f2b2

BAND 1

DIE EIGENE LEISTUNGSFÄHIGE KI

Ein Sprachmodell auf deinem Mac.
Ohne Cloud. Ohne Abo. Ohne Gnade.

Frank Tentler.ai · Stand 17.08.2026

f2b2

Eine Reihe für Menschen,
die keine Zeit für Vorworte haben.

Jeder Band bringt dich in einem Abend von null auf funktionierend. Kein Grundlagenkapitel über die Geschichte der Informatik, keine Motivationsprosa, kein Glossar am Ende, das niemand liest. Was du wissen musst, steht dort, wo du es brauchst. Was schiefgehen wird, steht direkt daneben.

Alle Angaben in diesem Band wurden am Erscheinungstag recherchiert und geprüft. Modellnamen und Programmversionen altern in diesem Feld innerhalb von Monaten. Die Denkwerkzeuge dahinter halten Jahre. Der Band ist so gebaut, dass du auch mit dem Nachfolgemodell noch damit arbeiten kannst.

f2b2 · Band 1
Die eigene leistungsfähige KI
Stand: 17. August 2026
Frank Tentler.ai
[Created with human/ai co-intelligence]

INHALT

0	Lies das zuerst, es dauert eine Minute	4
1	Die Maschine, auf der das laufen soll	6
2	Fünf Vokabeln, dann ist Schluss mit Theorie	8
3	Der erste Abend	10
4	Drei Regler entscheiden über alles	12
5	Vier Momente, in denen dein Mac kaputt wirkt	14
6	Der zweite Weg: Ollama	16
7	Was du realistisch bekommst	18
8	Datenschutz ohne Märchenstunde	20
9	Was zuerst verfällt	22

0

LIES DAS ZUERST, ES DAUERT EINE MINUTE

Erwartungsmanagement, bevor du 17 Gigabyte lädst

Du wirst heute Abend ein Sprachmodell auf deinem Mac installieren, das vollständig dir gehört. Es heißt Qwen3.8-27B, kommt aus Alibabas Forschungslabor, wurde am 14. August 2026 veröffentlicht und steht unter Apache 2.0, einer Lizenz, die dir auch die kommerzielle Nutzung ausdrücklich erlaubt. Du darfst es benutzen, verändern, in Kundenprojekte einbauen und musst niemanden fragen.

Bevor du loslegst, eine Wahrheit, die dir kein Werbetext sagt: Dieses Modell verliert jeden direkten Qualitätsvergleich mit ChatGPT oder Claude. Es hat 27,78 Milliarden Parameter, die Cloud-Systeme haben ein Vielfaches und ganze Rechenzentren dahinter. Diese Lücke schließt kein Arbeitsspeicher-Upgrade der Welt. Wer nach fünf Minuten enttäuscht ist, hat die falsche Frage gestellt.

Die richtige Frage lautet, was dieses Ding kann, was kein Cloud-Dienst je können wird. Vier Antworten. Kein einziges Zeichen verlässt deinen Schreibtisch, weshalb Mandantendaten, Personalakten und Kundenunterlagen ohne Auftragsverarbeitungsvertrag hineindürfen. Es läuft ohne Internet, im Zug, im Funkloch, im Blackout. Es kostet nach dem Kauf der Maschine null Euro pro Anfrage, du kannst es zwanzigtausend Dokumente durchkauen lassen, ohne auf eine Rechnung zu schauen. Und niemand in Kalifornien kann es abschalten, verteuern oder in seinen Antworten verändern.

Wer eines dieser vier Dinge braucht, braucht es dringend. Wer keines braucht, behält sein Abo und liest hier trotzdem weiter, weil man erst versteht, was Cloud-KI ist, wenn man einmal eine ohne Cloud betrieben hat.

MACH ES GENAU SO

Plane für den ersten Durchlauf einen Abend mit anderthalb Stunden, davon rund vierzig Minuten reines Warten auf den Download. Leg dir vorher eine echte Aufgabe aus deiner Arbeit zurecht, ein Dokument zum Zusammenfassen, eine Mail zum Umformulieren. Du willst am Ende wissen, ob dir das Ding nützt, und das erfährst du nicht mit einem Gedicht über Katzen.

