basebox
KI-Plattform

Nachhaltigkeitsbericht basebox GmbH

Berichtsjahr 2026 · Erster Nachhaltigkeitsbericht des Unternehmens | Stand: 03.09.2026, aktualisiert am 08.10.2026

1. Strategie

basebox entwickelt die Rechtssichere KI-Plattform für Healthcare & Government. Die Mission – KI in regulierte Organisationen zu bringen, sicher, beherrschbar und bezahlbar – bedeutet, KI unter voller Kontrolle nutzbar zu machen. Verantwortungsvoller Umgang mit Ressourcen ist Teil dieses Selbstverständnisses: Sicherheit, Kontrolle und Nachhaltigkeit sind für basebox keine Gegensätze, sondern greifen ineinander.

Dieser Bericht ist der erste Nachhaltigkeitsbericht von basebox. Er schafft die Grundlage, um Fortschritte in den kommenden Jahren nachvollziehbar zu machen. basebox verzichtet bewusst auf komplexe Zielmarken, die zum jetzigen Zeitpunkt nicht belastbar unterlegt werden können, und legt stattdessen transparent dar, was heute bereits umgesetzt ist und wo Verbesserungspotenzial besteht.

Der Bericht behandelt zwei Bereiche getrennt: den Unternehmensbetrieb von basebox (Abschnitt 3) und den Betrieb der KI-Plattform einschließlich der genutzten Rechenzentren und KI-Modelle (Abschnitt 4). Der zweite Bereich ist für Kunden der relevantere, weil dort der überwiegende Teil des Energieverbrauchs entsteht.

Ein zentraler Aspekt: Die technologischen Grundsatzentscheidungen von basebox – der vollständige Einsatz von Rust als Implementierungssprache sowie intelligentes Model-Routing beim Einsatz von KI-Modellen – wurden primär aus Sicherheits- und Effizienzgründen getroffen. Beide erzeugen einen ökologischen Zusatznutzen, der in Abschnitt 5 beschrieben wird.

2. Prozessmanagement

Die Verantwortung für Nachhaltigkeit liegt bei René Herzer, Geschäftsführer und Gründer von basebox.

Folgende Richtlinien sind seit Gründung 2022 etabliert:

- **Bürostrom:** 100 % Ökostrom (EWS Schönau) seit Aufnahme der Geschäftstätigkeit 2022
- **Reisen:** Geschäftsreisen im DACH-Raum grundsätzlich mit der Bahn; Car-Sharing (Mitgliedschaft in Utting am Ammersee) nur in Ausnahmefällen
- **Fuhrpark:** Kein eigener Fuhrpark, außer Fahrrädern
- **Homeoffice:** Alle Mitarbeitenden können im Homeoffice arbeiten
- **Arbeitszeit:** Flexible Zeiteinteilung nach individueller Leistungsfähigkeit; Überstunden werden nicht empfohlen
- **Rechenzentrumsauswahl:** Bei der Auswahl von Rechenzentren und Modellanbietern werden Energiemix, PUE-Wert und vorliegende Zertifizierungen als Kriterien berücksichtigt

3. Unternehmensbetrieb

3.1 Bürobetrieb

Das Büro in Utting am Ammersee (ca. 150 qm) wird vollständig mit Ökostrom des Anbieters EWS Schönau betrieben – seit Unternehmensgründung 2022.

3.2 Mobilität

Geschäftsreisen finden überwiegend per Bahn statt: ca. 2 Reisen pro Monat mit je 2 Personen quer durch den DACH-Raum. Hinzu kommen ca. 4 Flugreisen pro Jahr zwischen Irland und Deutschland, bedingt durch den Head of Engineering aus Irland. Ein eigener Fuhrpark existiert nicht.

3.3 Ressourcen

Der Papierverbrauch ist minimal, da im Arbeitsalltag nur selten gedruckt wird. Die Mülltrennung erfolgt nach deutschem Standard.

4. Betrieb der KI-Plattform

Dieser Abschnitt betrifft die angebotene Leistung selbst: die Rechenzentren, in denen die Plattform und die KI-Modelle betrieben werden, sowie den Energieverbrauch und die Emissionen, die bei der Nutzung entstehen.

4.1 Genutzte Rechenzentren

Die Plattform wird ausschließlich in Rechenzentren innerhalb des Europäischen Wirtschaftsraums betrieben. Für jeden Standort sind der PUE-Wert (Power Usage Effectiveness), der Anteil erneuerbarer Energien und die vorliegenden Zertifizierungen ausgewiesen.

