



AUF DER SUCHE NACH  
**RELEVANZ IN**  
**MITGELTENDEN**  
**UNTERLAGEN**

**REConf 2026**

Nevres Balic, FORVIA Hella

Andreas Willert, SodiusWillert

# KI-Souveränität: Vom Hype zur industriellen Realität

Ein aktueller Beitrag in der Elektronik  
Praxis bringt es auf den Punkt:

Die Industrie braucht keine allwissenden  
Cloud-LLMs, sondern kontrollierbare,  
ressourcenschonende Modelle direkt  
vor Ort. „Self-Hosted LLM“



[KI & Intelligent Edge](#) > Der Weg zur KI-Souveränität in der Industrie

Raus aus der Blackbox

## Der Weg zur KI-Souveränität in der Industrie

21.04.2026 · Von [Manuel Christa](#) · 3 min Lesedauer · 

Die Industrie braucht keine allwissenden Sprachmodelle aus der US-Cloud, sondern smarte, ressourcenschonende LLMs direkt an der Maschine. Eine aktuelle Debatte zeigt: Um Vendor-Lock-ins zu vermeiden und das eigene geistige Eigentum zu schützen, führt an lokaler Inferenz und Open-Source-Strategien kein Weg vorbei.

# AI funktioniert gut, solange die Informationen linear strukturiert sind

**Mitgeltende Unterlagen** sind keine linearen Texte:

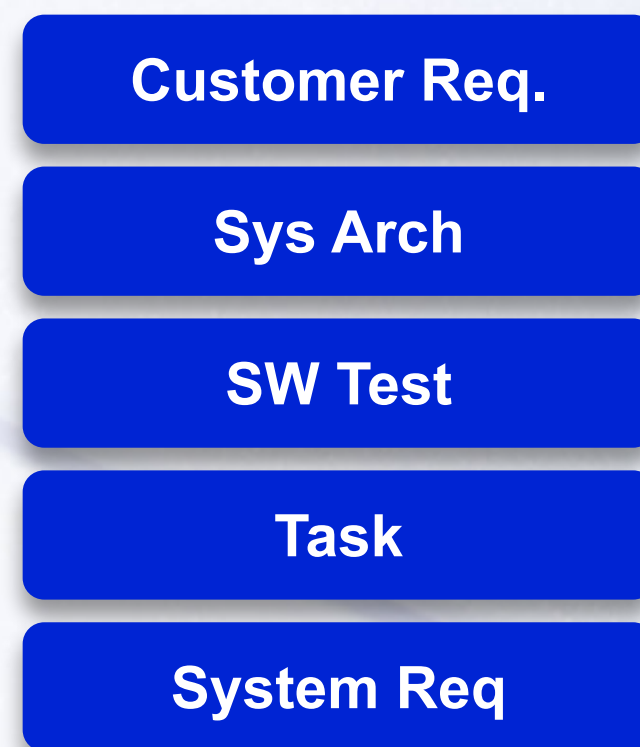
- Verweise über mehrere Kapitel hinweg
- Kaskadierte Tabellen als zentrale Wissensquelle (z. B. *Spektren, Klassifikationen*)
- Überlagerte Bedingungen, Ausnahmen, Geltungsbereiche
- Abhängigkeiten zwischen Dokumentteilen