HIER ENDET DIE FREUDE

Der häufigste Abbruchgrund ist keine Technikpanne, sondern eine Erwartung. Wer die lokale Installation gegen sein Cloud-Abo antreten lässt, verliert die Lust nach einer Viertelstunde und erzählt danach allen, lokale KI tauge nichts. Die Installation läuft dann einwandfrei und verstaubt trotzdem. Gegenmittel: Miss das Modell an den vier Punkten oben, nicht an der Prosaqualität von Claude.

1

DIE MASCHINE, AUF DER DAS LAUFEN SOLL

Eine einzige Zahl entscheidet, und es ist nicht die, die du denkst

Auf Windows-Rechnern gibt es Arbeitsspeicher und getrennt davon Grafikspeicher, und Sprachmodelle wollen in den Grafikspeicher. Auf deinem Mac gibt es diese Trennung nicht. Prozessor und Grafikeinheit teilen sich denselben Speicherpool, Apple nennt das Unified Memory. Die einzige Zahl, die für dich zählt, ist deshalb der Gesamtarbeitsspeicher deines Mac. Minus dem, was schon belegt ist.

Und belegt ist immer etwas. macOS nimmt sich ungefragt rund 3,5 Gigabyte, das war schon immer so, gewöhn dich dran. Ein offener Browser mit zwanzig Tabs nimmt sich gern weitere vier bis acht. Das Modell selbst braucht in seiner sparsamen Fassung rund 17 Gigabyte am Stück. Jetzt kannst du selbst rechnen, und genau diese Rechnung ist das ganze Geheimnis der Hardware-Frage.

DIE STAFFELUNG

Bei 24 Gigabyte Gesamtspeicher läuft das Modell, aber es wird eng wie in einer Studentenküche. Vor jedem Start räumst du auf, schließt Chrome, Docker und alles andere, was Speicher hortet. Bei 32 Gigabyte wird es entspannt. Bei 48 oder 64 Gigabyte musst du nie wieder über Speicher nachdenken und kannst zusätzlich das Kontextfenster aufdrehen, dazu kommen wir in Kapitel 2.

Die zweite Zahl, die zählt und die dir kein Verkäufer nennt, ist die Speicherbandbreite, also wie schnell Daten zwischen Speicher und Rechenwerk fließen. Sie bestimmt, wie flott das Modell Wörter produziert. Der M4 Pro schafft 273 Gigabyte pro Sekunde, der einfache M4 nur 120. Das ist der Unterschied zwischen flüssigem Mitlesen und Warten aufs Wort. Deshalb gilt: Ein M4 Pro mit 48 Gigabyte schlägt einen einfachen M4 mit 64. Mehr Speicher auf langsamer Anbindung ist ein größerer Stau.

MACH ES GENAU SO

Finde zuerst heraus, was du hast: Apfel-Menü, "Über diesen Mac". Dort stehen Chip und Speicher. Notiere beide Werte, du brauchst sie in Kapitel 4 für die Kontexteinstellung. Wenn du neu kaufst: Mac mini M4 Pro mit 48 Gigabyte ist Stand August 2026 der Preis-Leistungs-Punkt für genau diesen Zweck. Die M5-Generation steckt zwar schon in anderen Macs, der mini wird offiziell weiter mit M4 und M4 Pro verkauft.

HIER ENDET DIE FREUDE

Apple-Speicher ist festgelötet. Was du beim Kauf konfigurierst, ist die Ausstattung, mit der du in fünf Jahren noch arbeitest, ein Nachrüsten gibt es nicht. Der zweite Klassiker: 16-Gigabyte-Macs. Dort passt dieses Modell schlicht nicht hinein, egal was du schließt. Für 16 Gigabyte gibt es kleinere Modelle mit 4 bis 9 Milliarden Parametern, die sind erstaunlich brauchbar, aber sie sind nicht Thema dieses Bandes.

