Bundesministerium für Wirtschaft und Energie

Mittelstand-Digital

Mittelstand-Digital Zentrum Spreeland

Leitfaden

Automatisierte Antwortentwürfe für Buchungsanfragen



www.digitalzentrum-spreeland.de

Stand November 2026

Diese Anleitung hat zwei Teile.

TEIL A

Teil A richtet sich an alle, die Anfragen beantworten. Zwei Seiten, ohne Technik.

TEIL B

Teil B beschreibt die einmalige Einrichtung Schritt für Schritt. Er richtet sich an den IT-Dienstleister und ist für die tägliche Arbeit nicht nötig.

Inhalt

1 Für die tägliche Arbeit 3

1.1 Was das System tut 3
1.2 Ihr Arbeitsablauf 3
1.3 Was das System nicht tut 3
1.4 Wenn kein Entwurf entsteht 4
1.5 Die Texte ändern 4
1.6 Wenn etwas nicht funktioniert 4

2 Einrichtung Schritt für Schritt 5

2.1 Wie das System aufgebaut ist 5
2.2 Andere Mail-Dienste 6
Schritt 1 Voraussetzungen prüfen 6
Schritt 2 Ollama installieren und Sprachmodell laden 7
Schritt 3 n8n starten 7
Schritt 4 Zugang zum Postfach einrichten (Beispiel Gmail) 8
Schritt 5 Gmail-Zugang in n8n hinterlegen 8
Schritt 6 Ablauf anlegen 9
Schritt 7 Filter und Stichtag setzen 9
Schritt 8 Nachricht laden 10
Schritt 9 Sprachmodell anbinden 10
Schritt 10 Systemanweisung hinterlegen 11
Schritt 11 Entwurf ablegen 11
Schritt 12 Testlauf 12
Schritt 13 Dauerbetrieb einrichten 12
Schritt 14 Begleiteter Testbetrieb 12

3 Fehlerbehebung 13

4 Was regelmäßig zu tun ist 14

TEIL A

1 Für die tägliche Arbeit

Für Mitarbeitende, die Anfragen beantworten

1.1 Was das System tut

Wenn eine Anfrage per E-Mail eingeht, liest das System sie und legt eine fertig formulierte Antwort in Ihren Entwürfe-Ordner. Sie öffnen den Entwurf, lesen ihn durch, ergänzen bei Bedarf etwas und klicken auf Senden.

Das gilt für Anfragen über das Kontaktformular Ihrer Webseite ebenso wie für direkt geschriebene E-Mails.

Es wird nichts automatisch verschickt. Jede Nachricht verlässt das Haus erst, wenn ein Mensch darauf geschaut und auf Senden geklickt hat.

1.2 Ihr Arbeitsablauf

1. Den Ordner „Entwürfe“ in Gmail öffnen.
2. Entwurf anklicken und durchlesen.
3. Passt alles? Auf Senden klicken.
4. Passt etwas nicht? Text ändern und senden — oder den Entwurf löschen und die Anfrage wie bisher von Hand beantworten.

Der Empfänger ist bereits eingetragen. Das System übernimmt die Absenderadresse der Anfrage, und der Entwurf erscheint im selben Gesprächsverlauf. Sie müssen die Adresse nicht mehr heraussuchen und kopieren.

1.3 Was das System nicht tut

- Es sagt nicht verbindlich zu, ob etwas frei ist. Die Antwort verweist auf die Onlinebuchung oder die zuständige Stelle.
- Es bestätigt keine Buchungen eigenständig und macht keine Preiszusagen für Sonderfälle.
- Es formuliert die Antwort selbst, auf Grundlage der Vorgaben, die Sie hinterlegt haben. Ein Sprachmodell kann sich dabei irren. Prüfen Sie Angaben zu Preisen, Terminen und Namen deshalb vor dem Senden.
- Es verschiebt und löscht keine Nachrichten. Alle Anfragen bleiben im Posteingang.

1.4 Wenn kein Entwurf entsteht

Das ist kein Fehler, sondern Absicht. Das System bearbeitet nur Nachrichten, die zum eingestellten Filter passen, etwa ungelesene Mails im Posteingang ab einem Stichtag. Eigene Mails und Nachrichten außerhalb des Filters bleiben unberührt. Beantworten Sie diese Mails wie bisher.