**Die relevante Antwort steht selten an einer Stelle** sie entsteht aus der Verknüpfung mehrerer.

# VON DATEN ZU BEDEUTUNG

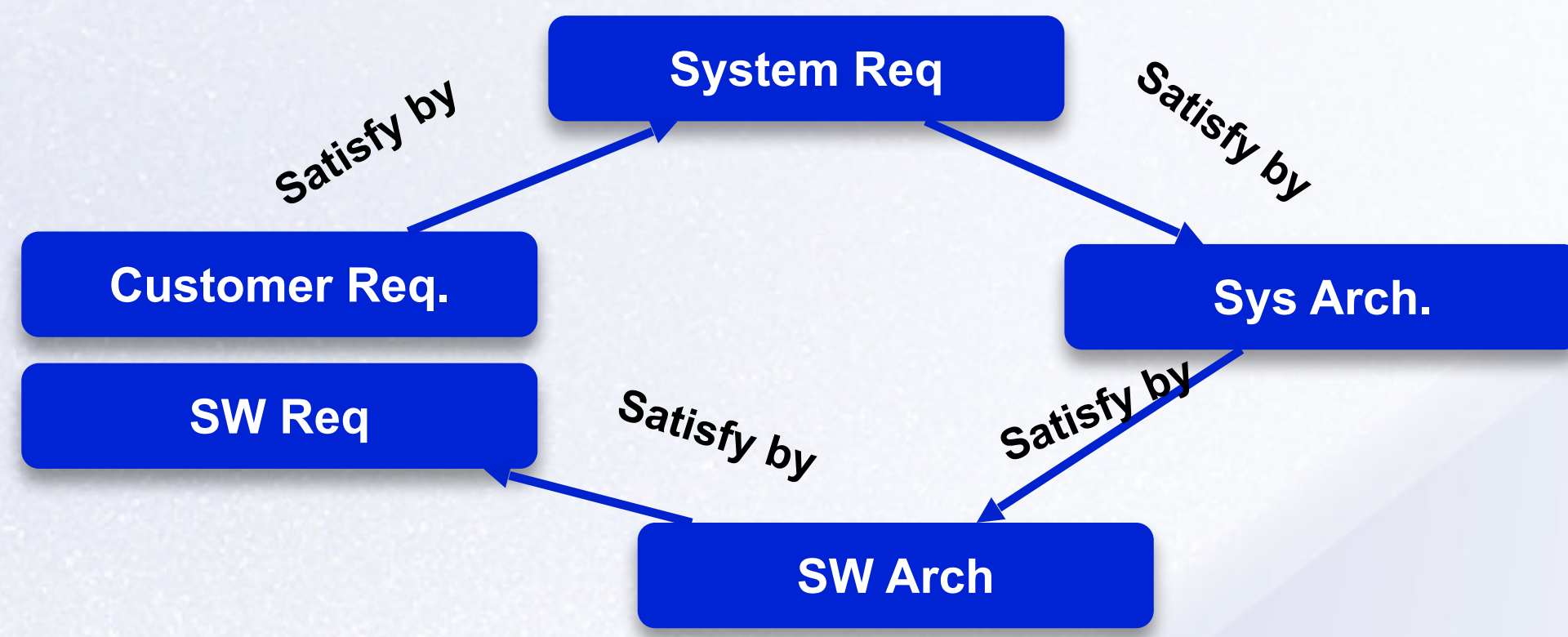
*Realität wird über Beziehungen beschrieben – nicht über Datenmenge*

## Datensammlung



- Isolierte Artefakte
- Kontext implizit
- Relevanz unklar

## Beziehungsnetz



- Knoten = Entitäten
- Kanten = Bedeutung
- Kontext explizit

***Relevanz entsteht durch Beziehungen - nicht durch mehr Dokumente***

# WARUM GRAPHEN IM ANFORDERUNGSMANAGEMENT HELFEN

**Anforderungsmanagement = Beziehungsmanagement**

# AUF DER SUCHE NACH RELEVANZ IN MITGELTENDEN UNTERLAGEN

Anforderungsmanagement als explizite und implizite Ontologie

## Impliziter Kontext (strukturell & kulturell)

- Kapitel und Dokumentstruktur
- Abstraktionsebenen
- Sprache und Terminologie
- Unternehmens und Kundenkultur
- „Wie Anforderungen gemeint sind“

## Expliziter Kontext (normativ)

- Traceability Anforderungen
- Definierte Artefakt-Typen
- Zulässige Beziehungen
- Compliance Regeln (z.B. ASPICE)

***Relevanz entsteht aus beiden Ebenen – auch wenn nur eine formal modelliert ist***

# AUF DER SUCHE NACH RELEVANZ IN MITGELTENDEN UNTERLAGEN

## Was lässt sich konkret verbessern?

Wenn Anforderungsmanagement bereits durch **explizite und implizite** Ontologien geprägt ist, muss dieses **Beziehungswissen** den **LLMs gezielt nutzbar** gemacht werden.