- **Cloud: Plattform und Eigenbetrieb des KI-Modells in basebox (ab 01.11.2026) —** noris network AG, Rechenzentrum München-Ost (Aschheim bei München), Server des Hardware-Partners FAST LTA. PUE unter 1,2 · 100 % erneuerbare Energien (CMS-Standard „Erzeugung EE") · ISO/IEC 27001 auf Basis IT-Grundschutz, ISO 14001, KyotoCooling
- **KI-Modell in basebox bis 31.10.2026, danach zusätzliche Rechenleistung —** STACKIT GmbH & Co. KG, Deutschland. PUE unter 1,3 (flottenweit) · 100 % erneuerbare Energien · BSI C5 Typ 2, ISO/IEC 27001, ISO 50001, EN 50600
- **Option auf Wunsch des Kunden statt STACKIT —** Mistral AI SAS, Frankreich. Keine veröffentlichten Kennzahlen des Anbieters
- **App mit externem Modell, nur bei Freischaltung —** Google Cloud (Vertex AI), Frankfurt (europe-west3). Flotten-PUE 1,09 (TTM 2025) · 68 % CO_2-freie Energie (CFE, 2024) · BSI C5:2020-Testat, ISO/IEC 27001, ISO 50001
- **Demoversion —** Hetzner Online GmbH, Deutschland (Nürnberg). PUE 1,10 bis 1,16 · 100 % erneuerbare Energien · ISO/IEC 27001
- **Websuche, nur bei Freischaltung —** Staan.ai (European Search Perspective SAS), EU. Keine veröffentlichten Kennzahlen des Anbieters, keine Zertifizierung

Veränderung zum 01.11.2026: Ab diesem Zeitpunkt betreibt basebox das KI-Modell in basebox selbst im Rechenzentrum München-Ost der noris network AG, auf Servern von FAST LTA. Damit verlagert sich der größte Teil der Modellverarbeitung von externen Anbietern in den Eigenbetrieb. Der Standort wurde unter anderem wegen des Energiemixes, des PUE-Werts und der KyotoCooling-Technologie ausgewählt.

Transparenzhinweis zu Mistral AI und der Websuche: Für diese beiden Dienste liegen keine veröffentlichten Rechenzentrumskennzahlen vor. basebox weist dies aus, statt Schätzwerte anzugeben. Beide Dienste werden innerhalb der EU erbracht. Mistral AI wird nur eingesetzt, wenn ein Kunde es wünscht; die Websuche ist nur aktiv, wenn sie ausdrücklich eingeschaltet wird.

4.2 Energieverbrauch und Emissionen pro Anfrage

Die folgenden Angaben weisen den geschätzten Energieverbrauch beziehungsweise CO_2-Ausstoß je Anfrage für die eingesetzten Modelle aus. Alle Werte beziehen sich auf eine Antwort von rund 400 Token.

- **KI-Modell in basebox, Eigenbetrieb ab 01.11.2026 —** Pro Anfrage: ca. 0,0042 kWh. Daraus ergeben sich 0,15 g CO_2eq bei Ökostrombetrieb und 2,08 g CO_2eq bei deutschem Standardnetzstrom. Modelltraining: offene Gewichte, Trainings-CO_2 nicht veröffentlicht
- **KI-Modell in basebox bei STACKIT, Deutschland —** Pro Anfrage: ca. 0,11 g CO_2eq. Modelltraining: offene Gewichte, Trainings-CO_2 nicht veröffentlicht
- **Option Mistral AI, Frankreich —** Pro Anfrage: ca. 1,14 g CO_2eq (Hersteller-LCA). Modelltraining: 20,4 kt CO_2eq über den Lebenszyklus von 18 Monaten. Der Wert stammt aus der Hersteller-LCA (Mistral, 2025) und dient als konservativer Referenzwert der Modellfamilie.
- **App mit externem Modell, Google Vertex AI, Frankfurt —** Pro Anfrage: ca. 1,24 g CO_2eq. Modelltraining: vom Hersteller nicht veröffentlicht

Wie diese Werte zustande kommen

Die Angaben sind Schätzungen auf Grundlage von Energieverbrauch, PUE-Wert des Rechenzentrums und Emissionsfaktor des bezogenen Stroms. Die Rechnung folgt durchgängig dem Schema:

```
Energieverbrauch der Anfrage × PUE des Rechenzentrums × Emissionsfaktor des Strommixes
```

Im Einzelnen:

- **KI-Modell in basebox, Eigenbetrieb:** ca. 0,0042 kWh, entsprechend rund 15 Sekunden Rechenzeit auf zwei H200 und einer RTX PRO 6000. Es sind zwei Werte ausgewiesen, weil der Emissionsfaktor je nach Strombezug erheblich abweicht.
- **KI-Modell in basebox bei STACKIT:** ca. 0,003 kWh × PUE 1,2 × 30 g/kWh. Der niedrige Emissionsfaktor ergibt sich aus dem tatsächlichen Ökostrombezug am Standort und erklärt den deutlichen Unterschied zu den anderen Modellen.
- **App mit externem Modell (Google):** ca. 0,003 kWh × PUE 1,09 × 380 g/kWh. Für den Emissionsfaktor wird der deutsche Netzmix angesetzt und nicht der bilanzielle Ausgleich des Anbieters, weil am Standort Frankfurt physisch Netzstrom anliegt. Das ist die konservativere Annahme.