1.5 Die Texte ändern

Was im Entwurf steht, bestimmt eine Systemanweisung im Feld „AI Agent“ in n8n: Ton, Anrede, feste Hinweise, Kontaktdaten und Signatur. Diese Vorgaben gehören Ihnen. Ändern sich Öffnungszeiten, Preise oder eine Formulierung, werden sie dort angepasst. Kenntnisse des Systems sind dafür nicht nötig.

Legen Sie fest, wer bei Ihnen dafür zuständig ist. Diese Person sollte wissen, wo die Systemanweisung steht.

Wichtig beim Bearbeiten: Ausdrücke in doppelten geschweiften Klammern wie `{{ $json.text }}` setzt n8n beim Ausführen durch echte Inhalte ein. Sie müssen unverändert stehen bleiben.

1.6 Wenn etwas nicht funktioniert

Entstehen über längere Zeit keine Entwürfe mehr, melden Sie sich bei Ihrem IT-Dienstleister oder der am Ende genannten Ansprechperson. Ihre Arbeit ist dadurch nicht blockiert: Alle Anfragen liegen weiterhin vollständig im Posteingang und können wie zuvor beantwortet werden.

TEIL B

2 Einrichtung Schritt für Schritt

Für den IT-Dienstleister · einmalig · etwa 90 Minuten

Diese Schritte werden einmalig durchgeführt. Beschrieben ist der Weg unter Windows mit PowerShell am Beispiel eines Gmail-Postfachs. Andere Mail-Dienste lassen sich ebenso anbinden, siehe „Andere Mail-Dienste“. Rechnen Sie mit etwa 90 Minuten, einschließlich des Herunterladens des Sprachmodells.

2.1 Wie das System aufgebaut ist

Es besteht aus zwei Programmen, die zusammenarbeiten:

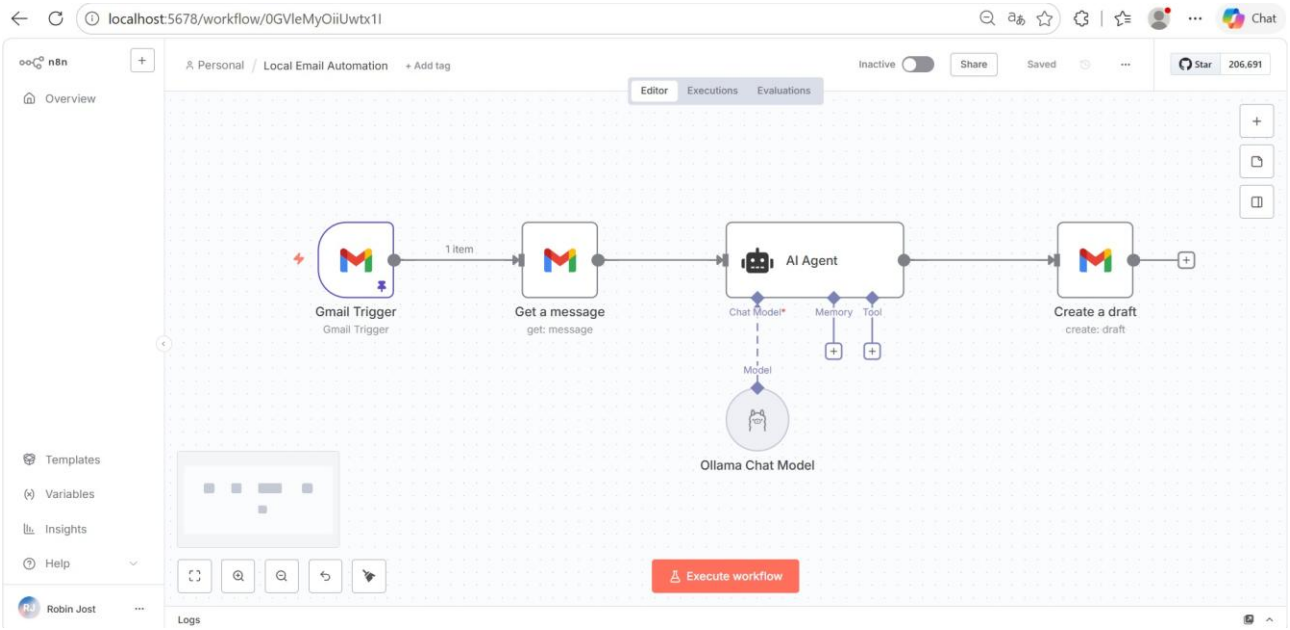
`n8n`

Steuert den Ablauf: holt neue Nachrichten aus Gmail, übergibt sie an das Sprachmodell und legt die Antwort als Entwurf ab.

