HACOY

Wie ein kleines Modehaus mit lokaler KI arbeitet

Wie ein Modehaus in München seinen Alltag mit KI Modellen auf eigenen Rechnern bestreitet, was wir gemessen haben und was nicht funktioniert hat.

130	224	24.124	99,8 %
KI Rollen	geplante Aufgaben	Modellaufrufe in 30 Tagen	davon gelungen, alle im eigenen Haus

HACOY Research Note 01 · München · Oktober 2026

Jede Zahl stammt aus unseren eigenen Unterlagen, gelesen am 18. Oktober 2026, und ist in den Quellen am Ende genannt. Diese Notiz nennt keine andere Marke.

Kurz gesagt

HACOY ist ein Modehaus mit einem Gründer und einem kleinen Team. Dahinter arbeitet ein System aus 130 KI Rollen und 224 geplanten Aufgaben, die Mails lesen, Antworten entwerfen, Produktangaben prüfen, Fragen recherchieren und die Unterlagen ordnen. Alles davon läuft auf sechs Apple Rechnern, die uns gehören, in München. In den letzten 30 Tagen hat das System 24.124 Mal ein Sprachmodell aufgerufen. Jeder Aufruf in unserem Protokoll wurde von einem Modell auf unseren eigenen Rechnern beantwortet, und 99,8 Prozent davon gelangen.

Diese Notiz beschreibt, wie das System gebaut ist, was wir gemessen haben und was nicht funktioniert hat. Wir schreiben sie auf, weil wir das meiste davon auf dem langsamen Weg gelernt haben, und weil man in einem kleinen Haus gut sehen kann, wie das im Alltag aussieht.

Warum lokal

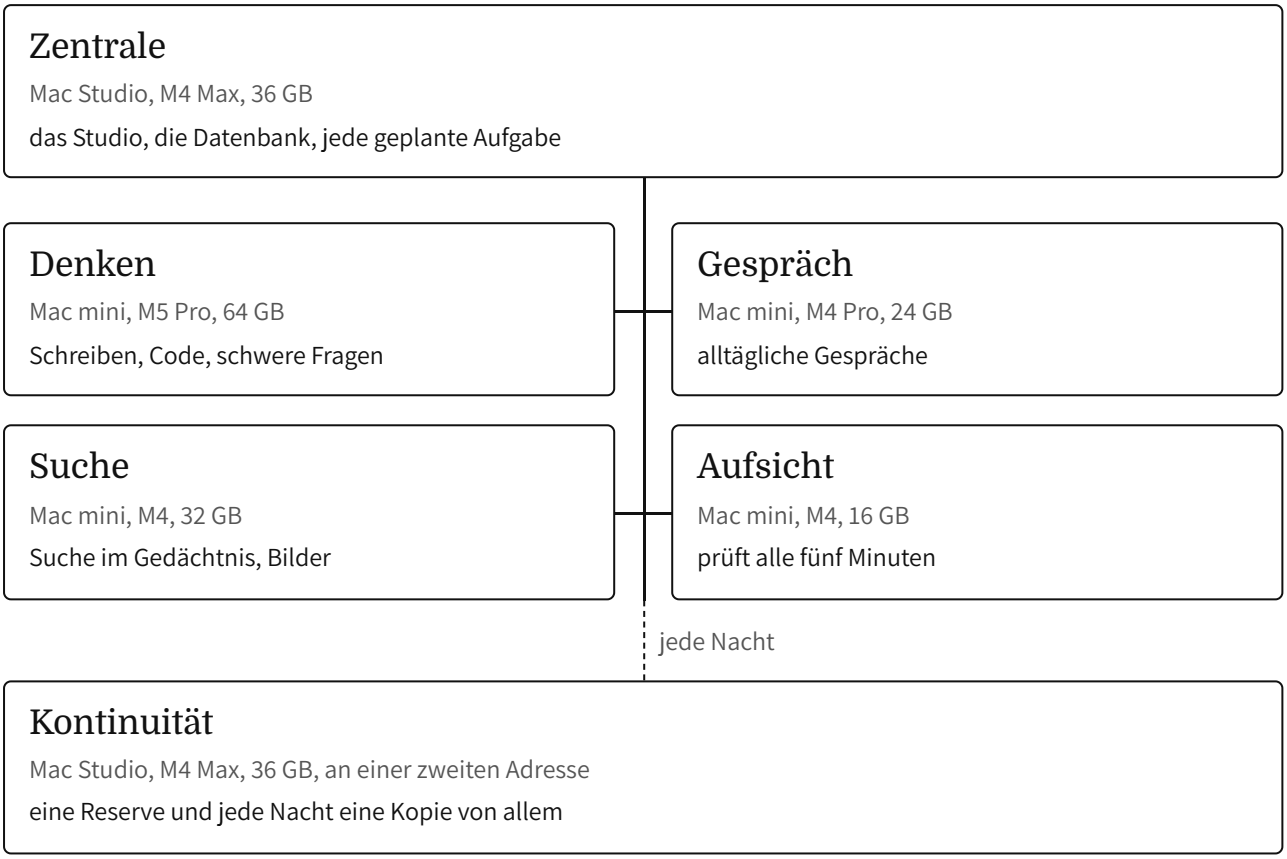
Wir haben früh entschieden, dass die Agenten, die von selbst und rund um die Uhr arbeiten, nur Modelle auf unseren eigenen Rechnern nutzen. Dafür gibt es drei Gründe.

