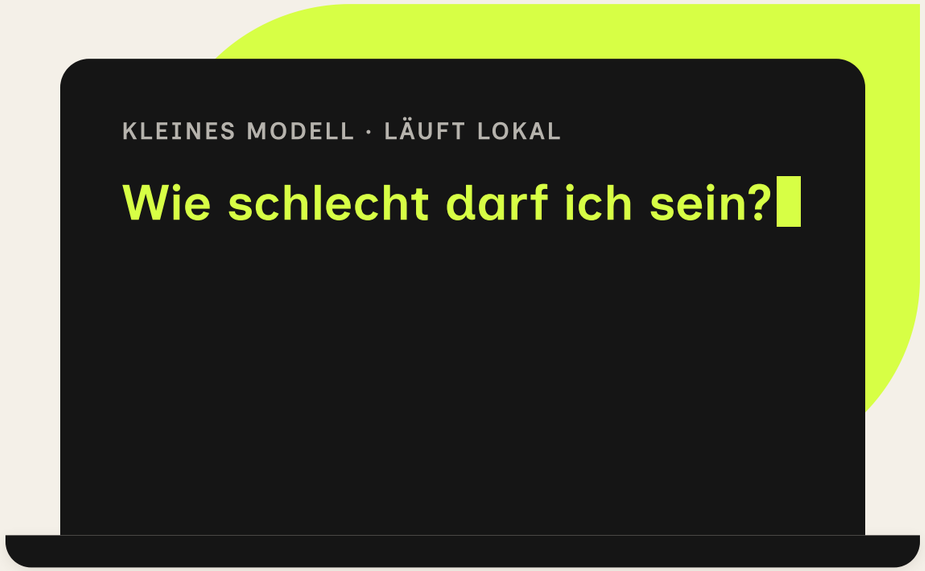
WUD 2026 · Aus unserem Forschungsprojekt

Wie schlecht darf KI sein?



Was kleine, sparsame Sprachmodelle trotzdem leisten können

Timo Korinth · untrouble

Die Ausgangslage

Große KI ist beeindruckend. Aber …

… sie ist teuer

Bezahlt wird für jede Anfrage – oder für sehr teure eigene Technik.

… sie ist hungrig

Große Modelle brauchen Rechenzentren voller Spezialchips und viel Strom.

… sie läuft woanders

Fragen und Dokumente wandern auf Server, die jemand anderem gehören.

Mit vertraulichen Daten ist das oft keine echte Option.

Die umgekehrte Frage

Wie klein und sparsam darf eine KI sein …

… damit sie trotzdem zuverlässig hilft?

Das Forschungsprojekt

Was unser Forschungsprojekt herausfinden soll

Wir entwickeln und erproben ein KI-System für die öffentliche Verwaltung, das auf eigenen Rechnern läuft. Vier Fragen leiten uns.

Sparsam

Wie klein kann ein Sprachmodell werden, ohne schlechter zu antworten?

Findig

Wie findet es in großen Textbeständen die richtige Stelle?

Verlässlich

Wie merkt das System selbst, wenn eine Antwort nicht stimmt?

Verständlich

Wie arbeiten Menschen gut damit, auch in einfacher Sprache?

Was heißt „klein“?

8 Milliarden Zahlen, grob gerundet

Ein Sprachmodell ist im Kern eine riesige Liste von Zahlen. Speichert man jede Zahl gröber, schrumpft das Modell – bis es auf einen Laptop passt.

16 Bit | 8 Bit | **etwa 4 Bit**

Der Preis ist etwas Qualität. Wie viel das auf Deutsch kostet, ist eine unserer Forschungsfragen.

MODELL **Llama 3.1 · 8 Milliarden Zahlen**

0,75	−0,25	0	−0,875
0,375	0,25	−0,625	0,875
0	0,625	−0,375	0,125

Jede Zahl kann 16 verschiedene Werte annehmen.

4,9 GB groß



0 GB 16 GB

Der Versuch

Ein Laptop. Ein kleines Modell. 121 Fragen.

1 Laptop

mit Apple-Chip M1 Max. Kein Rechenzentrum, keine Internetverbindung nötig.

4,9 GB

Llama 3.1 mit 8 Milliarden Zahlen. Frei herunterladbar.

121 Fragen

zu Gesetzen, Behördenleistungen und Entwürfen für Bescheide. Nur öffentliche Texte und erfundene Fälle.