`Ollama`

Führt das Sprachmodell auf Ihrem eigenen Rechner aus. Es formuliert den Antworttext nach Ihren Vorgaben.

Beide müssen laufen. Fehlt eines, entstehen keine Entwürfe.



Der Ablauf in n8n: „Gmail Trigger“ erkennt eine neue Nachricht, „Get a message“ lädt sie vollständig, der „AI Agent“ formuliert mit dem lokalen Sprachmodell („Ollama Chat Model“) die Antwort, „Create a draft“ legt sie als Entwurf ab.

n8n und das Sprachmodell laufen lokal auf Ihrem eigenen Rechner beziehungsweise Server. Der Inhalt der Anfragen wird an keinen KI-Anbieter übertragen. Das Postfach selbst liegt weiterhin bei Google; n8n greift über die offizielle Gmail-Schnittstelle darauf zu.

2.2 Andere Mail-Dienste

Gmail dient in diesem Leitfaden als Beispiel. Der Ablauf funktioniert auch mit anderen Mail-Diensten. Dann werden die drei Gmail-Felder durch die passenden Felder ersetzt; Sprachmodell, Systemanweisung und Testlauf bleiben gleich.

Mail-Dienst	Felder in n8n
Gmail / Google Workspace	Gmail Trigger, Get a message, Create a draft — wie in diesem Leitfaden beschrieben.
Microsoft 365 / Outlook	Microsoft Outlook Trigger und Microsoft Outlook mit „Draft / Create“. Der Zugang wird über Microsoft Entra statt über die Google Cloud Console eingerichtet.
Andere Anbieter über IMAP	„Email Trigger (IMAP)“ liest neue Nachrichten mit Server, Benutzername und Passwort. Entwürfe kann n8n über IMAP nicht selbst ablegen; dafür ist ein kleiner Zusatzdienst nötig, der die Antwort per IMAP im Entwürfe-Ordner speichert.

Schritt 1 Voraussetzungen prüfen

PowerShell öffnen: Windows-Taste drücken, „PowerShell“ eingeben, Enter. Prüfen, ob Node.js vorhanden ist:

EINGABE

```
node -v
```

ERWARTETE AUSGABE — EINE VERSIONSNUMMER, MINDESTENS V20

```
v22.19.0
```

Falls „Der Befehl … wurde nicht gefunden“ erscheint: Node.js von nodejs.org herunterladen, LTS-Version wählen, installieren, PowerShell schließen und neu öffnen.

Tipp: Für das Sprachmodell werden mindestens 16 GB Arbeitsspeicher und 10 GB freier Speicherplatz empfohlen. Auf schwächeren Rechnern ein kleineres Modell wählen, zum Beispiel `llama3.2:3b`. Die Antworten werden dann etwas einfacher.

Schritt 2 Ollama installieren und Sprachmodell laden

Ollama von ollama.com/download für Windows herunterladen und installieren. Das Programm startet danach automatisch im Hintergrund, erkennbar am Symbol in der Taskleiste. Nach einem Neustart des Rechners startet es von selbst.

EINGABE — INSTALLATION PRÜFEN

```
ollama --version
```

EINGABE — SPRACHMODELL LADEN

```
ollama pull llama3.1:8b
```

Das Modell ist etwa 5 GB groß. Der Download dauert je nach Verbindung einige Minuten.

EINGABE — MODELL TESTEN

```
ollama run llama3.1:8b "Schreibe einen kurzen Gruß auf Deutsch."
```

Es erscheint nach einigen Sekunden ein kurzer Text. Die erste Antwort dauert länger, weil das Modell erst geladen wird.