### Heute vorhanden

- Explizite Beziehungen  
(Traceability zwischen Anforderungen, Architektur, Tests)
- Impliziter Kontext  
(Kapitelstruktur, Sprache, Abstraktion, Kultur)
- Wissen verteilt & schwer nutzbar

### Gezielt nutzbar durch Beziehungen & LLMs

- Zusammenhänge erklären
- Ableitungen unterstützen
- Auswirkungen verständlich machen

*Der Nutzen entsteht direkt in bestehenden Aufgaben:  
Analyse · Ableitung · Review · Change · Kommunikation*

# AUF DER SUCHE NACH RELEVANZ IN MITGELTENDEN UNTERLAGEN

## Konkrete Anwendungsfälle im Engineering-Alltag

Beziehungen machen Zusammenhänge sichtbar – LLMs helfen, sie zu analysieren, zu erklären und zu erweitern

### Typische Problemstellungen

- Zusammenhänge visualisieren
- Anforderungen clustern
- Inkonsistenzen & Lücken erkennen
- Impliziten Kontext erklären

### Konkreter Mehrwert

- **Verbesserungsvorschläge** ableiten
- **Tiefergehende Analysen** auf Ontologie – Basis
- **Angebotsunterstützung** (RFQ – Analyse)
- **Reuse** aus bestehenden Projekten

*Der Nutzen entsteht direkt im bestehenden Workflow – ohne zusätzliche Dokumente oder neue Prozesse*

# AUF DER SUCHE NACH RELEVANZ IN MITGELTENDEN UNTERLAGEN

## POC Use Case

### Herausforderung

- Angebotsdokumentation mit hoher Zahl an Unterlagen (OEM ~170 Dokumente)
- Mischung aus Kundenvorgaben und globalen gesetzlichen Anforderungen.
- Nicht alle Dokumente relevant (z. B. Schraubenvorschrift, verklebtes Steuergerät)
- Teils impliziter Einfluss der Unterlagen (z.B. Absicherung nur durch Fahrversuche)

### Zieldefinition

- Automatisierte Analyse von mitgeltenden Unterlagen
- Relevanzanalyse der mitgeltenden Unterlagen
- Beziehungsanalyse zwischen RFQ-Anforderungen zu den mitgeltenden Unterlagen

*Ziel: Relevanz früh erkennen – bevor Kosten erstellen*

# **FORVIA**

## **FORVIA**



# Mehrfache Verweise *(Mult-Hop Problem)*

## 8.2 Temperature load spectra

In addition to the specification of minimum ambient temperature  $T_{\min}$  and maximum ambient temperature  $T_{\max}$ , the distribution indicating how long the component is exposed to the various temperatures between  $T_{\min}$  and  $T_{\max}$  must also be specified, in order to fully describe the temperature load to which the component is exposed at the point of use in the vehicle.

For vehicles with an alternative powertrain, a distinction must be made between operating situations "Driving," "Charging," "Preconditioning," and "On-grid parking" and the respective **temperature load spectra** must be specified both for the ambient temperature and for the coolant circuit temperature.

Table 45: Temperature load spectrum

| Temperature in °C               | Distribution |
|---------------------------------|--------------|
| $T_{\text{field},1} = T_{\min}$ | $p_1$        |
| $T_{\text{field},2}$            | $p_2$        |
| ...                             | ...          |
| $T_{\text{field},n} = T_{\max}$ | $p_n$        |

This is based mainly on field measurements and technical experience.

Typical **temperature load spectra** for "Driving" operating situation in terms of different installation areas are indicated in Appendix B.