2

FÜNF VOKABELN, DANN IST SCHLUSS MIT THEORIE

Alles, was du verstehen musst, um nie wieder ein Forum lesen zu müssen

01 QUANTISIERUNG

Ein Modell besteht aus Milliarden Zahlen, und jede wird beim Training mit hoher Genauigkeit gespeichert. In voller Auflösung belegt unser Modell rund 56 Gigabyte. Quantisierung rundet diese Zahlen gröber, so wie du 19,99 im Kopf als 20 abspeicherst. Bei einer Rundung auf vier Bit schrumpft das Ganze auf etwa 17 Gigabyte und verliert dabei erstaunlich wenig Können. Im Dateinamen steht dann `4bit` oder `Q4`. Das ist der Normalfall, nimm für den Anfang nichts anderes.

02 UNIFIED MEMORY

Kapitel 1, du erinnerst dich. Der Punkt, der jeden Anfänger einmal erwischt: Ein Modell, das "in 17 Gigabyte passt", braucht eine Maschine mit deutlich mehr als 17 Gigabyte, weil Betriebssystem, Programme und der Gesprächsspeicher denselben Topf plündern.

03 KONTEXTFENSTER

Wie viel Text das Modell gleichzeitig im Blick behalten kann, gemessen in Token. Ein Token entspricht grob einem halben bis ganzen deutschen Wort. Unser Modell beherrscht von Haus aus 262.144 Token, das ist ungefähr ein dickes Buch. Merke dir den entscheidenden Satz dazu: Diese Zahl ist eine Fähigkeit des Modells und keine Eigenschaft deines Rechners. Das Modell kann so viel, dein Speicher muss es erst bezahlen.

04 KV-CACHE

Der Zwischenspeicher fürs laufende Gespräch, der mit jedem Wort wächst. Hier hat unser Modell einen echten Bautrick: Von seinen 64 Verarbeitungsschichten

führen nur 16 überhaupt so einen wachsenden Speicher, die anderen 48 arbeiten mit einem Zustand fester Größe, der bei zweitausend Wörtern genauso viel kostet wie bei zweihunderttausend. Als Faustwert fallen etwa 64 Kilobyte pro Token an, rund ein Viertel dessen, was ein konventionell gebautes Modell gleicher Größe verlangen würde. Genau deshalb sind lange Dokumente auf einem Schreibtischrechner überhaupt machbar.

05 THINKING

Das Modell kann vor der Antwort intern nachdenken und produziert dieses Nachdenken als Text, den du meistens gar nicht sehen willst. Ab Werk steht diese Funktion auf der höchsten Stufe. Das erklärt neunzig Prozent aller Berichte, das Modell sei quälend langsam. Es ist nicht langsam, es monologisiert.

MACH ES GENAU SO

Rechne deine Maschine einmal durch, mit Stift, es dauert dreißig Sekunden: Gesamtspeicher minus 3,5 für macOS, minus 17 für das Modell, minus das, was deine offenen Programme nehmen. Was übrig bleibt, geteilt durch 0,064 Megabyte, ist grob dein Budget an Kontexttoken. Bei 48 Gigabyte und mäßig vollem Schreibtisch landet das Budget weit jenseits von allem, was du je brauchst. Bei 24 Gigabyte landest du bei einer Zahl, die dir Kapitel 4 als Reglerempfehlung wiederbegegnet.

HIER ENDET DIE FREUDE

Foren und YouTube sind voll von Zahlen zu diesem Modell, die für andere Modelle oder ältere Versionen gelten. Zwischen Qwen 3.5, 3.6 und 3.8 lagen keine sechs Monate, und die Anleitungen dazu sehen zum Verwechseln ähnlich aus. Bevor du irgendeinem Tuning-Tipp folgst, prüfe das Datum des Beitrags. Alles vor dem 14. August 2026 beschreibt ein anderes Modell.

3

DER ERSTE ABEND

Anderthalb Stunden, davon vierzig Minuten Warten. Der Rest ist Klicken.

Du installierst jetzt LM Studio. Das ist ein kostenloses Programm mit grafischer Oberfläche, das Modelle findet, herunterlädt, korrekt konfiguriert und dir ein Chatfenster hinstellt. Seit Juli 2025 ist es auch für die berufliche Nutzung kostenlos, eine Registrierung verlangt niemand. Es gibt puristischere Wege über die Kommandozeile, und einer davon kommt in Kapitel 6. Heute Abend zählt, dass es funktioniert.

