Raus aus der Cloud: Wie offene KI-Modelle unseren NAO unabhängig machen

Snacktime Digitalisierung „NAO Upcycling Teil 2“ vom 21. Oktober 2026 – der Rückblick

Im Mai brauchte unser Roboter NAO für eine Antwort 16 bis 19 Sekunden, und jede Frage verließ das Haus. Heute denkt derselbe Roboter auf einem Laptop neben ihm: offene KI-Modelle, keine Cloud, keine laufenden Kosten, Antwort in rund zwei Sekunden. In der Snacktime hat Erik Dethier gezeigt, wie das geht, was dabei gehakt hat und was kleine und mittlere Unternehmen daraus mitnehmen können.

Im Mai: ein Roboter mit Gehirn in der Cloud

NAO ist ein humanoider Roboter von 2008: 58 Zentimeter groß, 25 Freiheitsgrade, der Hersteller Aldebaran ist seit Juli 2025 insolvent. In Teil 1 im Mai haben wir gezeigt, dass so ein Gerät mit moderner KI wieder sinnvoll sprechen kann. Damals saß sein Gehirn beim Anbieter: Spracherkennung über die Whisper-API, das Sprachmodell gpt-5-nano von OpenAI, die Stimme ebenfalls von OpenAI.

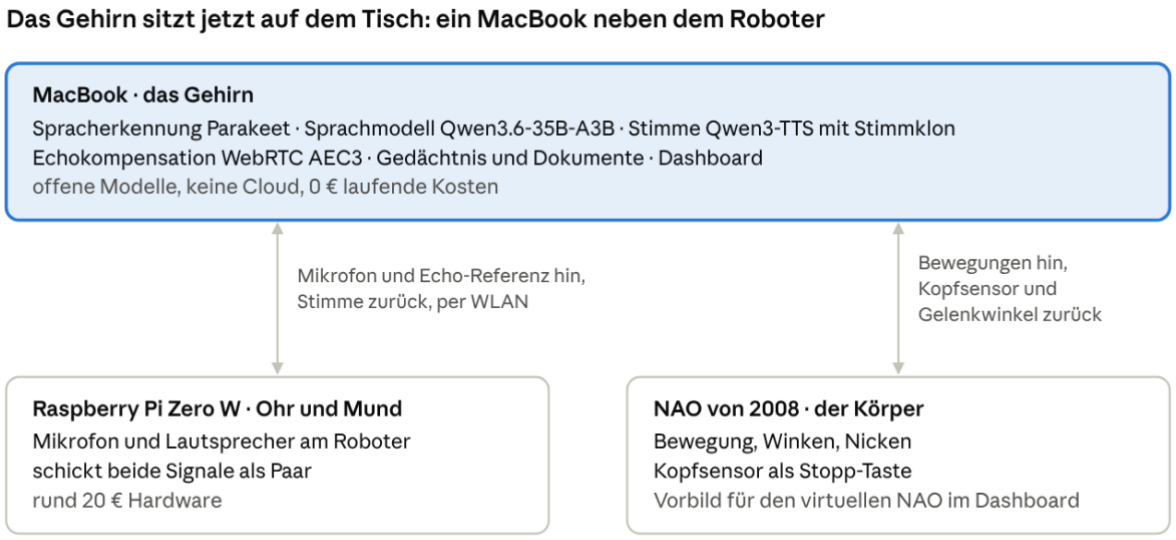
Das funktionierte, hatte aber seinen Preis. Eine Antwort dauerte 16 bis 19 Sekunden, jedes Gespräch kostete Geld, und jede Frage verließ das Haus. Daraus wurde die Frage für Teil 2: Geht das auch ohne Cloud, und zwar mit Modellen, die jeder herunterladen kann?

Die Frage ist für den Mittelstand alles andere als akademisch. Laut Bitkom nutzen 2026 bereits 57 Prozent der Unternehmen ab 20 Beschäftigten KI, ein Jahr zuvor waren es 36 Prozent. Zugleich nennen 66 Prozent den Datenschutz als Hemmnis, 45 Prozent fürchten, dass Daten abfließen. 85 Prozent halten Deutschland für zu abhängig von US-Cloud-Anbietern. Lokale KI setzt genau an dieser Hürde an: Kein Wort und kein Dokument verlässt den Raum.

Quellen: Bitkom, 14.09.2026 (57, 66 und 45 Prozent); Bitkom Cloud Report, 17.06.2026 (85 Prozent).