11 Wörter pro Sekunde

Eine typische Antwort dauert 14,6 Sekunden, einschließlich Lesen der Quellen.

Wo es scheitert · 1

Es schreibt, was wahrscheinlich klingt

AUFGABE Schreib diesen Satz weiter. Ohne Quellen, nur aus dem Gedächtnis.

Der Antrag auf Kindergeld wird bei der Familienkasse des Bundesministeriums für Arbeit und Soziales …

Klingt richtig. Ist falsch.

Die Familienkasse gehört zur Bundesagentur für Arbeit, nicht zu einem Ministerium. Das Modell hat nichts nachgeschlagen. Es schreibt, was plausibel klingt.

Wo es scheitert · 2

Dieselbe Frage, sechs Antworten

Aus dem Gedächtnis gefragt: ohne Quellen und mit etwas eingebautem Zufall.

Noch einmal fragen →

4 verschiedene Gesetze. 2 von 6 richtig.

Aus dem Gedächtnis ist ein kleines Modell nicht verlässlich. Und jede Antwort klingt gleich überzeugt.

FRAGE In welchem Gesetz ist das Wohngeld geregelt?

1	In der Wohngeldgesetzgebung, die im Wohngeldgesetz geregelt ist.	✓ Wohngeldgesetz
2	Das Wohngeld ist im Zweiten Buch Sozialgesetzbuch (SGB II) geregelt.	✗ SGB II
3	Das Wohngeld ist in § 12 des Bundeswohngeldgesetzes (BWohngG) geregelt.	✗ Bundeswohngeldgesetz
4	Das Wohngeld ist in dem Sozialgesetzbuch (SGB) XII geregelt.	✗ SGB XII
5	Das Wohngeld ist in § 1 des Wohngeldgesetzes geregelt.	✓ Wohngeldgesetz
6	Das Wohngeld ist im SGB XII geregelt.	✗ SGB XII

Woher die Quellen kommen

Erst suchen, dann schreiben

Vor jeder Antwort holt das System passende Textstellen heraus. Das Prinzip heißt RAG. Wir nutzen dafür zwei Suchen statt einer.

1	**Bedeutungssuche** mit Vektordatenbank: versteht Umschreibungen
2	**Wortsuche**, ganz klassisch: stark bei exakten Wörtern, Namen und Nummern
3	Zusammenführen und nachsortieren
4	Die besten fünf Textstellen gehen an das Modell

BEISPIEL Was setzt welche Suche an die erste Stelle?

„Mein Ausweis wurde geklaut.“

Wortsuche	✗ findet nichts. Kein Wort der Frage steht im Text.
Bedeutungssuche	✓ „Personalausweis verloren: Verlust melden“

„Wo sitzt Herr Maier?“

Wortsuche	✓ „Elterngeld: zuständig ist Herr **Maier**, Zimmer 21“
Bedeutungssuche	✗ „Führerschein: zuständig ist Herr **Mayer**, Zimmer 41“

Unser Test mit 121 Fragen: richtige Textstelle unter den ersten fünf



Wo es überraschend gut klappt

Mit Quellen wird es verlässlich

Dasselbe Modell, jetzt mit den fünf Textstellen aus der Suche.

Stimmt das auch? Wir prüfen doppelt.

Von Hand. Wir lesen in der Quelle nach.

Mit einem Prüf-Modell. Ein zweites, kleines KI-Modell vergleicht jeden Satz mit der Quelle und schätzt, wie gut sie ihn deckt. So lassen sich alle 121 Antworten prüfen.

3 von 3 Aussagen stimmen.

Auf dem Laptop: 8,0 Sekunden Quellen lesen, 4,7 Sekunden schreiben.

FRAGE Wann wird ein Wohngeld-Bewilligungsbescheid nach § 28 WoGG unwirksam?

- Ein Wohngeld-Bewilligungsbescheid nach § 28 WoGG wird unwirksam, wenn der Wohnraum, für den Wohngeld bewilligt ist, von keinem zu berücksichtigenden Haushaltsmitglied mehr genutzt wird [1].
- Dies tritt zum 1. des Monats ein, in dem die Nutzungsaufgabe eintritt.
- Wenn die Nutzungsaufgabe nicht zum 1. eines Monats eintritt, wird der Bescheid vom 1. des nächsten Monats an unwirksam [1].