KONTROLLE — LÄUFT DER DIENST?

```
Invoke-RestMethod http://localhost:11434
```

ERWARTETE AUSGABE

```
Ollama is running
```

Schritt 3 n8n starten

EINGABE — EIN NEUES POWERSHELL-FENSTER

```
npx n8n
```

Beim ersten Start werden Programmteile heruntergeladen, das dauert einige Minuten. Danach erscheint:

```
Editor is now accessible via:
http://localhost:5678
```

Diese Adresse im Browser öffnen. Beim ersten Aufruf wird ein Benutzerkonto angelegt. Die Zugangsdaten notieren und beim Betrieb hinterlegen.

Dieses Fenster muss geöffnet bleiben. Wird es geschlossen, ist n8n beendet und es entstehen keine Entwürfe mehr. Für den Dauerbetrieb siehe Schritt 13.

Schritt 4 Zugang zum Postfach einrichten (Beispiel Gmail)

Damit n8n auf das Gmail-Postfach zugreifen darf, wird einmalig ein Zugang in der Google Cloud Console angelegt. Kosten entstehen dabei nicht.

1. console.cloud.google.com öffnen und mit dem Google-Konto des Postfachs anmelden.
2. Oben ein neues Projekt anlegen, zum Beispiel „n8n Antwortentwürfe“.
3. Unter „APIs & Dienste“, „Bibliothek“ nach „Gmail API“ suchen und auf „Aktivieren“ klicken.
4. Unter „OAuth-Zustimmungsbildschirm“ einen App-Namen und eine Support-Adresse eintragen. Nutzertyp „Intern“ bei Google Workspace, sonst „Extern“. Bei „Extern“ die Adresse des Postfachs als Testnutzer eintragen.
5. Unter „Anmeldedaten“ auf „Anmeldedaten erstellen“, dann „OAuth-Client-ID“. Anwendungstyp „Webanwendung“ wählen.
6. Bei „Autorisierte Weiterleitungs-URIs“ die folgende Adresse eintragen und speichern.

```
http://localhost:5678/rest/oauth2-credential/callback
```

Google zeigt anschließend Client-ID und Clientschlüssel an. Beide Werte notieren, sie werden im nächsten Schritt gebraucht.

Hinweis: Bei Nutzertyp „Extern“ im Status „Testen“ läuft die Anmeldung nach sieben Tagen ab. Für den Dauerbetrieb die App unter „OAuth-Zustimmungsbildschirm“ auf „In Produktion“ stellen oder ein Google-Workspace-Konto mit Nutzertyp „Intern“ verwenden.

Schritt 5 Gmail-Zugang in n8n hinterlegen

1. In n8n links auf „Overview“, oben auf den Reiter „Credentials“, dann „Create credential“.
2. „Gmail OAuth2 API“ auswählen.
3. Client-ID und Clientschlüssel aus Schritt 4 eintragen.

4. Auf „Sign in with Google“ klicken, das Konto des Postfachs wählen und den Zugriff erlauben.
5. Es muss „Account connected“ erscheinen. Danach auf „Save“.

Schritt 6 Ablauf anlegen

In n8n links auf „Overview“ klicken, dann oben rechts auf „Create Workflow“. Es öffnet sich eine leere Arbeitsfläche. Über das Plus-Zeichen rechts oben werden die Felder nacheinander hinzugefügt; n8n verbindet jedes neue Feld automatisch mit dem vorherigen.

Feld	So wird es hinzugefügt	Aufgabe
Gmail Trigger	Nach „Gmail Trigger“ suchen, Ereignis „Message Received“ wählen.	Prüft regelmäßig, ob neue Nachrichten eingegangen sind.
Get a message	Nach „Gmail“ suchen, dann die Aktion „Get a message“.	Lädt die vollständige Nachricht mit Absender, Betreff und Text.
AI Agent	Nach „AI Agent“ suchen.	Formuliert die Antwort nach der Systemanweisung.
Ollama Chat Model	Am Anschluss „Chat Model“ unter dem AI Agent auf das Plus, dann „Ollama Chat Model“.	Das lokale Sprachmodell, das der AI Agent verwendet.
Create a draft	Nach „Gmail“ suchen, dann die Aktion „Create a draft“.	Legt die Antwort als Entwurf in Gmail ab.

Die Hauptreihe lautet Gmail Trigger → Get a message → AI Agent → Create a draft. Das Ollama Chat Model hängt unterhalb am AI Agent und nicht in der Hauptreihe. Das Ergebnis entspricht der Abbildung unter „Wie das System aufgebaut ist“.