SCHRITT 1: PROGRAMM HOLEN

Lade LM Studio von `lmstudio.ai` und zieh es in den Programme-Ordner. Beim ersten Start fragt macOS, ob du dem Programm aus dem Internet vertraust. Ja, tust du.

SCHRITT 2: MODELL FINDEN

Klick auf das Lupensymbol und such nach `Qwen3.8-27B`. Du bekommst eine Liste von Varianten, und jetzt kommt die einzige Entscheidung des Abends: Nimm eine Fassung, in deren Namen `MLX` und `4bit` vorkommen. MLX ist Apples hauseigenes Rechenverfahren für die M-Chips. Es belegt rund zehn Prozent weniger Speicher und läuft fünfzehn bis dreißig Prozent schneller als das plattformübergreifende Format GGUF, das dir in derselben Liste begegnet. Auf einem Mac gibt es keinen Grund, das langsamere Format zu wählen.

SCHRITT 3: WARTEN

Der Download umfasst rund 17 Gigabyte. Bei 100 Megabit Leitung sind das gut zwanzig Minuten, bei DSL im Bestandsbau eher eine Stunde. Koch was. Wichtig: Lass den Mac nicht in den Ruhezustand fallen, sonst pausiert der Download kommentarlos.

SCHRITT 4: NICHT SOFORT LOSTIPPEN

Wenn der Download fertig ist, willst du sofort etwas eingeben. Tu es nicht. Erst kommen die drei Regler aus Kapitel 4, und die dauern zwei Minuten. Diese Reihenfolge ist der Unterschied zwischen einem guten und einem frustrierenden ersten Eindruck.

MACH ES GENAU SO

Such exakt so: `Qwen3.8-27B MLX 4bit`. In den Ergebnissen tauchen mehrere Anbieter auf, etwa `lmstudio-community` und `mlx-community`. Beide sind seriöse Standardquellen, nimm die, die LM Studio dir zuoberst anbietet. Prüfe vor dem Klick die angezeigte Dateigröße: Sie muss bei ungefähr 16 bis 18 Gigabyte liegen. Steht dort 55 oder mehr, hast du die unquantisierte Vollfassung erwischt, und die passt nicht auf deine Maschine.

HIER ENDET DIE FREUDE

Drei Abbruchstellen an diesem Abend. Erstens der Ruhezustand, siehe oben. Zweitens die Festplatte: 17 Gigabyte Download brauchen während des Entpackens zeitweise das Doppelte an freiem Platz, prüfe vorher, ob mindestens 40 Gigabyte frei sind. Drittens findest du im Netz Anleitungen mit dem Hinweis, für dieses Modell gebe es noch keine MLX-Fassung, nur GGUF. Diese Anleitungen sind älter als der 15. August 2026 und damit Geschichte. Die MLX-Fassungen erschienen innerhalb von Stunden nach dem Release.

4

DREI REGLER ENTSCHEIDEN ÜBER ALLES

Zwei Minuten Konfiguration, die dir Wochen Frust ersparen

REGLER 1: KONTEXTFENSTER RUNTER

Öffne die Modelleinstellungen und such den Regler für die Kontextlänge, englisch Context Length. Der Vorgabewert steht oft absurd hoch, weil das Modell nun mal 262.144 Token kann. Stell ihn auf 8.192. Wenn deine Rechnung aus Kapitel 2 üppig ausfiel, auf 16.384. Erhöhen kannst du später jederzeit, und zwar dann, wenn du wirklich ein langes Dokument verarbeiten willst und die Maschine stabil läuft. Der umgekehrte Weg, erst abstürzen und dann suchen warum, ist der übliche und der dümmere.

REGLER 2: THINKING RUNTER

Such in den Einstellungen nach Thinking, Reasoning oder Reasoning Effort. Ab Werk steht der Wert auf der höchsten Stufe, das Modell denkt dann vor jeder Antwort ausgiebig nach. Stell ihn auf mittel oder niedrig, für den allerersten Test gern ganz aus. Du erlebst damit den Unterschied zwischen einem Modell, das nach vierzig Sekunden Selbstgespräch zu antworten beginnt, und einem, das sofort schreibt. Für harte Nüsse, komplizierte Analysen, kniffligen Code, drehst du ihn gezielt wieder hoch. Das ist kein Entweder-oder, das ist ein Gangschalthebel.