Heute: drei Schichten, alles im Raum

Der Aufbau hat dieselben drei Schichten wie im Mai, nur ist das Gehirn vom Rechenzentrum auf den Tisch gezogen. Ein MacBook erkennt die Sprache, formuliert die Antwort und spricht sie, ein Raspberry Pi für rund 20 Euro ist Ohr und Mund, NAO selbst ist der Körper.



NAOtonomous in drei Schichten · alles lokal

Alle drei Modelle sind offen: herunterladbar, austauschbar, ohne Konto beim Anbieter. Die Spracherkennung ist Parakeet, das Sprachmodell Qwen3.6-35B-A3B, die Stimme Qwen3-TTS mit dem Klon einer echten Stimme.

Eine unscheinbare Zutat ist die Echokompensation. Der Pi schickt das Mikrofonsignal und das, was gerade aus dem Lautsprecher kommt, als Paar zum Laptop. Der rechnet NAOs eigene Stimme heraus. Ohne das hört NAO sich selbst und hält seine eigenen Worte für eine Frage.

Das Ergebnis: Kurze Antworten kommen nach rund zwei Sekunden, lange nach drei bis vier, gemessen vom Ende der Frage bis zum ersten Ton. Lange Antworten brauchen länger, weil NAO sie erst ganz formuliert und dann am Stück spricht. So bleibt die Stimme über die ganze Antwort gleich.

Vier Szenen am Roboter

In der Snacktime haben wir vier kurze Szenen gezeigt, hier als Standbilder aus den Aufzeichnungen. Zwischen Frage und Antwort ist nichts geschnitten, die gemessene Antwortzeit steht im Bild.

1. Ohne Cloud



Szene 1: „Nao, wo läuft eigentlich dein Gehirn?“ – die Antwort kommt nach 1,9 Sekunden vom Laptop.

Die erste Frage ist zugleich die Pointe: „Alles läuft lokal auf meinem MacBook. Spracherkennung, Sprachmodell und Stimme – keine Cloud, keine Abhängigkeit vom Anbieter.“ Danach nennt NAO die Uhrzeit. Auch das ist ein Werkzeug, das ohne Internet auskommt.

