auf der anderen Seite kann KI den Energieverbrauch drosseln, indem sie Menschen unterstützt, einen nachhaltigeren Lebensstil zu pflegen. Genau hinzuschauen lohnt sich also, gerade auch bei ChatGPT und anderen Chatbots. Denn nicht jede Anfrage kostet gleich viel Energie. Eine Untersuchung im Auftrag des BFE hilft, den Stromverbrauch besser zu beurteilen.



Startseite des Nachhaltigkeits-Chatbots Zü-Re. Zü-Re stand im Zentrum einer Fallstudie, mit der Forscher den Stromverbrauch von Chatbots auf der Grundlage Künstlicher Intelligenz untersucht haben. Screenshot: B. Vogel

Fachbeitrag zu den Erkenntnissen aus einem Forschungsprojekt im Bereich Elektrizitätstechnologien, das vom Bundesamt für Energie finanziell unterstützt wurde. Der Beitrag ist unter anderem im Fachmagazin eTrends (Ausgabe Oktober 2026) erschienen.

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Bundesamt für Energie BFE

Dank maschinenbasierten Dialogsystemen wie ChatGPT, Gemini, Copilot, Grok oder Claude ist künstliche Intelligenz heute in aller Munde. Eine kurze Anfrage (‹Prompt›) zu jedem möglichen Thema genügt, und schon füllen sich die Zeilen. Viele Antworten erscheinen verblüffend plausibel und werden von Nutzerinnen und Nutzern gern akzeptiert. Und dies, obwohl sie von einer Maschine stammen, genauer genommen von einem Sprachmodell (Large Language Modells), das durch Auswertung unzähliger Texte in die Lage versetzt wurde, selbst sinnvolle Texte zu formulieren. Welcher Rechenaufwand für die Beantwortung eines Prompts nötig ist und welcher Stromverbrauch damit einhergeht, bleibt den meisten Userinnen und Usern verborgen.