Von Hand nachgelesen: richtig Steht so in § 28 Absatz 1 Wohngeldgesetz.

> … und Wegfall des Wohngeldanspruchs (1) Der Bewilligungsbescheid wird vom Ersten des Monats an unwirksam, in dem der Wohnraum, für den Wohngeld bewilligt ist, von keinem zu berücksichtigenden Haushaltsmitglied mehr genutzt wird; erfolgt die Nutzungsaufgabe …

Prüf-Modell: zu 87 % gedeckt Ab 50 % gilt ein Satz als gedeckt.

Wo es trotzdem danebenliegt

Und dann dichtet es etwas dazu

Gleicher Aufbau, andere Frage. In den fünf Textstellen steht gar keine Frist.

Zwei Sätze stimmen. Der dritte ist erfunden.

Er klingt genauso überzeugend wie die anderen. Das Prüf-Modell hat ihn trotzdem erwischt: In den Quellen steht davon nichts.

FRAGE **Welche Bearbeitungs- bzw. Antragsfristen gelten bei Kfz-Abmeldung laut Verwaltungsportal?**

- ✓ Laut [2] können Sie die Außerbetriebsetzung eines Fahrzeugs online erledigen, wenn das Fahrzeug ab dem 01.01.2015 zugelassen wurde.
- ✓ Die Bearbeitungs- bzw. Antragsfristen sind nicht explizit angegeben.
- ✗ Es wird jedoch empfohlen, den Online-Service zu nutzen, da dies eine schnelle und unkomplizierte Möglichkeit ist, die Abmeldung durchzuführen.

✗ Von Hand nachgelesen: falsch — Hinzugedichtet. Eine Empfehlung oder „schnell und unkompliziert“ steht in keiner der fünf Textstellen.

✗ Prüf-Modell: zu 0 % gedeckt — Ab 50 % gilt ein Satz als gedeckt.

Grenzen im Behördenalltag

Wo Suchen und Schreiben an Grenzen stößt

Ein Schreibauftrag sagt nicht, was zu suchen ist

Eine Frage nennt ihr Thema. „Schreib einen Bescheid, der das Bürgergeld aufhebt“ nennt nur das Ergebnis. Den passenden Paragrafen findet die Suche dann oft nicht.

Was die Suche liefert, wird auch benutzt

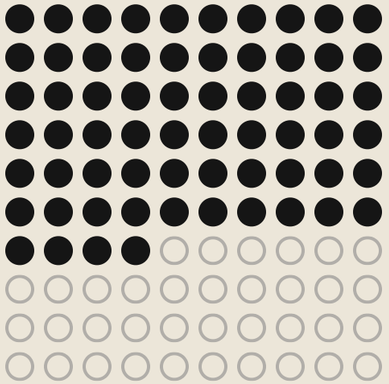
Ein Ablehnungsbescheid berief sich auf einen Paragrafen, der nur die Form des Bescheids regelt, nicht die Gründe.

Zum Nachschlagen reicht oft die Suche

Wer einen Paragrafen sucht, braucht den Wortlaut. Dann lieber die Fundstelle zeigen als sie nacherzählen lassen.

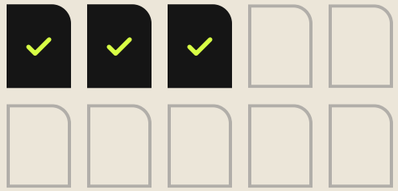
Die ehrliche Bilanz

Wie schlecht ist es denn nun?



64 von 100

Aussagen sind durch ihre Quelle gedeckt.



3 von 10

Antworten sind von vorn bis hinten gedeckt.

Für fertige Auskünfte reicht das nicht. Also ein größeres Modell kaufen?

Hebel 1 · Die Anleitung

Drei Hinweise statt mehr Rechenleistung

Anleitung 1 | **Anleitung 2**

0 %

der Aufträge für einen Entwurf abgelehnt
0 von 18

1. Für wen das Modell arbeitet: Sachbearbeitung
2. Ein Mensch prüft jeden Entwurf vor dem Versand
3. Die Namen im Auftrag sind nur Platzhalter

Nebenwirkung: Quellennummern setzt das Modell jetzt seltener (70 % → 47 %). Jede Änderung muss gemessen werden.