2. Unterbrechen

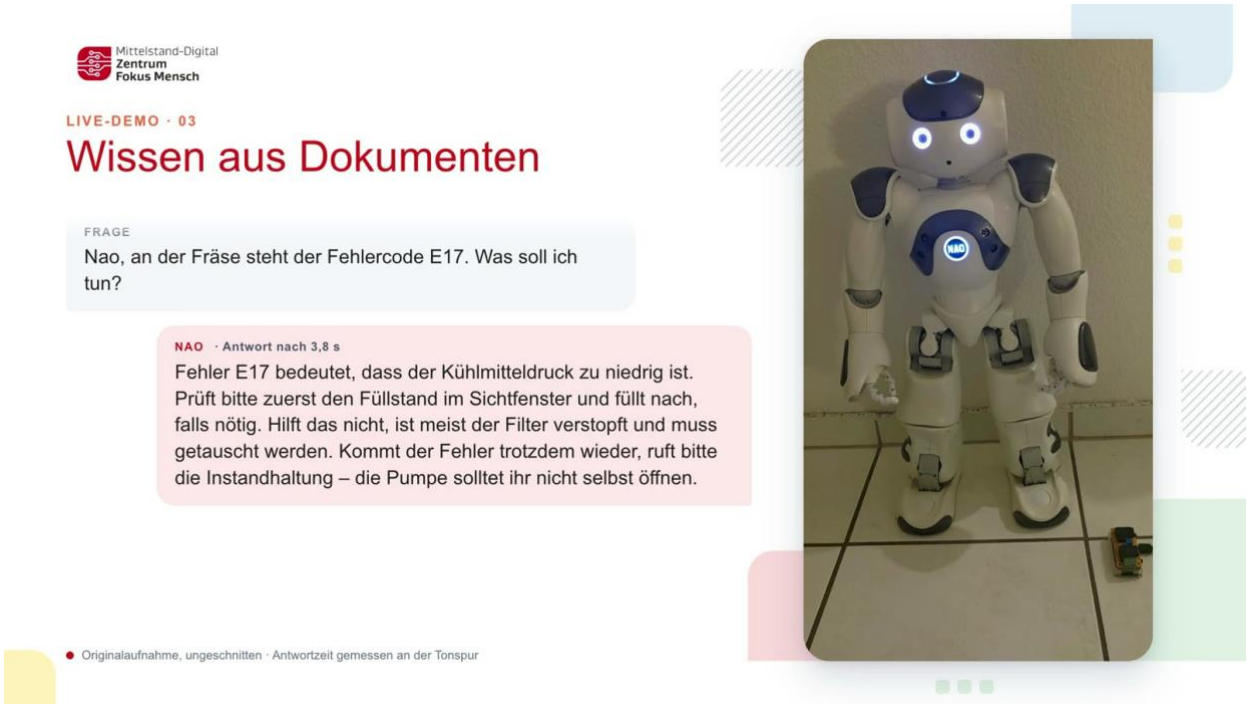


Szene 2: Einmal auf den Kopf getippt, und NAO hört mitten im Wort auf.

Wer etwas anderes wissen will, muss nicht warten, bis die Antwort zu Ende ist. Ein Zuruf wie „Stopp“ genügt, ebenso ein kurzes Antippen des Kopfsensors. Damit der Zuruf überhaupt

ankommt, muss der Laptop NAOs eigene Stimme aus dem Mikrofon herausrechnen. Sonst übertönt der Roboter jeden Menschen im Raum.

3. Wissen aus Dokumenten



Szene 3: Die Antwort auf Fehler E17 stammt aus dem Wartungsplan, der vorher im Dashboard abgelegt wurde.

Für die Demo haben wir einen Wartungsplan für eine CNC-Fräse erfunden und im Dashboard abgelegt. Auf die Frage „An der Fräse steht Fehler E17. Was soll ich tun?“ antwortet NAO: Der Kühlmitteldruck ist zu niedrig, erst den Füllstand prüfen, sonst den Filter tauschen. Auch „Wen rufe ich in der Spätschicht an?“ beantwortet er aus dem Dokument: Frau Yilmaz, Durchwahl 217.

Dahinter steckt mehr als ein abgelegtes PDF. NAO zerlegt das Dokument in Abschnitte und sucht nach Bedeutung und nach Wörtern zugleich. So findet er auch Fachbegriffe wie Fehlercodes oder Schmiermittel zuverlässig.

4. Gedächtnis



Szene 4: „Weißt du noch, was ich suche?“ – NAO erinnert sich an das frühere Gespräch.

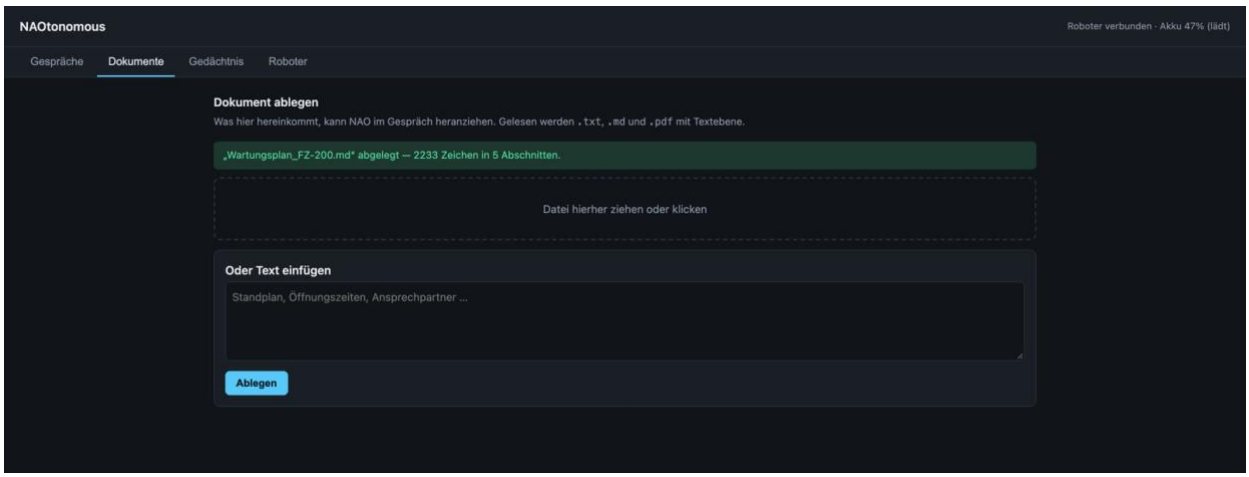
Erik stellt sich vor: Instandhaltungsleiter bei einem Metallbauer, auf der Suche nach einem Assistenten für Wartungsfragen, am Donnerstag kommt er mit seinem Chef wieder. Ist danach 90 Sekunden niemand mehr da, beendet NAO das Gespräch und notiert sich Rolle, Anliegen und Zusagen. Als Erik später fragt, ob NAO noch weiß, was er sucht, kommt die Antwort sofort.