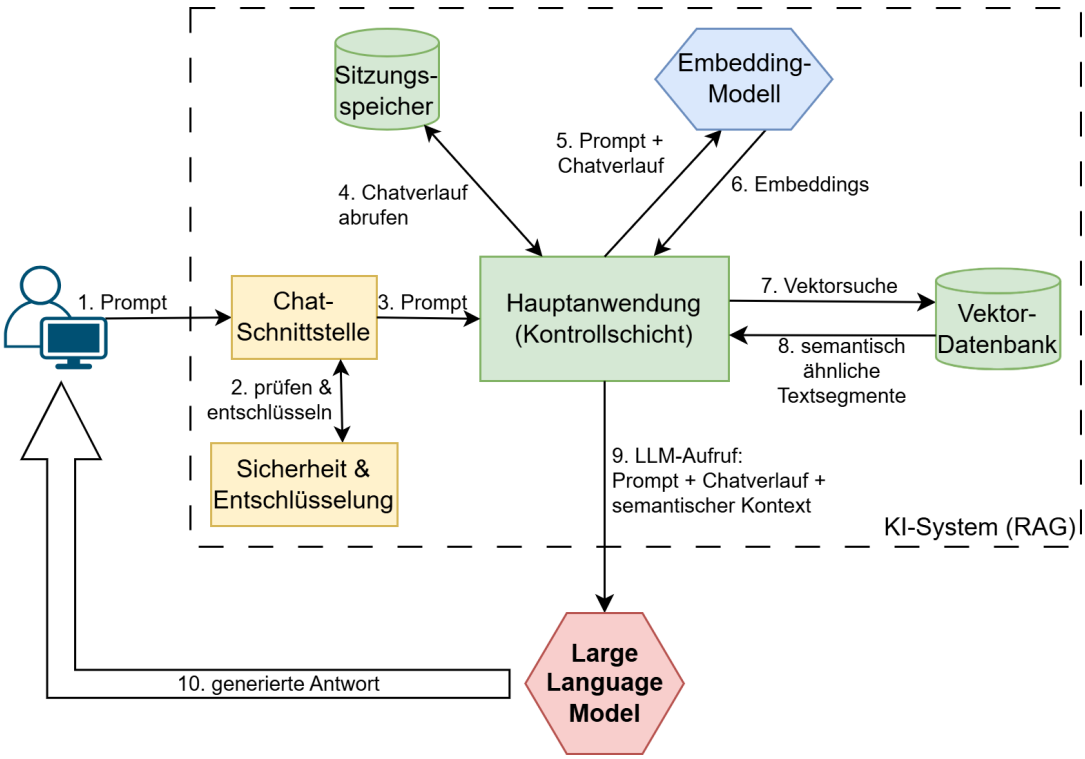
Mehr als nur das Sprachmodell

Auch für Fachpersonen ist es nicht einfach, den Stromverbrauch von Chatbots zu quantifizieren. «Bei grossen Sprachmodellen, auf denen ChatGPT und Co. beruhen, wissen nur die Anbieter, wie viel Strom das Training der Sprachmodelle gekostet hat und wie viel Strom für die Beantwortung eines Prompts erforderlich ist», sagt Vlad Coroama. Der ausgebildete Computerwissenschaftler und Inhaber des Roegen Centre for Sustainability (Zürich) untersucht Fragen rund um den Energieverbrauch von Digitalisierung. Im Auftrag des BFE hat er verschiedene Aspekte des Stromverbrauchs generativer KI erforscht.

In der Studie mit dem Akronym SAFE-AI hat Coroama gemeinsam mit Daniel Schien (University of Bristol) einen methodischen Rahmen erarbeitet, der dabei hilft, den Stromverbrauch von KI-Anwendungen besser zu verstehen und zu quantifizieren. KI-Sprachmodelle konsumieren Strom während des Trainings und später bei der Beantwortung der Prompts. Doch nicht nur das KI-Modell braucht Strom, sondern auch das KI-System. Das KI-System, so wie es Coroama versteht, umfasst alle IT-Komponenten, die nötig sind, damit das Sprachmodell eine Nutzeranfrage beantworten kann. Dazu gehören verschiedene interne Datenbanken, zum Beispiel ein Speicher für die Inhalte früherer Chats, auf die bei neuen Prompts zurückgegriffen wird. Dazu gehören aber auch Sicherheits- und Entschlüsselungs-Komponenten oder ein Werkzeug, das im Zuge von Prompts bedeutungsgleiche Wörter (Synonyme) generiert.

Fallstudie aus Zürich

Die Komponenten des KI-Systems können für den Stromverbrauch einer Chatanfrage wichtiger sein als der Verbrauch des Sprachmodells für sich genommen. Das veranschaulicht eine Fallstudie, die die beiden Forscher mit Unterstützung der Stadt Zürich durchführten: Im März 2025 nahm die Stadt den Nachhaltigkeits-Chatbot Zü-Re in Betrieb. Dieser beantwortet Fragen aus der Bevölkerung rund um das Thema Kreislaufwirtschaft: Er sucht für die Userin zum Beispiel den nächst-



Schematische Darstellung des KI-Modells (Large Language Model/Grosses Sprachmodell) und des KI-Systems. Letzteres umfasst die Komponenten, die benötigt werden, um die Benutzeranfrage (Prompt) durch das KI-Modell beantworten zu lassen. Illustration: V. Coroama