Einordnung, was diese Zahlen leisten und was nicht

Die Werte sind Schätzungen, keine gemessenen Verbrauchsdaten. Belastbare, unabhängig verifizierte Zahlen zum Energieverbrauch von LLM-Inferenz sind branchenweit nur begrenzt verfügbar; die meisten Modellanbieter veröffentlichen weder Trainings- noch Inferenzdaten. basebox legt daher die Berechnungsgrundlage offen, damit die Zahlen nachvollziehbar und überprüfbar sind, statt sie als gemessene Werte darzustellen.

Der tatsächliche Verbrauch einer einzelnen Anfrage schwankt erheblich – mit der Länge der Eingabe, der Länge der Antwort, der Nutzung von Reasoning-Funktionen und der Frage, ob zusätzliche Datenquellen abgefragt werden. Die Werte oben beschreiben einen typischen Fall, keine Obergrenze.

basebox baut ein eigenes Energie- und Auslastungsmonitoring auf, um diese Schätzungen schrittweise durch gemessene Werte zu ersetzen. Ab dem Eigenbetrieb im November 2026 ist das für die selbst betriebene Inferenz technisch möglich.

5. Effizienz durch Produktentscheidungen

Zwei technologische Entscheidungen prägen den Energiebedarf der Plattform.

5.1 Rust als alleinige Implementierungssprache

Die Plattform ist zu 100 % in Rust entwickelt. Diese Entscheidung wurde primär aus Sicherheits- und Zuverlässigkeitsgründen getroffen – zentrale Anforderungen im Umgang mit kritischen Daten aus Gesundheitswesen und öffentlicher Verwaltung – bringt aber Effizienzvorteile mit sich:

- Native Kompilierung zu Maschinencode ohne Laufzeitinterpreter oder virtuelle Maschine. Studien ordnen Rust energetisch auf einem Niveau mit C/C++ ein, bei zusätzlicher Speichersicherheit.
- Kein Garbage Collector, dadurch entfallen wiederkehrende CPU- und Energiezyklen im Dauerbetrieb.
- Geringerer Ressourcenbedarf ermöglicht kleiner dimensionierte Hardware, reduzierten Kühlbedarf und einen kleineren Hardware-Fußabdruck.

Grundlage: Untersuchungen zur Energieeffizienz von Programmiersprachen der Universitäten Coimbra und Minho. Das konkrete Einsparpotenzial hängt von Implementierung und Auslastung ab.

5.2 Model-Routing und Modellauswahl

Nicht jede Anfrage benötigt das größte verfügbare Modell. Die Plattform ist darauf ausgelegt, Anfragen an das jeweils kleinste Modell zu leiten, das die Aufgabe zuverlässig löst:

- **Right-Sizing der Rechenleistung:** Einfache Aufgaben nutzen kleinere, ressourcenschonendere Modelle; nur komplexe Anfragen beanspruchen große Modelle.
- Vermeidung von Redundanz durch Caching bereits beantworteter oder ähnlicher Anfragen.
- Reduzierter Bedarf an energieintensiver Beschleuniger-Hardware für den Großteil des Datenverkehrs.

Wie groß der Effekt ist, zeigen die Werte in Abschnitt 4.2: Zwischen dem sparsamsten und dem verbrauchsintensivsten Textmodell liegt bei vergleichbarer Anfrage etwa der Faktor zehn. Die Wahl des Modells ist damit der wirksamste einzelne Hebel zur Reduktion des Energieverbrauchs im laufenden Betrieb.

Administratoren können festlegen, welche Modelle in ihrer Organisation zur Verfügung stehen, und dabei Energieeffizienz als Kriterium berücksichtigen.

6. Reporting gegenüber Kunden

basebox stellt Kunden jährlich ein Nachhaltigkeitsreporting zur Nutzung der Plattform bereit. Es umfasst:

- aggregierte Energie- und CO_2-Werte für die Nutzung der Plattform im Berichtszeitraum
- eine Aufschlüsselung nach eingesetzten Modellen und Diensten
- die aktuellen Rechenzentrumskennzahlen der genutzten Standorte: PUE-Wert, Energiemix, Zertifizierungen
- die Entwicklung gegenüber dem Vorjahr, sobald mehr als ein Berichtszeitraum vorliegt
- die im Berichtszeitraum umgesetzten und geplanten Maßnahmen zur Reduktion von Energieverbrauch und Emissionen

Das Reporting wird erstmals für das erste volle Vertragsjahr erstellt und danach jährlich fortgeschrieben.