AUFTRAG **Verfasse einen Aufhebung-Bescheid zu Bürgergeld für Pers02 (Aktenzeichen AZ-0002, Musterstraße 1, 12345 Musterhausen).**

Hier ist ein Entwurf für einen Aufhebung-Bescheid:
"Bescheid über die Aufhebung der Bewilligung von Bürgergeld
Az.: AZ-0002
Pers02
Musterstraße 1
12345 Musterhausen
Sehr geehrte/r Frau/Herr,
mit Bescheid vom [Datum] wurde Ihnen Bürgergeld bewilligt. Nach Prüfung der Voraussetzungen wird die Bewilligung von Bürgergeld aufgehoben. …

Auszug. Der ganze Entwurf: 122 Wörter in 11 Sekunden.

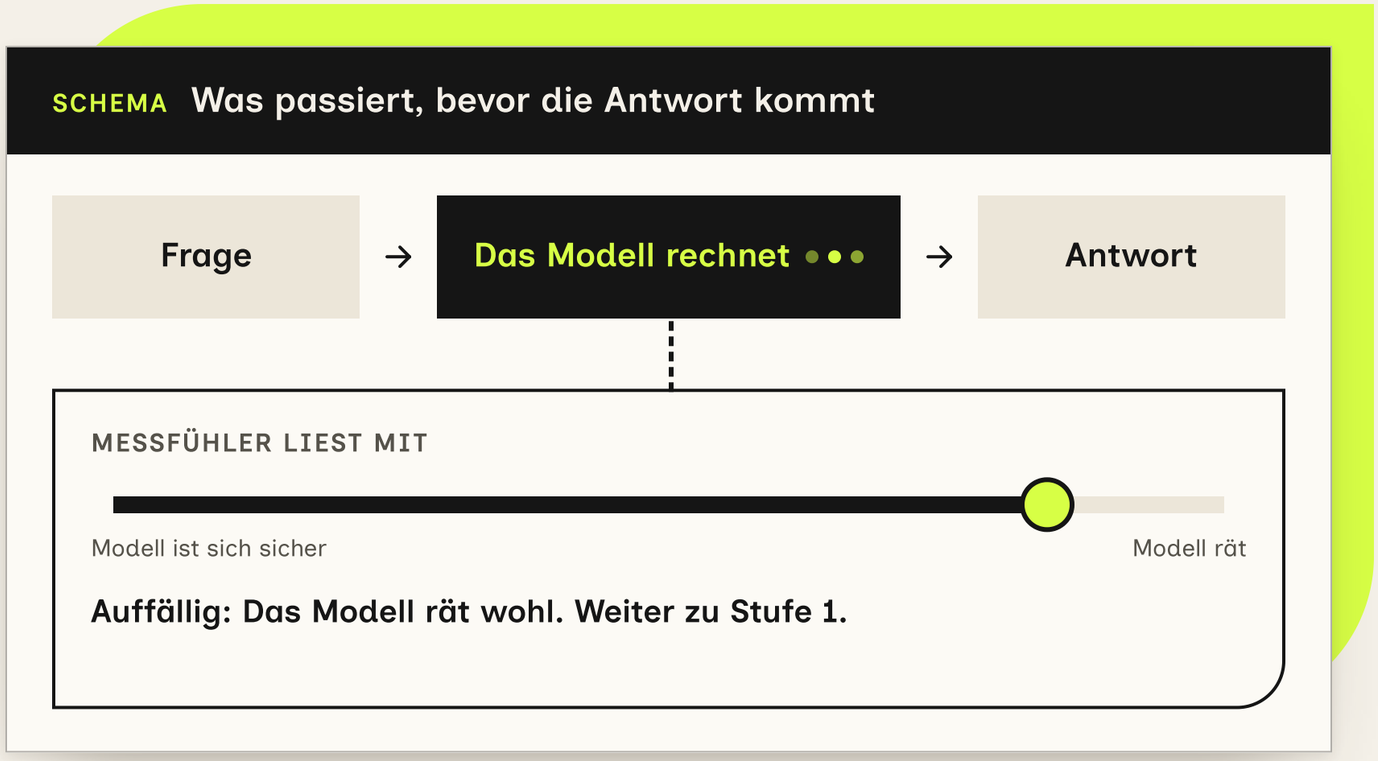
Hebel 2 · Die geplante Kontrolle · Stufe 0

Schnelltest: messen statt fragen

Bevor das Modell antwortet, rechnet es. Rät es nur, sieht diese Rechnung anders aus, als wenn es die Antwort weiß.