In jedem der drei Gmail-Felder oben bei „Credential to connect with“ den Zugang aus Schritt 5 auswählen. Anschließend oben rechts auf „Save“ und dem Ablauf einen Namen geben, zum Beispiel „Antwortentwürfe“. Die Einstellungen der einzelnen Felder folgen in den Schritten 7 bis 11.

Schritt 7 Filter und Stichtag setzen

Doppelklick auf „Gmail Trigger“. „Poll Times“ auf „Every Minute“, „Event“ auf „Message Received“. Unter „Filters“ das Feld „Search“ hinzufügen. Dort steht der Filter, der bestimmt, welche Nachrichten bearbeitet werden:

```
in:inbox is:unread -from:me after:2026/09/01
```

Begriff	Bedeutung
in:inbox	Nur Nachrichten im Posteingang.
is:unread	Nur ungelesene Nachrichten.
-from:me	Keine eigenen Nachrichten.
after:	Der Stichtag. Nachrichten davor bleiben unberührt.

Das Datumsformat ist Jahr/Monat/Tag, also `2026/09/01` — nicht `01.09.2026`. Soll nur eine bestimmte Art von Anfragen bearbeitet werden, lässt sich der Filter erweitern, etwa um `subject:Anfrage` für Mails aus dem Kontaktformular.

Anschließend oben rechts auf „Save“.

Schritt 8 Nachricht laden

Doppelklick auf „Get a message“. Einstellungen:

Einstellung	Wert
Resource / Operation	Message / Get
Message ID	`{{ $json.id }}`
Simplify	Ausgeschaltet, damit der vollständige Text übergeben wird.

Schritt 9 Sprachmodell anbinden

Doppelklick auf „Ollama Chat Model“. Bei „Credential to connect with“ einen neuen Zugang anlegen und als Base URL eintragen:

```
http://localhost:11434
```

Läuft n8n später in Docker, lautet die Adresse stattdessen `http://host.docker.internal:11434`. Unter „Model“ das in Schritt 2 geladene Modell `llama3.1:8b` auswählen.

Schritt 10 Systemanweisung hinterlegen

Doppelklick auf „AI Agent“. „Source for Prompt“ auf „Define below“ stellen und bei „Prompt“ eintragen:

```
Absender: {{ $json.from.text }}
Betreff: {{ $json.subject }}

{{ $json.text }}
```

Unter „Options“ das Feld „System Message“ hinzufügen. Hier stehen die Vorgaben des Betriebs. Die Angaben in eckigen Klammern ersetzen:

```
Du schreibst Antwortentwürfe auf E-Mail-Anfragen für [Name des Betriebs].
Schreibe auf Deutsch, freundlich und in der Sie-Form.
Mache keine verbindlichen Zusagen zu Verfügbarkeit, Preisen oder Terminen.
Für Buchungen verweise auf: [Buchungslink]
Telefon: [Telefonnummer] · Erreichbar: [Zeiten]
Unterschreibe mit: [Name oder Team]
Gib nur den Text der Antwort aus, ohne Betreff.
```

Die Anschlüsse „Memory“ und „Tool“ bleiben leer. Anschließend „Save“.

Schritt 11 Entwurf ablegen

Doppelklick auf „Create a draft“. Einstellungen:

Einstellung	Wert
Resource / Operation	Draft / Create
Subject	`Re: {{ $('Get a message').item.json.subject }}`
Message	`{{ $json.output }}`
Options · To Email	`{{ $('Get a message').item.json.from.value[0].address }}`
Options · Thread ID	`{{ $('Get a message').item.json.threadId }}`

Die Thread ID sorgt dafür, dass der Entwurf im Gesprächsverlauf der Anfrage erscheint. Anschließend „Save“.

Schritt 12 Testlauf

1. In n8n unten auf „Execute workflow“ klicken.
2. Erst jetzt von einer anderen Adresse eine Testanfrage an das Postfach schicken.
3. Bis zu einer Minute warten. Die Felder färben sich nacheinander grün. Das Formulieren dauert je nach Rechner einige Sekunden bis eine Minute.
4. In Gmail den Ordner „Entwürfe“ öffnen und den Entwurf prüfen: Stimmen Empfänger und Betreff? Sind Ton und Inhalt passend? Erscheint er im Gesprächsverlauf der Anfrage?