Maßnahmen zur Reduktion

Umgesetzt:

- 100 % Ökostrom im Bürobetrieb seit 2022
- Auswahl von Rechenzentren mit hohem Anteil erneuerbarer Energien und niedrigem PUE-Wert
- Rust als Implementierungssprache, dadurch geringerer Rechenbedarf im Dauerbetrieb
- Caching zur Vermeidung redundanter Modellaufrufe
- Bahnbasierte Geschäftsreisen, kein eigener Fuhrpark

Geplant:

- Eigenes Energie- und Auslastungsmonitoring, um Schätzwerte durch gemessene Werte zu ersetzen (ab Eigenbetrieb November 2026)
- Ausbau des Model-Routings, um den Anteil der über kleine Modelle beantworteten Anfragen zu erhöhen
- Verlagerung weiterer Inferenz in den Eigenbetrieb mit vollständigem Ökostrombezug
- Ausweitung der Datengrundlage auf Anbieter, für die derzeit keine Kennzahlen vorliegen

7. Gesellschaft

7.1 Team-Struktur

basebox beschäftigt ein internationales Team, unter anderem mit Mitgliedern aus Deutschland, Brasilien und Indien (in Deutschland ansässig).

- **Vollzeit:** 4 Frauen (davon 3 in Tech-Funktionen), 3 Männer (davon 2 in Tech-Funktionen)
- **Teilzeit:** 1 Frau, 2 Werkstudenten (männlich)
- **Freie Mitarbeiter in Tech-Funktionen:** 2 Männer (über eigene Gesellschaften in Irland und Estland)

7.2 Arbeitsbedingungen

- Sehr niedriger Krankenstand, keine nennenswerten Krankheitsvorkommnisse
- Keine Arbeitsunfälle im Berichtszeitraum
- 30 Tage Urlaub pro Jahr
- Flexible Arbeitszeiteinteilung; Überstunden werden nicht empfohlen
- Durchgängige Homeoffice-Möglichkeit

7.3 Fluktuation

Im Berichtszeitraum haben zwei Mitarbeitende das Unternehmen verlassen: ein Senior Backend Developer (Wechsel in die Selbstständigkeit) und ein Full Stack Engineer.

7.4 Weiterbildung

Aktuell besteht kein gesondertes Weiterbildungsbudget. basebox prüft, dies mit wachsender Unternehmensgröße zu etablieren.

7.5 Vergütung

Die Vergütung erfolgt nach transparenten, funktionsbezogenen Kriterien. Alle Mitarbeitenden können ihre Arbeitszeit flexibel einteilen.

8. Unternehmensführung

8.1 Beschaffung

Arbeitsmittel werden teils als refurbished Geräte, teils neu beschafft. Neuanschaffungen sind im Produktkontext erforderlich, da aktuelle Hardware für Tests mit neuesten Geräten notwendig ist.

8.2 Compliance

Im Berichtszeitraum gab es keine rechtlichen Verstöße und keine Bußgelder.

9. Zusammenfassung

Strategie: Erster Nachhaltigkeitsbericht; Sicherheit und Effizienz als verbindendes Prinzip.

Prozesse: Verantwortung bei der Geschäftsführung; Energiemix und PUE als Auswahlkriterien für Rechenzentren.

Plattformbetrieb: Alle Rechenzentren innerhalb des EWR. Für vier von sechs genutzten Standorten liegen PUE-Werte zwischen 1,09 und unter 1,3 vor, drei davon mit 100 % Ökostrom (Ausnahme Google Cloud Frankfurt: 68 % CO_2-freie Energie). Energieverbrauch und Emissionen sind je Modell mit offengelegter Berechnungsgrundlage ausgewiesen.

Effizienz: Rust und Model-Routing als strukturelle Hebel. Zwischen dem sparsamsten und dem verbrauchsintensivsten Textmodell liegt etwa Faktor zehn.

Reporting: Jährliches Nachhaltigkeitsreporting für Kunden mit Aufschlüsselung nach Modellen, Rechenzentrumskennzahlen und Entwicklung im Zeitverlauf.

Unternehmen: 100 % Ökostrom im Büro seit 2022; bahnbasierte Mobilität; diverses Team; keine Vorfälle im Berichtszeitraum.

Verantwortlich für den Inhalt: René Herzer, Geschäftsführer, basebox GmbH. Rückfragen zu diesem Bericht: support@basebox.ai