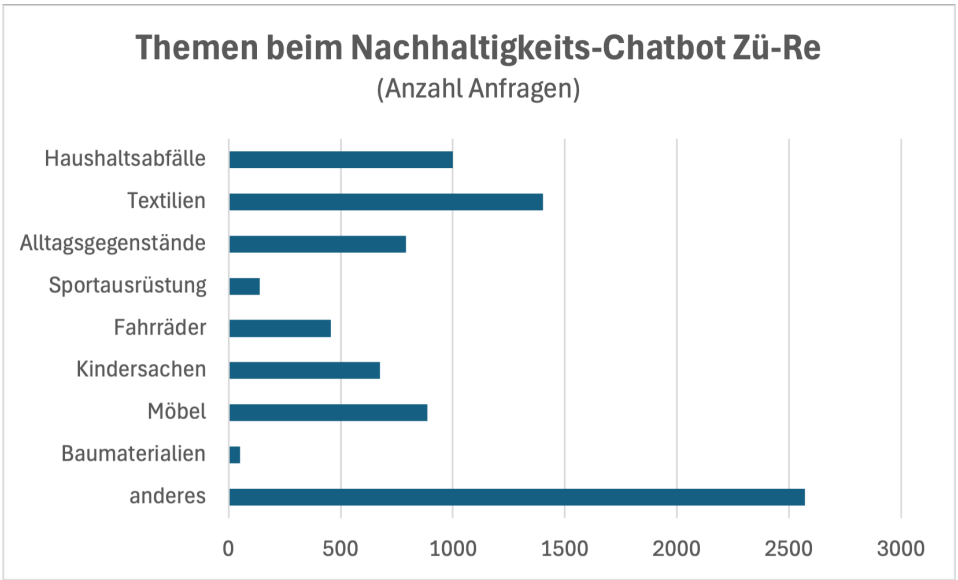
Rechenzentren sind durch KI-Anwendungen zu einem Gegenstand öffentlicher Diskussionen geworden. Doch auch unhabhängig von KI brauchen Unternehmen z.B. aus der Finanz- oder Pharmabranche oder Forschungseinrichtungen wie das CERN (Foto) weiterhin viel Rechenkapazität für die Datenverarbeitung. Foto: Shutterstock

gelegenen Secondhand-Laden oder führt den User zu einer auf E-Bikes spezialisierten Fahrradwerkstatt. Im Rahmen von SAFE-AI werteten die Forschenden rund 1'100 Chats aus, die der Chatbot im Juli 2025 beantwortet hat. Jeder Chat bestand in der Regel aus mehreren Prompts und den zugehörigen Antworten.

Dabei zeigte sich: Vom Stromverbrauch der 1'100 Chats entfallen 99 % auf das KI-System, also die elektronischen Komponenten, die im Colocation-Rechenzentrum bei einem externen Dienstleister in der Umgebung von Zürich untergebracht sind, der das System hostet. Der Strom wird dort unter anderem benötigt, um auf Servern Dokumente mit lokalen Informationen zur Stadt Zürich vorzuhalten. Auf diese Daten greift der Chatbot Zü-Re zurück, wenn er Anfragen zu lokalen Themen beantwortet: Formuliert eine Zürcher Userin einen Prompt, werden diese lokalen Informationen zusammen mit dem Prompt an das Sprachmodell übermittelt, das sich physisch heute in der Regel in einem US-Rechenzentrum befindet. Dort wird der Prompt unter Einbezug der lokalen Kontextinformationen verarbeitet und die Antwort anschliessend an die Userin in Zürich übermittelt. Nur rund 1 % des Stromverbrauchs, die diese Chatbot-Anfrage verursacht, entfallen auf das Sprachmodell im US-Rechenzentrum. Der grosse Rest geht auf der IT-Infrastruktur im Raum Zürich.

Eine halbe Kilowattstunde pro Chat

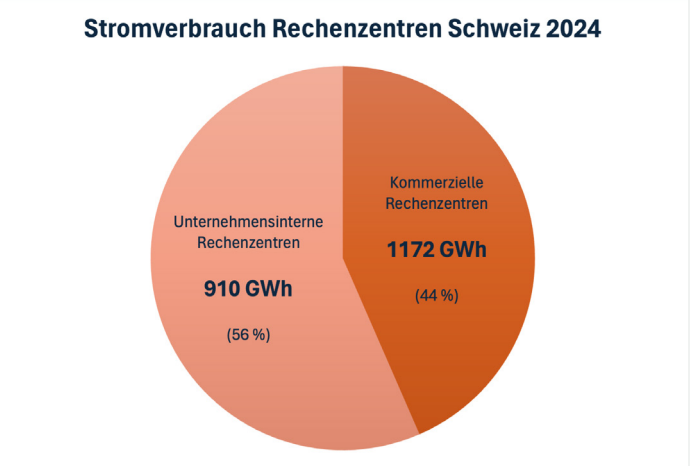
Gemäss den Berechnungen der beiden Forscher verbrauchten die 1'107 Chats im Juli 2025 insgesamt 550 kWh Strom, also rund 0,5 kWh pro Chat. Ein Chat, der jeweils aus einem