Passt der Inhalt nicht, die Systemanweisung aus Schritt 10 anpassen und den Test wiederholen.

Schritt 13 Dauerbetrieb einrichten

Im Testmodus arbeitet n8n nur, solange der Editor die Ausführung anstößt. Für den laufenden Betrieb den Schalter oben rechts von „Inactive“ auf „Active“ stellen.

Die Durchläufe erscheinen dann nicht mehr im Editor, sondern oben im Reiter „Executions“. Dort ist jeder Vorgang mit allen Zwischenschritten nachvollziehbar.

Automatischer Start nach einem Neustart

Ollama startet unter Windows von selbst. n8n muss nach einem Neustart des Rechners ebenfalls wieder starten. Drei Möglichkeiten, in aufsteigender Zuverlässigkeit:

Variante	Eigenschaften
Autostart-Ordner	Eine Verknüpfung in den Autostart legen. Einfach, aber das Fenster bleibt sichtbar und kann versehentlich geschlossen werden.
Windows-Dienst	n8n mit einem Werkzeug wie NSSM als Dienst registrieren. Läuft unsichtbar im Hintergrund und startet automatisch.
Docker	n8n und Ollama als Container. Empfohlen für einen Server. Startet nach Stromausfall selbstständig neu.

Bei Docker ist zusätzlich ein fester Verschlüsselungsschlüssel zu setzen (`N8N_ENCRYPTION_KEY`). Ohne ihn sind nach einem Verlust des Datenbereichs alle hinterlegten Zugangsdaten unbrauchbar.

Schritt 14 Begleiteter Testbetrieb

Vor der Umstellung sollte das System einige Wochen mitlaufen. Die Mitarbeitenden vermerken bei jedem Entwurf kurz: passte, musste geändert werden, oder unbrauchbar. Daraus ergibt sich, ob die Systemanweisung und der Filter stimmen, bevor Routine einkehrt.

3 Fehlerbehebung

Beobachtung	Ursache und Abhilfe
Es kommen keine Nachrichten an, obwohl der Ablauf läuft	Filter und Stichtag prüfen. Im Testmodus muss der Ablauf zuerst laufen, danach die Testmail eintreffen. Bereits gelesene Mails werden wegen is:unread übergangen.
ECONNREFUSED beim Feld „Ollama Chat Model“	Ollama läuft nicht. Symbol in der Taskleiste prüfen oder Ollama über das Startmenü neu starten.
„model not found“	Das Modell ist nicht geladen oder anders benannt. Mit ollama list prüfen, ggf. ollama pull wiederholen.
Antwort dauert sehr lange oder bricht ab	Der Rechner ist für das Modell zu schwach. Ein kleineres Modell laden und in Schritt 9 auswählen.
„invalid_grant“ oder Gmail-Anmeldung abgelaufen	Die Google-App steht noch im Status „Testen“ (siehe Schritt 4). In n8n erneut „Sign in with Google“ ausführen und die App auf „In Produktion“ stellen.
Entwurf erscheint nicht im Gesprächsverlauf	Die Thread ID in Schritt 11 fehlt oder ist falsch eingetragen.
Entwurf wird doppelt angelegt	Der Ablauf läuft zweimal — etwa im Test- und im Dauerbetrieb gleichzeitig. Eine der beiden Ausführungen beenden.
Ton oder Inhalt passt nicht	Die Systemanweisung in Schritt 10 genauer fassen, etwa mit einem Beispiel für eine gute Antwort.

4 Was regelmäßig zu tun ist

- Nach einer Passwortänderung oder wenn der Google-Zugang widerrufen wurde: in n8n erneut „Sign in with Google“ ausführen.
- Bei geänderten Angaben wie Öffnungszeiten, Preisen oder Ansprechpersonen: die Systemanweisung prüfen.
- Gelegentlich in den Reiter „Executions“ schauen, ob Fehler aufgetreten sind.
- Nach Aktualisierungen von n8n oder Ollama einen Testlauf durchführen.

ANSPRECHPARTNER

Robin Jost

Wissenschaftlicher Mitarbeiter und Trainer für Künstliche Intelligenz

Mittelstand-Digital Zentrum Spreeland

Telefon +49 (0)355 69-2474 · robin.jost@b-tu.de