REGLER 3: ALLES ANDERE IN RUHE LASSEN

Es gibt weitere Regler mit Namen wie Temperatur, Top-p und Top-k. Sie steuern, wie kreativ oder wie strikt das Modell formuliert, und LM Studio setzt sie passend zur gewählten Betriebsart automatisch. Fass sie nicht an. Irgendein Forenbeitrag wird dir empfehlen, an ihnen zu drehen. Der Beitrag wurde für ein anderes Modell geschrieben.

MACH ES GENAU SO

Reihenfolge für den ersten Prompt danach: Lade das Modell, öffne die Aktivitätsanzeige (Programme, Dienstprogramme) und wirf einen Blick auf den Tab Speicher. Der Speicherdruck-Graph unten sollte grün sein. Gelb heißt eng, rot heißt, dein Kontextfenster ist zu groß oder zu viele Programme laufen. Dann tippe deine vorbereitete echte Aufgabe aus Kapitel 0. Nicht "Hallo, wer bist du". Das kann jedes Modell, und es sagt dir nichts.

HIER ENDET DIE FREUDE

Der klassische Doppelfehler: Kontext auf Maximum und Thinking auf Werkseinstellung, gleichzeitig. Das Modell lädt minutenlang, der Lüfter dreht, die erste Antwort kommt nach einer gefühlten Ewigkeit, und du bist sicher, dass die Maschine zu schwach ist. Ist sie nicht. Wenn nach dem Zurückdrehen beider Regler immer noch alles zäh ist, dann erst lohnt der Blick auf die Hardware, und zwar in dieser Reihenfolge: Speicherdruck prüfen, andere Programme schließen, Mac neu starten, Modell neu laden. In neun von zehn Fällen ist das Thema damit durch.

5

VIER MOMENTE, IN DENEN DEIN MAC KAPUTT WIRKT

Ist er nicht. Es sieht nur exakt so aus.

Diese vier Situationen erzeugen zusammen den größten Teil des Anfängerfrusts, und alle vier haben dieselbe fiese Eigenschaft: Sie fühlen sich an wie ein zu schwacher Rechner und sind in Wahrheit eine Einstellung oder eine fehlende Datei. Wer sie kennt, bleibt gelassen. Wer nicht, gibt nach zwei Abenden auf.

MOMENT 1: SEITENWEISE SELBSTGESPRÄCH

Das Modell schreibt vor der eigentlichen Antwort lange innere Monologe und braucht ewig. Das ist der Thinking-Regler auf Werkseinstellung, Kapitel 4, Regler 2. Runterdrehen, Ruhe ist.

MOMENT 2: LADEABBRUCH MIT SPEICHERFEHLER

Beim Laden des Modells erscheint eine Fehlermeldung, irgendwas mit Memory. Entweder ist das Kontextfenster zu groß eingestellt oder die Maschine zu voll. Kontext runter, Chrome und Docker zu, neu laden. Auf einem 24-Gigabyte-Mac gehört dieses Ritual zum Betrieb dazu wie das Vorheizen zum Backen.

MOMENT 3: DAS IGNORIERTE BILD

Du ziehst ein Foto oder einen Screenshot in den Chat, und das Modell tut, als wäre nichts. Keine Fehlermeldung, einfach nichts. Das Modell versteht Bilder und sogar Video, aber diese Fähigkeit steckt in einer zweiten Datei von rund 0,9 Gigabyte, dem sogenannten Projektor. Bei den MLX-Fassungen für LM Studio ist er normalerweise dabei. Bei GGUF-Downloads von Hand muss er separat geladen werden, und wenn er fehlt, schweigt das System, statt sich zu beschweren.

MOMENT 4: PLÖTZLICHE AMNESIE

Nach etlichen Gesprächsrunden vergisst das Modell, was ihr vorher besprochen habt, obwohl dein Kontextfenster locker reichen müsste. Das ist ein bekannter Fehler der mitgelieferten Gesprächsvorlage aus den ersten Tagen nach dem Release: Sie wickelt jede vergangene Runde in leere Denkblöcke ein und schneidet dabei den Verlauf ab. Es liegt weder an dir noch an der Quantisierung. Abhilfe schafft eine korrigierte Vorlage aus der Community oder schlicht eine aktualisierte Programmversion. Bis dahin: lange Vorhaben in frischen Chats beginnen.