Zwei Grenzen gehören zur Ehrlichkeit dazu: NAO erkennt niemanden an der Stimme, der Name muss fallen. Smalltalk merkt er sich absichtlich nicht.

Das Dashboard: sehen, was NAO weiß

Alles, was NAO hört, liest und behält, lässt sich in einem Dashboard im Browser einsehen. Es läuft auf demselben Laptop wie die KI, ohne Konto und ohne Server im Netz. Vier Ansichten:

Dokumente ablegen. Eine Datei wird ins Feld gezogen, Text, Markdown oder PDF mit Textebene. NAO zerlegt sie in Abschnitte und kann sie ab sofort im Gespräch heranziehen.



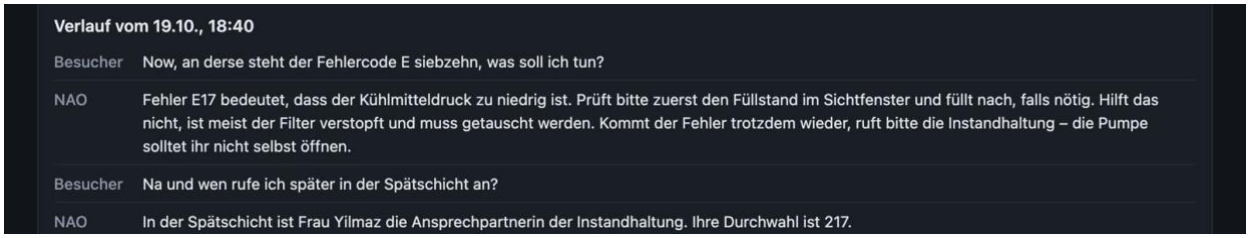
Der Wartungsplan aus Szene 3 im Dashboard: abgelegt in fünf Abschnitten.

Gedächtnis. Hier steht, was NAO behalten hat: Notizen aus Gesprächen und abgelegte Dokumente. Jede Zeile lässt sich löschen, das wirkt sofort.



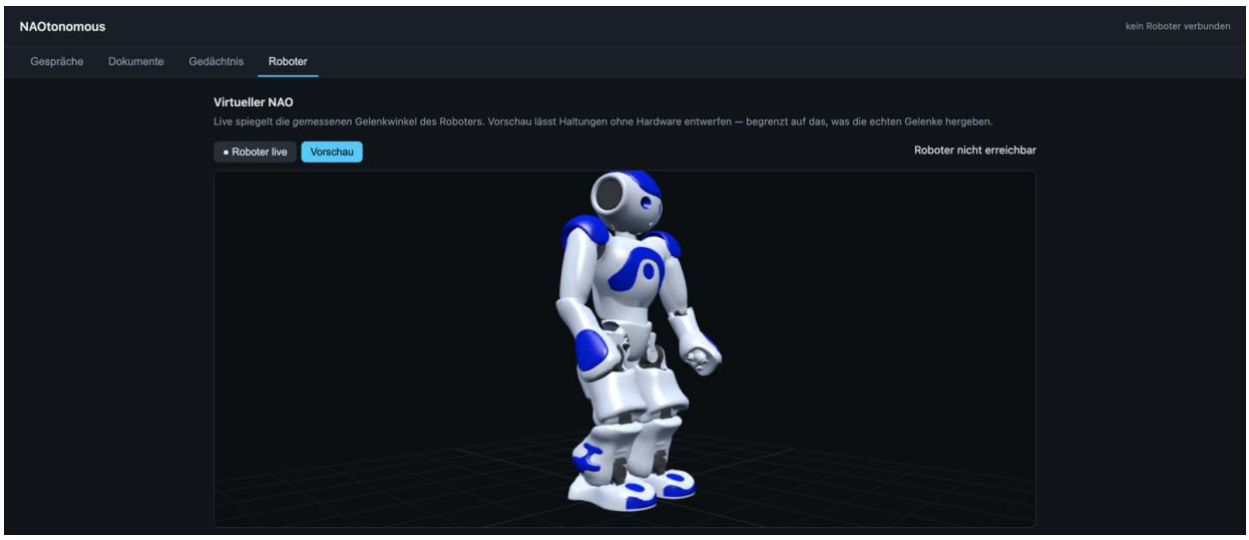
Was NAO aus dem Gespräch mit Erik behalten hat, darunter der Wartungsplan als ein Eintrag.