Der erste ist Privatsphäre. Kundenmails, Bestellungen und unsere eigenen Notizen verlassen das Haus nicht, um von einem fremden Modell gelesen zu werden. Der zweite sind Kosten, die nicht mit der Nutzung wachsen: Eine Schleife, die alle fünf Minuten läuft, kostet denselben Strom, ob sie nützt oder nicht, und es gibt keine Rechnung pro Wort. Der dritte ist das Lernen. Jeder Prompt, der an einem kleinen Modell scheitert, lehrt uns etwas über die Aufgabe, und die Antwort ist meist, die Aufgabe kleiner zu machen, nicht das Modell größer.

Das System selbst bauen und reparieren wir mit einem Programmierassistenten in der Cloud, und ein Mensch liest jede Änderung. Was das Unternehmen Tag für Tag trägt, läuft lokal.

Die Rechner

Sechs Rechner, geplant wie ein Team



Quelle: unser Rechnerregister und Rollenplan, 23. und 24. September 2026. Jeder Rechner hat eine Aufgabe, eine Vertretung und ein Ziel.

Rolle	Rechner	Speicher	Aufgabe
Zentrale	Mac Studio, M4 Max	36 GB	die Studio App, die Datenbank, jede geplante Aufgabe, das eine Eingabefeld
Denken	Mac mini, M5 Pro	64 GB	das Schreibmodell, das Programmiermodell, das Modell für schwierige Fragen
Gespräch	Mac mini, M4 Pro	24 GB	alltägliche Gespräche im Hintergrund
Suche	Mac mini, M4	32 GB	die Embeddings, mit denen das System in seinem eigenen Gedächtnis sucht, und das Lesen von Bildern
Aufsicht	Mac mini, M4	16 GB	prüft alle fünf Minuten, dass nichts still stehen geblieben ist
Kontinuität	Mac Studio, M4 Max	36 GB	eine Reserve an einer zweiten Adresse, jede Nacht eine Kopie von allem

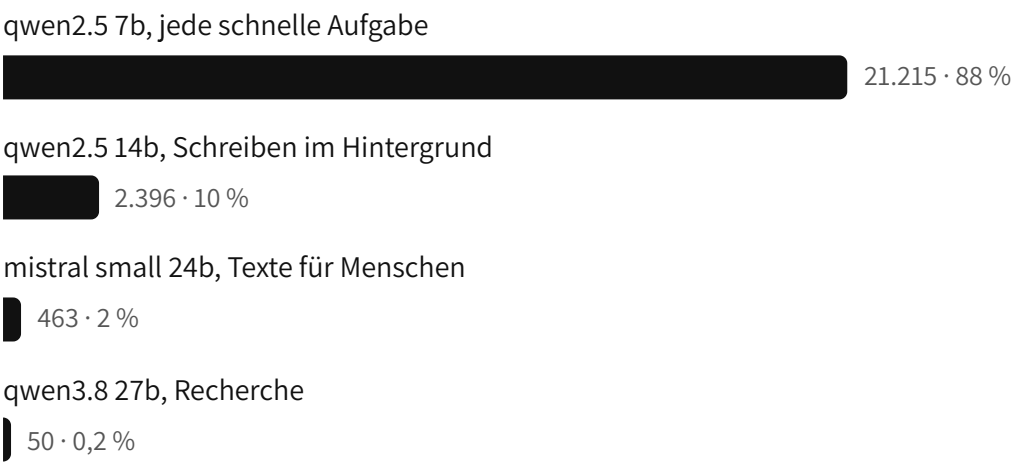
Wir haben die Rechner geplant wie ein Team. Jeder hat eine einzige Aufgabe, eine Liste dessen, was er nie tut, eine Vertretung, wenn er ausfällt, und ein Ziel, das sich prüfen lässt. Nur die Zentrale hat viele Aufgaben, und deshalb hat sie einen Zwilling.