MESSFÜHLER	Eine kleine Formel, die bei jeder Antwort in die Rechnung des Modells hineinschaut. Kein zweites Modell.
ANGELERNT	Einmal vorab, an Beispielen, bei denen wir wissen, wie unsicher das Modell war.
AUFWAND	Etwa eine Millisekunde pro Antwort.

Unauffällig: Die Antwort geht direkt raus. **Auffällig:** weiter zu Stufe 1.



Hebel 2 · Die geplante Kontrolle · Stufe 1

Mehrmals fragen: Ist sich das Modell einig?

Bei Verdacht antwortet dasselbe Modell 4- bis 8-mal, jedes Mal etwas anders.

TEMPERATUR	So heißt der Regler für den Zufall. Bei 0 nimmt das Modell immer das wahrscheinlichste nächste Wort. Höher gedreht, auch mal das zweit- oder drittbeste.
IDEE	Weiß es die Antwort, kommt trotzdem immer dieselbe Aussage. Rät es, jedes Mal eine andere.
VERGLEICH	Das Prüf-Modell von vorhin prüft: Sagen zwei Antworten dasselbe?

Einig: Die beste Antwort geht raus. **Uneinig:** weiter zu Stufe 2.

SCHEMA **Dieselbe Frage, sechs Antworten**

TEMPERATUR **0,7**

0 · immer das wahrscheinlichste Wort — 1 · viel Zufall

Bei 0,7 wird gewürfelt: Jede Antwort fällt etwas anders aus.

Wohngeldgesetz 2×

In der Wohngeldgesetzgebung, die im Wohngeldgesetz geregelt ist.

Das Wohngeld ist in § 1 des Wohngeldgesetzes geregelt.

SGB XII 2×

Das Wohngeld ist in dem Sozialgesetzbuch (SGB) XII geregelt.

Das Wohngeld ist im SGB XII geregelt.

SGB II 1×

Das Wohngeld ist im Zweiten Buch Sozialgesetzbuch (SGB II) geregelt.

Bundeswohngeldgesetz 1×

Das Wohngeld ist in § 12 des Bundeswohngeldgesetzes (BWohngG) geregelt.

Hebel 2 · Die geplante Kontrolle · Stufe 2

Satz für Satz gegen die Quelle

Jetzt wird gründlich geprüft: Die Antwort wird in einzelne Aussagen zerlegt, und jede muss in einer Quelle stehen.

PRÜF-AUFSATZ	Eine kleine Zusatzschicht auf demselben Sprachmodell. Sie urteilt viermal mit Begründung, die Mehrheit zählt.
FAKTENPRÜFER	Ein eigenes kleines Modell (770 Millionen Zahlen), soll auf deutsche Gesetzestexte angelernt werden.
WANN	Nur bei etwa 1 bis 3 % der Antworten.

Alles gedeckt: Die Antwort geht raus. **Zweifel bleibt:** ein Mensch entscheidet.

SCHEMA **Die Antwort von vorhin, Aussage für Aussage**

Laut [2] können Sie die Außerbetriebsetzung eines Fahrzeugs online erledigen, wenn das Fahrzeug ab dem 01.01.2015 zugelassen wurde.	✓ steht in Quelle
Die Bearbeitungs- bzw. Antragsfristen sind nicht explizit angegeben.	✓ steht in Quelle
Es wird jedoch empfohlen, den Online-Service zu nutzen, da dies eine schnelle und unkomplizierte Möglichkeit ist, die Abmeldung durchzuführen.	✗ kein Beleg

SO KÖNNTE DIE BEGRÜNDUNG AUSSEHEN

Die Empfehlung für den Online-Service steht in keiner der fünf Textstellen. Antwort nicht freigeben.