MACH ES GENAU SO

Häng dir diese Diagnose-Reihenfolge an den Monitor, sie löst fast alles: Erstens Thinking prüfen. Zweitens Kontextgröße prüfen. Drittens Speicherdruck in der Aktivitätsanzeige prüfen. Viertens bei Bildproblemen die Frage, ob der Projektor mitgeladen ist. Fünftens LM Studio auf Updates prüfen. Erst wenn alle fünf sauber sind, darfst du anfangen, an deinem Mac zu zweifeln.

HIER ENDET DIE FREUDE

Die gefährlichste Falle ist Moment 4, weil er wie ein Qualitätsproblem aussieht. Wer nicht weiß, dass es die Gesprächsvorlage ist, zieht den einzig logischen falschen Schluss: Das Modell sei eben doch zu klein und zu dumm. Danach wird deinstalliert und in der Community erzählt, lokale Modelle könnten sich nichts merken. Der Fehler steckt in einer Textschablone, nicht im Modell, und er ist mit einem Update erledigt.

6

DER ZWEITE WEG: OLLAMA

Für alle, die mehr wollen als ein Chatfenster. Aber erst, wenn Weg eins läuft.

LM Studio ist ein Fenster zum Reinschreiben, und für viele bleibt es das völlig zurecht. Wer aber das Modell in andere Programme einbinden will, es im Heimnetz für mehrere Leute bereitstellen oder eigene Dokumentensammlungen durchsuchbar machen möchte, braucht etwas, das im Hintergrund läuft und eine Schnittstelle anbietet. Das ist Ollama. Andere Software dockt daran an, zum Beispiel die Chat-Oberfläche Open WebUI, die aus deinem Mac einen kleinen KI-Server für den Hausgebrauch macht.

Ein Wort zur Geschichte, weil sie dir in Foren begegnen wird: Ollama war für Modelle mit Bildverständnis lange eine schlechte Wahl, die getrennte Projektordatei wurde nicht mitgeführt, Bilder blieben stumm. Das ist seit Version 0.32.12 vorbei. Es gibt inzwischen eine eigens für Apple-Chips optimierte MLX-Fassung des Modells direkt in Ollamas Bibliothek, Download rund 18 Gigabyte.

DIE ZWEI BEFEHLE

Ollama bedienst du über das Terminal. Falls du es noch nie geöffnet hast: Programme, Ordner Dienstprogramme, Programm Terminal. Es ist ein Textfenster, es beißt nicht. Nach der Installation von Ollama (Download von `ollama.com`, Doppelklick, fertig) tippst du:

```
ollama pull qwen3.8:27b-mlx
ollama run qwen3.8:27b-mlx
```

Der erste Befehl lädt, der zweite startet einen Chat direkt im Terminal. Mehr Kommandozeile brauchst du nicht. Alles Weitere, etwa Open WebUI als hübsche Oberfläche davor, ist Stoff für Band 2.

MACH ES GENAU SO

Geh diesen Weg erst, wenn Kapitel 3 bis 5 hinter dir liegen und LM Studio zuverlässig läuft. Der Grund ist schlichte Fehlersuche: Wer zwei Systeme parallel einrichtet, weiß bei einem Problem nie, wo es wohnt. Und bedenke die Platte: Ollama und LM Studio verwalten ihre Modelle getrennt, zwei Wege bedeuten zwei Kopien von je 17 bis 18 Gigabyte.

HIER ENDET DIE FREUDE

Der Klassiker hier heißt Versionsstand. Wenn `ollama pull` das Modell nicht findet oder Bilder wieder mal ignoriert werden, ist fast immer ein veraltetes Ollama schuld, das vor dem Modellrelease installiert wurde. Erst `ollama -v` tippen und die Version prüfen, dann aktualisieren, dann erneut versuchen. Die Reihenfolge spart dir einen Abend Forenlektüre.