Die Modelle

Modell	Größe	Aufgabe	Aufrufe, letzte 30 Tage
qwen2.5 7b	4,7 GB	jede schnelle Aufgabe: Sortieren, kurze Entwürfe, Gespräche der Rollen	21.215
qwen2.5 14b	9,0 GB	Schreiben im Hintergrund, das ein paar Sekunden länger dauern darf	2.396
mistral small 24b	14,3 GB	Texte, die ein Mensch lesen wird	463
qwen3.8 27b	30 GB	Recherche und Fragen, die erst Nachdenken brauchen	50
qwen3 coder 30b	18,6 GB	der Programmierassistent	
qwen2.5 vl 7b	6,0 GB	Bilder lesen	
nomic embed, bge m3	0,3 und 1,2 GB	Text in Vektoren für die Suche verwandeln	

Neun von zehn Aufrufen gehen an das kleinste Modell. Das war am Anfang nicht der Plan, es ist das, was uns die Messungen gelehrt haben.

Wohin die 24.124 Modellaufrufe der letzten 30 Tage gingen



Quelle: das Modellprotokoll des Studio, die 30 Tage bis zum 17. Oktober 2026. Jeder Aufruf auf unseren eigenen Rechnern beantwortet, 99,8 Prozent gelangen.

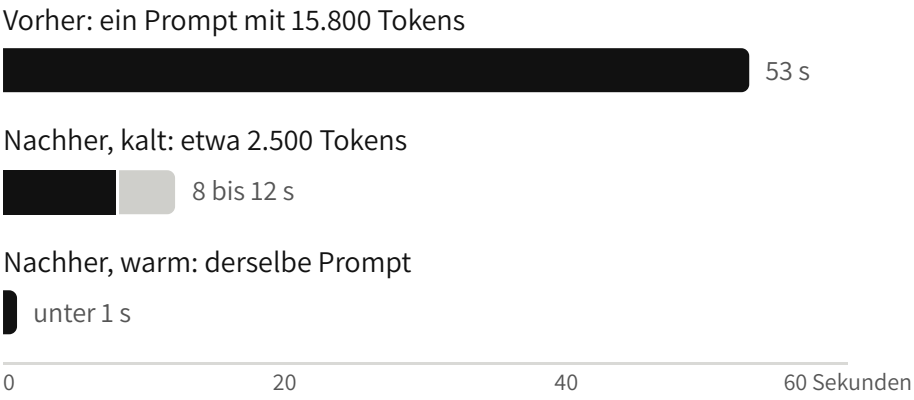
Was wir gemessen haben

Ein kleines Modell, das warm bleibt, schlägt ein größeres, das kalt startet. Auf einem Mac mini antwortet das 7b Modell kalt in 3,1 Sekunden und geladen in 0,44 Sekunden. Das 14b Modell braucht etwa 60 Sekunden, um neu zu laden, wenn es aus dem Speicher verdrängt wurde, und war damit auf einem Rechner mit 36 GB für Gespräche nicht brauchbar. Am 25. Juni 2026 haben wir jede Gesprächsrolle auf das 7b Modell gelegt und es geladen gehalten.

Das größte Modell, das hineinpasst, ist nicht das größte, das funktioniert. Ein 32b Modell auf der Zentrale mit 36 GB lief aus dem Speicher und gab leere Antworten. Am 17. Juni 2026 haben wir es aus jedem Ablauf genommen. Von den 36 GB bleiben etwa 12 GB für Modelle, wenn der Rest des Systems läuft, also hält die Zentrale ein Modell zur Zeit.

Der Prompt ist der langsame Teil. Unser hauseigener Assistent begann jede Antwort einmal mit einer Rolle und einem Handbuch von 15.800 Tokens. Auf dem Rechner mit 64 GB hieß das 53 Sekunden Lesen vor dem ersten Wort. Gekürzt auf etwa 2.500 Tokens beginnt eine warme Antwort in unter einer Sekunde und eine kalte in 8 bis 12 Sekunden.

Sekunden bis zum ersten Wort einer Antwort



Quelle: unser Assistent auf dem Mac mini mit 64 GB, gemessen am 4. Oktober 2026. Der graue Teil eines Balkens ist die Spanne.

Tempo auf dem Rechner mit 64 GB, gemessen am 4. Oktober 2026: Das Schreibmodell beginnt nach etwa 2 Sekunden und schreibt etwa 20 Tokens pro Sekunde, das Programmiermodell etwa 90, das Recherchemodell etwa 10, nachdem es nachgedacht hat.