Die Anfragen beim Chatbot Zü-Re berühren eine breite Palette von Themen. Grafik: SAFE-AI/B. Vogel

WACHSENDER STROMBEDARF VON RECHENZENTREN

Rechenzentren für Anwendungen der Künstlichen Intelligenz (KI) werden stark ausgebaut. Die vier US-Tech-Unternehmen Amazon, Meta, Microsoft und Alphabet – auch als Hyperscaler bezeichnet – wollen 2026 laut Medienberichten 700 Mrd. US-Doller investieren, vor allem in Rechenzentren, in denen KI-Modelle trainiert und betrieben werden. Um den Strombedarf zu decken, dienen unter anderem leistungsstarke Gaskraftwerke. Auch Deutschland plant Gaskraftwerke in erheblichem Umfang, um den Stromhunger neuer Rechenzentren zu decken.



In einer vom BFE beauftragten Studie hat das Beratungsunternehmen TEP Energy den Stromverbrauch von Schweizer Rechenzentren für das Jahr 2024 auf 2'082 GWh (bzw. 2,1 TWh) geschätzt. Das entspricht 3,6 % des Schweizer Stromverbrauchs. Grafik: B. Vogel

Die Schweiz verfügt gemessen an der Bevölkerung über eine hohe Dichte an Rechenzentren. Diese dienen bislang aber nicht dem Training bzw. Betrieb von KI-Modellen, sondern gehören zur IT-Infrastruktur herkömmlicher Unternehmen. Eine im Frühjahr 2026 publizierte Studie im Auftrag des BFE beziffert den Stromverbrauch der Schweizer Rechenzentrum bezogen auf das Jahr 2024 mit 2,1 TWh (3,6 % des Schweizer Stromverbrauchs). Gegenüber dem Jahr 2019 entspricht das einer Zunahme um 18 %. Künftig dürfte der Stromverbrauch weiter kräftig wachsen: Bis 2030 rechnet die Studie mit einer Zunahme um 20 bis 50 Prozent, im Extremfall sogar um 70 Prozent.

Unternehmensinterne Rechenzentren verlieren unterdessen an Bedeutung, weil Unternehmen ihre IT vermehrt auslagern. Dabei nutzen sie die Rechenleistung von Cloud-Anbietern, oder sie betreiben ihre firmeneigene IT-Hardware in sogenannten Colocation-Rechenzentren, die keine Rechenleistung bereitstellen, aber die nötige Rechenzentren-Infrastruktur wie Stromversorgung, Kühlung, Brandschutz oder Monitoring. Firmenexterne Rechenzentren gewinnen an Bedeutung und konsumieren auch in der Schweiz einen immer grösseren Anteil des Stroms. Diese kommerziellen Rechenzentren haben teilweise grosse IT-Unternehmen (Hyperscaler) als Kunden, die in der Schweiz keine eigenen Rechenzentren besitzen.

Der Stromverbrauch von Rechenzentren ist durch verschiedene Untersuchungen schon recht gut erfasst. Eine ebenfalls im Auftrag des Bundesamts für Energie durchgeführte, neue Studie des Roegen Centre for Sustainability unter dem Titel ‹Synergien und Zielkonflikte der Kreislaufwirtschaft in Rechenzentren› (CEST-DC) hat den Fokus nun auf die Nachhaltigkeit von Rechenzentren erweitert. Die Untersuchung behandelte Fragestellungen rund um Energieeffizienz, Wasser- und Materialverbrauch sowie Werkstoffkreisläufe. Zudem analysierte sie die Auswirkungen von Rechenzentren auf das Stromnetz.