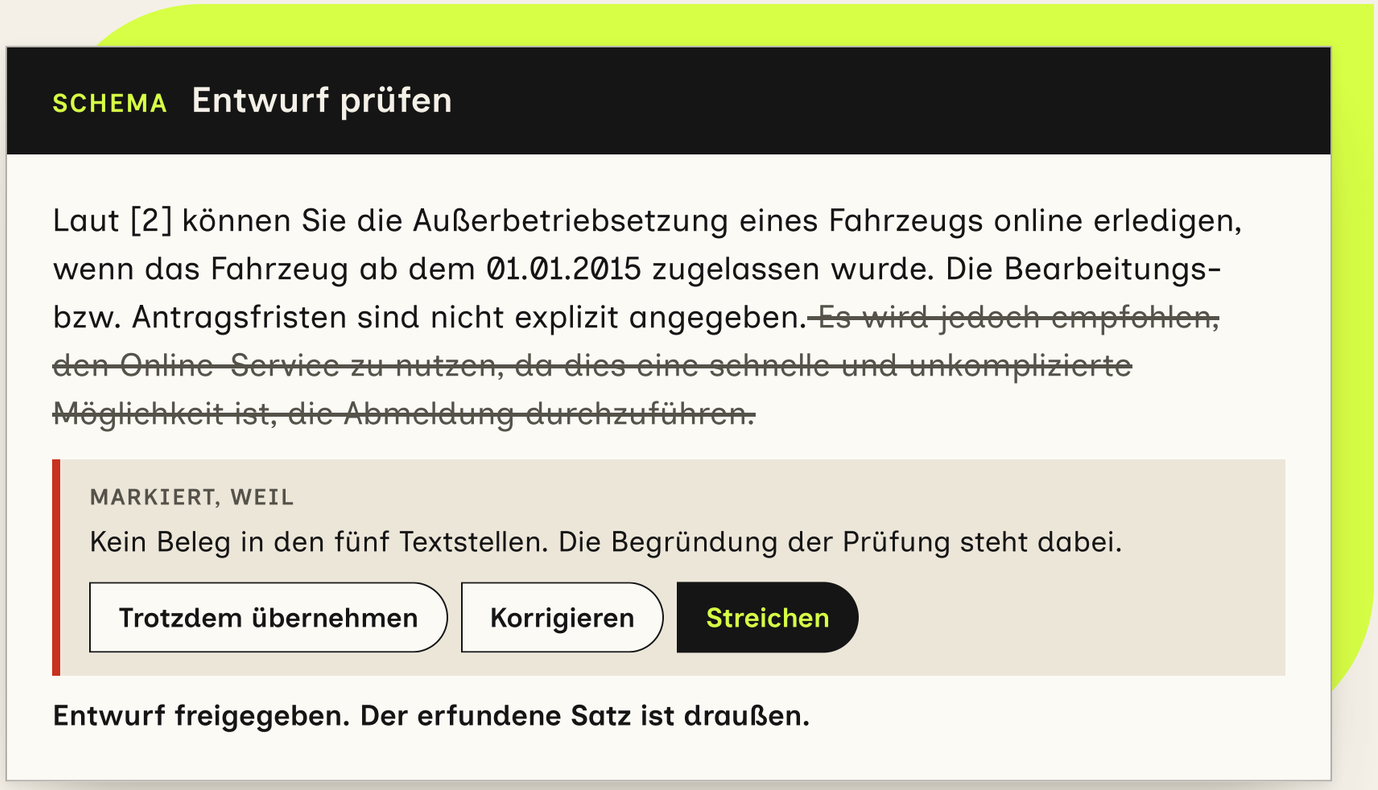
Hebel 2 · Die geplante Kontrolle · Am Ende

Ein Mensch entscheidet

Hier arbeitet keine KI mehr. Die Sachbearbeitung sieht den Entwurf mit den markierten Stellen und entscheidet selbst.

SIEHT	Nur die zweifelhaften Sätze, mit Begründung und Quelle.
ENTSCHEIDET	Übernehmen, korrigieren oder streichen.
VERANTWORTUNG	Bleibt beim Menschen. Die KI liefert Entwürfe, keine Entscheidungen.

Ergebnis: Der Mensch prüft wenige markierte Stellen statt jeden Text von vorn.



Fazit

Wie schlecht darf KI sein?

Ziemlich klein und alles andere als perfekt – wenn drei Dinge stimmen.

Quellen statt Gedächtnis

Erst suchen, dann schreiben lassen. Zum Nachschlagen reicht oft die Suche.

Eine Anleitung, die die Lage erklärt

Drei Hinweise: Ablehnungen von 33 % auf 0 %.

Kontrolle, die Zweifel zeigt

Die Maschine prüft vor, der Mensch entscheidet.

Danke. Fragen?

Timo Korinth · untrouble.de
Das Forschungsprojekt wird im Zentralen Innovationsprogramm Mittelstand (ZIM) gefördert.