7

WAS DU REALISTISCH BEKOMMST

Zahlen statt Adjektive, Herkunft immer dabei

Tempo zuerst. Auf Apple-Rechnern der M-Klasse liefert das Modell in der 4-Bit-Fassung gemessene rund 40 Token pro Sekunde Textausgabe. Diese Zahl stammt aus einer einzelnen veröffentlichten Messung, nimm sie als Anhaltspunkt und nicht als Garantie, aber die Größenordnung stimmt: schneller, als du lesen kannst, und spürbar langsamer als die Cloud.

Können als zweites. Der Hersteller nennt für einen agentischen Programmier-Benchmark 61,7 Prozent und für einen Terminal-Benchmark 73 Prozent, beides deutliche Sprünge gegenüber dem Vorgänger. Wichtig für deine Einordnung: Diese Zahlen kommen vom Hersteller selbst, teils aus eigenen oder angepassten Testverfahren, und unabhängige Bestätigung stand beim Druck dieses Bandes noch aus. Behandle sie als Werbung mit Substanz, nicht als Messergebnis.

Für deinen Alltag heißt das konkret: Zusammenfassen, Umformulieren, Übersetzen, Klassifizieren, Herausziehen von Informationen aus Dokumenten, Bildbeschreibung sowie einfache bis mittlere Programmieraufgaben laufen gut, teils auf einem Niveau, das vor achtzehn Monaten nur die Cloud bot. Anspruchsvolles mehrstufiges Denken, wirklich elegante deutsche Prosa und Recherchen über viele Quellen bleiben schwächer als dein Abo-Modell, und daran ändert auch der Thinking-Regler nur graduell etwas.

Die vernünftige Konsequenz ist Arbeitsteilung statt Ersatz. Alles Vertrauliche und alles Massenhafte läuft lokal, alles, was maximale Qualität braucht, läuft weiter in der Cloud. Wer das von Anfang an so plant, wird nicht enttäuscht, sondern wundert sich eher, wie viel vom Tagesgeschäft in die erste Kategorie fällt.

MACH ES GENAU SO

Führe in der ersten Woche eine simple Strichliste mit zwei Spalten: Aufgaben, die das lokale Modell gut erledigt hat, und Aufgaben, bei denen du zur Cloud zurück musstest. Nach zehn Einträgen weißt du, wie deine persönliche Arbeitsteilung aussieht, und die ist verlässlicher als jeder Benchmark, weil sie mit deinen Texten gemessen wurde.

HIER ENDET DIE FREUDE

Deutsch ist die Achillesferse aller offenen Modelle, auch dieses. Es schreibt korrektes, brauchbares Deutsch, aber es schreibt Übersetzerdeutsch: grammatisch sauber, rhythmisch tot. Für interne Zusammenfassungen völlig egal, für alles, was unter deinem Namen rausgeht, nicht. Plane für publikationsfähige deutsche Texte weiterhin die Cloud oder eine gründliche eigene Überarbeitung ein.

8

DATENSCHUTZ OHNE MÄRCHENSTUNDE

Der Satz "keine Daten verlassen den Rechner" stimmt. Mit Sternchen.

Das Versprechen gilt für das Modell und seine Antworten, und es gilt vollständig: Nach dem Download kannst du das Netzwerkkabel ziehen, und alles läuft weiter. Kein Prompt, kein Dokument, keine Antwort verlässt die Maschine. Für Mandantendaten, Patientenunterlagen, Personalakten und alles andere, was du keinem amerikanischen Server anvertrauen darfst oder willst, ist das der ganze Punkt der Übung.

Das Versprechen gilt nicht automatisch für den Rest der Maschine. LM Studio und Ollama laden beim Start Modelllisten aus dem Netz, die Modelldateien selbst kommen von Hugging Face, und Programm-Updates telefonieren naturgemäß nach Hause. Nichts davon betrifft deine Inhalte, aber wer gegenüber einem Kunden oder Datenschutzbeauftragten behauptet, die Maschine sei komplett offline, sollte sie dann auch komplett offline betreiben.

Und das Versprechen kippt ins Gegenteil, wenn die Maschine selbst ungesichert ist. Deine Gesprächsverläufe liegen im Klartext auf der Platte. Ein lokales Modell auf einem unverschlüsselten Notebook, das im Zug liegen bleibt, ist datenschutzrechtlich ein größerer Unfall als jeder Cloud-Dienst mit ordentlichem Auftragsverarbeitungsvertrag.