↗ Der englischsprachige **Schlussbericht** zum Projekt ‹Circular Economy Synergies and Trade-offs in Data Centres› (CEST-DC) ist hier abrufbar.

Die Studie CEST-DC hat verschiedene Nachhaltigkeitsaspekte von Rechenzentren untersucht. Im Bild: Das Rechenzentrum von Swisscom in Zürich-Herdern, das seit 2012 Wärme an einen regionalen Wärmeverbund liefert. Foto: Shutterstock

oder mehreren Prompts inkl. Antworten besteht, braucht somit etwa gleich viel Strom wie 50 E-Bike-Kilometer oder fünf Stunden Fernsehen. In diesen Werten ist der Verbrauch des KI-Modells (einschliesslich Training) enthalten, ebenso der Verbrauch des gesamten KI-Systems (einschliesslich Hardware von Rechenzentren und Usern, und einschliesslich graue Energie, die für die Herstellung von Hardware-Komponenten benötigt wurde).

Diesem Verbrauch sind die positiven Effekte des Nachhaltigkeits-Chatbots gegenüberzustellen. Die Forscher hätten diesen positiven Effekt gern quantifiziert, bekamen auf eine entsprechende Umfrage unter den Zü-Re-Nutzerinnen und -Nutzern aber nicht genügend Rückmeldungen. Trotzdem sieht es Vlad Coroama positiv: «0,5 kWh Strom pro Chat klingt nach viel. Wenn jemand dank der Chatbot-Auskunft aber sein Velo reparieren lässt statt ein neues zu kaufen, kann sich das ökologisch schnell lohnen. Das gilt insbesondere auch für die CO_2-Bilanz, da der Chatbot in der Schweiz Strom mit einem geringen CO_2-Abdruck nutzt.»

Starke Nutzung senkt Verbrauch pro Anfrage

Dass der Stromverbrauch von Zü-Re – wie oben dargestellt – so ungleich auf die Schweiz und die USA verteilt ist, gilt nicht prinzipiell für alle Chatanfragen, die hierzulande gemacht werden. Der relativ hohe lokale Stromverbrauch von Zü-Re rührt nämlich daher, dass der Chatbot in den ersten Monaten nach Inbetriebnahme und bedingt durch die Sommerpause relativ wenig genutzt wurde. Im Untersuchungszeitraum gab es lediglich durchschnittlich 36 Chats pro Tag. Der Stromverbrauch des KI-Systems ist im Fall von Zü-Re ziemlich konstant und hängt nur unwesentlich von der Zahl der Anfragen ab. Da es nur wenige Anfragen gab, war der Stromverbrauch pro Anfrage sehr hoch. Werden Zü-Re oder ähnliche Systeme in Zukunft stärker genutzt, wird der Stromverbrauch des KI-Systems pro Anfrage zurückgehen. Damit sinkt auch der Stromverbrauch einer Anfrage bei Zü-Re insgesamt, und das KI-Modell wird daran einen grösseren Anteil haben.

Vor diesem Hintergrund steht für Vlad Coroama fest: «Um den Stromverbrauch von Chatbots zu beurteilen, müssen wir das Gesamtsystem betrachten, nicht nur das Sprachmodell, das irgendwo auf einem Server in Kalifornien oder Irland liegt. Lokale KI-Systeme können den Energieverbrauch dominieren, zumal wenn sie unterbenutzt sind.» Der Einbezug ihres KI-Systems ist unter anderem wichtig für das Nachhaltigkeits-Reporting von Unternehmen. Wollen sie ihren ökologischen Fussabdruck ausweisen, müssen sie auch den Stromverbrauch des KI-Systems berücksichtigen, nicht nur jenen der benutzten KI-Modelle.