Mehr Speicher kauft Raum, nicht Schreibtempo. Apples eigene Zahlen für die M5 Generation zeigen, dass ein Prompt 3,3 bis 4 Mal schneller gelesen wird als auf dem M4, geschrieben wird aber nur etwa 1,2 Mal schneller, weil das Schreiben an der Speicherbandbreite hängt. Ein größerer Rechner erlaubt größere Modelle und längere Prompts. Dasselbe Modell schreibt darauf kaum schneller.

Wie viel schneller die M5 Generation als der M4 ist



Die senkrechte schwarze Linie ist das Tempo des M4. Quelle: Apple Newsroom, 25. August 2026. Das Schreiben hängt an der Speicherbandbreite, ein größerer Rechner kauft also eher Raum als Schreibtempo.

Wie hier ein Agent entsteht

Jeder Agent beginnt als ein Satz des Gründers, zum Beispiel „eine Research Abteilung, für HACOY und für mich selbst". Daraus wird eine Rolle mit einem Namen und einem kurzen Prompt, ein Kontext, der nur die Daten holt, die diese Rolle braucht, ein kurzes Kit mit den Regeln des Hauses ganz oben, eine Schleife, die nach Plan läuft, und eine Warteschlange, in der die Arbeit auf einen Menschen wartet.

Die Kits sind wichtiger als die Modelle. Jedes Kit ist eine kurze Datei pro Thema, mit einer datierten Zeile pro Regel. Das Kit für die Sprache sagt zum Beispiel, dass HACOY warm und genau schreibt, nie laut, und prüft jeden Text mit einem kleinen Validator, bevor er hinausgeht. Ein kleines Modell folgt einer kurzen, genauen Datei weit besser als einem langen Handbuch.

Der Gründer hat ein einziges Eingabefeld. Was er dort schreibt, legt das System ab, entwirft oder recherchiert es, und es antwortet mit einer Zeile. Nichts erreicht einen Kunden ohne seine Freigabe, außer ein benannter Schalter für genau diese eine Art von Nachricht wurde nach einer Prüfung eingeschaltet.

Was nicht funktioniert hat

Regeln, die nur gelesen werden, ändern nichts. Eine Prüfung am 30. August 2026 ergab, dass unsere Hausregeln 425 Mal in Sitzungen geladen worden waren und kein einziges Ergebnis verändert hatten. Regeln, die eine Maschine prüfen kann, wurden deshalb zu einem Validator, der blockiert. Ein Text, der eine davon bricht, geht nicht hinaus.

Eine Warteschlange, die niemand lesen kann, ist keine Aufsicht. Am 13. Oktober 2026 warteten etwa 19.700 Einträge auf eine Freigabe. Eine Schreibschleife hatte 14.364 Entwürfe erzeugt, und keiner wurde behalten. Einer von 102 entworfenen LinkedIn Beiträgen wurde freigegeben. Seit dem 15. Oktober misst eine nächtliche Prüfung, wie oft die Arbeit jedes Erzeugers behalten wird, und meldet jeden, der unter einmal in zehn fällt, damit er pausiert werden kann. Was alt ist, wird archiviert, nie gelöscht.

Agenten brauchen ein Register. Von 217 geplanten Aufgaben am 13. Oktober ließen sich nur 36 einer benannten Rolle zuordnen. Manche Schleifen liefen, ohne etwas hervorzubringen. Eine Zahl von Agenten misst Aufwand. Was wir heute zählen, ist, was jeder geliefert hat und ob ein Mensch es behalten hat.

Was wir daraus mitnehmen

Ein kleines Haus braucht kein großes Modell, um mit KI zu arbeiten. Es braucht kleine Aufgaben, kurze Regeln, einen Menschen bei jeder Nachricht, die hinausgeht, und Messungen, die es auch liest. Die Rechner kosten einmal, das Lernen wächst mit.

Weitere Notizen folgen, sobald wir mehr gemessen haben. Studierende und Forschende, die mit uns an solchen Fragen arbeiten möchten, sind auf unserer Research Seite willkommen.

Quellen

Die Zahlen stammen aus unseren eigenen Unterlagen: dem Modellprotokoll des Studio für die 30 Tage bis zum 17. Oktober 2026, dem Rechnerregister und dem Rollenplan vom 23. und 24. September 2026, der Übersicht der Agenten vom 22. September 2026, den Messungen des Assistenten vom 4. Oktober 2026, der Prüfung des Studio vom 13. Oktober 2026 und der Prüfung der Erzeuger vom 15. Oktober 2026. Apples Zahlen zur M5 Generation stammen aus dem Apple Newsroom vom 25. August 2026.