Bleibt die Lizenz, und die ist die gute Nachricht des Kapitels: Qwen3.8-27B steht unter Apache 2.0. Das erlaubt Nutzung, Veränderung und kommerziellen Einsatz ausdrücklich und ist damit sauberer als vieles, was sich am Markt "offen" nennt und im Kleingedruckten Bedingungen versteckt.

MACH ES GENAU SO

Drei Handgriffe, einmalig: FileVault einschalten (Systemeinstellungen, Datenschutz und Sicherheit), damit die Platte verschlüsselt ist. Automatische Sperre auf höchstens fünf Minuten. Und wenn du die Installation beruflich nutzt, notiere dir Modellname, Version, Lizenz und Bezugsquelle in einer Zeile. Das ist deine Antwort, wenn ein Kunde oder Auditor fragt, was da eigentlich läuft, und du wirkst damit vorbereiteter als achtzig Prozent der Branche.

HIER ENDET DIE FREUDE

Das überzogene Versprechen ist die eigentliche Gefahr dieses Kapitels. Wer im Kundengespräch "hundertprozentig offline, hundertprozentig DSGVO-erledigt" sagt, hat im Zweifel beides nicht belegt und verbrennt Vertrauen genau bei dem Publikum, für das lokale KI gedacht ist. Die ehrliche Formulierung lautet: Die Inhalte bleiben auf der Maschine, die Maschine ist verschlüsselt, die Software aktualisiert sich übers Netz. Das ist stark genug, es braucht keine Übertreibung.

9

WAS ZUERST VERFÄLLT

Die Haltbarkeitstabelle dieses Bandes

Zuerst verfällt der Modellname. Zwischen Qwen 3.5 im März, 3.6 im April und 3.8 am 14. August 2026 lagen fünf Monate, und der Nachfolger ist beim Lesen dieser Zeilen womöglich schon da. Wenn ja: Ersetze den Namen und behalte den Rest. Such nach dem aktuellen 27B-Modell in MLX und 4 Bit, prüfe die Dateigröße gegen deinen Speicher, stell die drei Regler ein, kenne die vier Momente. Das Verfahren überlebt das Produkt.

Danach verfallen die Programmversionen, insbesondere alles, was in diesem Band über Bildfunktionen und Gesprächsvorlagen steht. Solche Fehler leben Wochen, nicht Jahre, und genau deshalb steht auf der ersten Seite ein Datum. Ein Band ohne sichtbares Stand-Datum wäre in diesem Feld keine Hilfe, sondern eine Falle für alle, denen er in achtzehn Monaten in die Hände fällt.

Am längsten hält das Kapitel über Speicher und Bandbreite, weil es Physik beschreibt und keine Software. Die Rechnung Gesamtspeicher minus System minus Modell minus Gesprächsspeicher wird auch in fünf Jahren die einzige sein, die zählt. Die Zahlen darin werden andere sein, die Rechnung nicht.

MACH ES GENAU SO

Wenn du diesen Band weitergibst, gib ihn mit einem Satz weiter: "Stand August 2026, die Kapitel 1, 2, 4 und 8 stimmen noch, den Modellnamen musst du prüfen." Dieser eine Satz macht aus einem alternden Dokument ein brauchbares, und er kostet dich nichts.

Und damit bist du durch. Das Modell läuft, du weißt, warum es läuft, und du weißt, was du prüfst, wenn es das nicht tut. Mehr steht in keinem dicken Fachbuch, es steht dort nur langsamer.

Frank Tentler.ai

[Created with human/ai co-intelligence]

f2b2

IN EINEM ABEND VON NULL AUF FUNKTIONIEREND.

Ein Sprachmodell auf deinem eigenen Mac: ohne Cloud, ohne Abo, ohne dass ein einziges Zeichen deinen Schreibtisch verlässt. Dieser Band zeigt den Weg in anderthalb Stunden, benennt die vier Momente, in denen dein Rechner kaputt wirkt, und erklärt die fünf Begriffe, nach denen du nie wieder ein Forum brauchst.

Eine Reihe für Menschen, die keine Zeit für Vorworte haben.

f2b2 · Band 1 · Stand 17.08.2026
Frank Tentler.ai · [Created with human/ai co-intelligence]