Lange Prompts kosten viel Energie

Die beiden Forscher entwickeln in ihrer Studie einen Ansatz, um den generellen Stromverbrauch von Chatbot-Anfragen zu quantifizieren. Nach ihrer Beobachtung ist die Anzahl der Eingabe- und Ausgabetokens der aussagekräftigste Parameter zur Abschätzung des Energieverbrauchs. Tokens sind – vereinfacht ausgedrückt – die Silben, aus denen ein Prompt bzw. der Antworttext des Chatbots zusammengesetzt sind. Es gilt: Der Energieverbrauch des Ausgabetextes wächst mit der Zahl der Tokens linear, der Energieverbrauch des Eingabetexts im Quadrat (allerdings von einem niedrigen Wert startend).

Bei kurzen Prompt-Anfragen wird der Energieverbrauch somit hauptsächlich durch die Länge des Ausgabetextes bestimmt. Die Zusammenfassung eines Eingabetextes braucht also weniger Energie als dessen Übersetzung. Bei längeren Prompt-Anfragen hingegen fällt dann auch die Länge des Eingabetextes ins Gewicht – und ab einer Länge von ca. 100'000 Tokens wird diese sogar dominierend. Integriert ein User in den Prompt z.B. ein 300-seitiges PDF, nimmt der Stromverbrauch stark zu. Vergleichsweise viel Strom braucht auch die Herstellung von KI-generierten Bildern oder Videos, weil diese aus grossen Ausgabedateien bestehen.

INTERNATIONALER AUSTAUSCH

SAFE-AI und CEST-DC liefern unter anderem methodische Grundlagen, um den Strom- bzw. Ressourcenverbrauch von KI-Anwendungen und Rechenzentren korrekt zu bestimmen. «Die beiden Studien sind nicht nur für die Schweiz relevant, sondern sie liefern auch einen wichtigen Input an die internationale Forschercommunity sowie internationale Regulierungs-Behörden», sagt Roland Brüniger, externer Leiter des BFE-Forschungsprogramms Elektrizitätstechnologien. Brüniger vertritt die Schweiz im Technology Collaboration Programme (TCP) 4E der internationalen Energieagentur (IEA). Im TCP 4E arbeiten Fachpersonen aus 14 Ländern zusammen, um sich unter anderem über Fragen des Energieverbrauchs und der Energieeffizienz in verschiedenen Anwendungsbereichen auszutauschen und mit den Ergebnissen politische Entscheidungsträger zu unterstützen.

OpenAI und Google veranschlagen übrigens für einen typischen Prompt einen Stromverbrauch von 0,3 Wh. Das sind mehr als tausendmal weniger als die oben genannten 0,5 kWh pro Zü-Re-Chat. Dies ist vor allem auf die geringe Nutzung – und daher Ineffizienz – des lokalen KI-Systems im Falle von Zü-Re zurückzuführen. Wäre beispielsweise das System 100-fach intensiver genützt worden, so würde es lokal gleich viel Energie verbrauchen, da die Kapazität noch längst nicht ausgeschöpft war. Pro Chat würde die Energie damit 100-mal geringer ausfallen und ca. 5 Wh pro Chat betragen. Zudem beziehen sich die Angaben von OpenAI und Google auf einen Prompt, nicht auf einen Chat (der in der Regel aus mehreren Prompts besteht). Ferner sind die Verbräuche bei OpenAI und Google enger definiert, beinhalten zum Beispiel nicht die graue Energie zur Herstellung der IT-Komponenten.

- Der englischsprachige **Schlussbericht** zum Projekt ‹Sustainability Assessment Framework for the Environmental Impacts of Artificial Intelligence› (SAFE-AI).

- **Auskünfte** erteilt Roland Brüniger, externer Leiter des BFE-Forschungsprogramms Elektrizitätstechnologien.

- Weitere **Fachbeiträge** über Forschungs-, Pilot-, Demonstrations- und Leuchtturmprojekte im Bereich Elektrizitätstechnologien.

Autor: Benedikt Vogel, im Auftrag des Bundesamts für Energie (BFE)
Stand: Oktober 2026