Gesprächsverläufe. Jedes Gespräch wird im Wortlaut abgelegt, nicht als Zusammenfassung. Man sieht auch, wo die Spracherkennung danebenlag: „Now" statt „Nao", „an derse" statt „an der Fräse". NAO hat die Frage trotzdem richtig verstanden.



Die Mitschrift von Szene 3, mit den Hörfehlern der Spracherkennung.

Virtueller NAO. Ein 3D-Modell zeigt live die gemessenen Gelenkwinkel des echten Roboters. Im Vorschaumodus lassen sich Haltungen auch ohne Hardware entwerfen, begrenzt auf das, was die echten Gelenke hergeben.



Der virtuelle NAO im Vorschaumodus.

Was gehakt hat

Die offenen Modelle selbst waren selten das Problem, die Nähte dazwischen schon: ein Mikrofon, das den eigenen Lautsprecher hört, eine Stimme, die nicht aufhört. Jede Änderung haben wir vorher und nachher gemessen, im Projekt stehen rund 80 Messberichte.

Was hakte	Was jetzt hilft
NAO hörte sich selbst. Er hielt die eigene Stimme für eine Frage und antwortete sich, in Endlosschleife.	Er erkennt seine eigenen Worte und verwirft sie.
Die Stimme redete 5½ Minuten weiter. In 3 von 60 Läufen hörte die Sprachausgabe nach dem Satz nicht auf.	Eine Grenze, abgeleitet aus der gemessenen Sprechdauer, bremst sie.
„Stopp“ kam nicht an. Am Mikrofon war NAOs Echo 10 dB lauter als ein Mensch.	Eine stärkere Echokompensation auf dem Laptop brachte den Zuruf von 0 auf 3 von 4 Versuchen.
Satz für Satz klang jeder Satz anders. So war NAO zwar schneller, aber die Stimme wechselte mitten in der Antwort.	NAO spricht die Antwort am Stück. Das kostet rund 0,7 Sekunden, dafür bleibt die Stimme gleich.

Was der Mittelstand daraus mitnimmt

Der Wartungsplan aus der Demo ist das Muster: Was heute in Ordnern oder PDFs steht, kann ein lokaler Assistent per Sprache beantworten, ohne dass ein Dokument das Haus verlässt. Drei Anwendungen liegen nahe:

1. **Wissensassistent für die Werkhalle.** Wartungspläne, Betriebsanweisungen, Fehlercodes: per Sprache gefragt, aus den eigenen Dokumenten beantwortet, offline.

2. **Sprachschnittstelle für Bestandsanlagen.** Alte Maschine, neue Bedienung: ein Laptop daneben statt einer neuen Steuerung.
3. **Klein anfangen mit einem lokalen Piloten.** Ein Laptop, offene Modelle, ein klar abgegrenzter Anwendungsfall.

Vier Sätze fassen den Vormittag zusammen:

- **Offene Modelle reichen für echte Gespräche.** Spracherkennung, Sprachmodell und Stimme laufen auf einem Laptop.
- **Unabhängig heißt austauschbar.** Modell, Stimme und Hardware lassen sich einzeln wechseln.
- **Die Nähte reißen, nicht die Bausteine.** Messen statt vermuten, vor allem dort, wo Teile zusammenkommen.
- **Klein anfangen, lokal bleiben.** Ein abgegrenzter Pilot zeigt schnell, was geht.

Sie möchten prüfen, ob lokale KI für Ihren Betrieb taugt? Das Mittelstand-Digital Zentrum Fokus Mensch begleitet kleine und mittlere Unternehmen dabei kostenfrei. Ansprechpartner ist Erik Dethier, e.dethier@digitalzentrum-fokus-mensch.de, mehr unter www.digitalzentrum-fokus-mensch.de.

Bilder: Standbilder aus eigenen Videoaufnahmen, Dashboard-Screenshots aus NAOtonomous.