## Appendix B (normative)

### Typical temperature load spectra for different installation areas

Table 94: Overview of installation areas, typical spectra, and temperature rises

| Installation area of the component            | Spectrum no. | Temperature rise in K |
|-----------------------------------------------|--------------|-----------------------|
| Vehicle interior, without special requirement | 1            | 36                    |
| Body-mounted, without special requirements    | 1            | 36                    |
| Vehicle interior, exposed to solar radiation  | 2            | 46                    |
| Body-mounted, roof                            | 2            | 46                    |
| Engine compartment, but not on the engine     | 3            | 60                    |
| On the radiator                               | 3            | 60                    |
| Engine-mounted                                | 4            | 75                    |
| Transmission-mounted                          | 4            | 75                    |

## B.3 Temperature load spectrum 3

Table 97: Temperature load spectrum 3

| Temperature in °C | Distribution in % |
|-------------------|-------------------|
| -40               | 6                 |
| 23                | 20                |
| 65                | 65                |
| 115               | 8                 |
| 120               | 1                 |

# Gemini Halluzination

- **Funktionszustände (Functional States - FS):** Definieren das Verhalten während und nach einer Belastung:
  - **FS A:** Alle Funktionen müssen während und nach der Prüfung erfüllt sein.
  - **FS B:** Funktionen können eingeschränkt sein, müssen aber nach der Prüfung selbstständig wieder in den Sollzustand zurückkehren.
  - **FS C:** Funktionen können eingeschränkt sein, erfordern jedoch einen Reset oder Neustart.
  - **FS D:** Permanente Funktionsbeeinträchtigung/Schaden.
  - **FS E:** Funktionen nicht erfüllt, aber die Komponente darf keinen Brand oder Sicherheitsrisiko darstellen.

## 4.2.2 Functional state A

The DUT fulfills all functions as specified.

## 4.2.3 Functional state B

Not used.

## 4.2.4 Functional state C

The DUT does not fulfill one or more functions during exposure to the test parameters. After exposure to the test parameters, the DUT must immediately return to functional state A automatically or by means of the external triggers specified in the Performance Specification. Undefined functions are not permissible at any time.

# AI kann das – aber zu welchem Preis?

**Moderne LLMs liefern heute bereits relativ gute Ergebnisse**

auch bei komplexen, vernetzten Informationen.

**Aber:**

Mitgeltende Unterlagen sind vertraulich

Eine Verarbeitung in Public-Cloud-Systemen ist nicht zulässig

Nachvollziehbarkeit und Kontrolle sind zwingend erforderlich

Die Lösung muss lokal, kontrollierbar und reproduzierbar sein, der Betrieb großer Modelle on-prem ist jedoch oft wirtschaftlich nicht darstellbar.

**Genau hier setzt unser Ansatz an.**

# Lokale KI-Analyse vertraulicher OEM-Dokumente

1. Strukturtreues Chunking (Kapitel, Tabellen, Anhänge)
2. Semantische Analyse mit Qwen 3.5 (lokal, Open Weights)
3. Hybride Speicherung: Vector DB + Postgres
4. Multi-Step Retrieval & Evidenzaufbau
5. Antwortgenerierung mit Qwen 3.5 (lokal, evidenzbasiert)

## Technologie:

Qwen 3.5 (32B – 122B), lokal betreibbar

Vector DB + Postgres

## Ziel:

Hohe Qualität bei voller Datenhoheit und moderater Rechenleistung

# I. Strukturtreues Chunking – Kontext erhalten statt zerstören

## I. Preprocessing der Dokumente

- Zerlegung entlang natürlicher Strukturen (*Kapitel, Tabellen, Anhänge – kein Kontextverlust*)
- Abgrenzung zu token-basiertem Chunking (*kein „Zerschneiden“ von Bedeutung*)
- Grundlage für Multi-Hop / strukturorientiertes RAG
- Speicherung inkl. Positions- und Strukturmetadaten (*Postgres*)

| <b>Token-basiertes Chunking</b> | <b>Strukturtreues Chunking</b>     |
|---------------------------------|------------------------------------|
| rein technisch                  | semantisch + strukturell           |
| Kontext geht verloren           | Kontext bleibt erhalten            |
| einzelne Textstücke             | zusammenhängende Wissensseinheiten |

## 2. Semantische Vorerschließung – vom Text zur Wissensstruktur

**Analyse jedes Datensegments** (*Chunks*) **mit Qwen 3.5** (*lokal, Open Weights*)

Mehr als reine Vektorisierung: explizite semantische Extraktion

**Extraktion von:**

- Aussagen (*Requirements, Tests, Definitionen*)
- beantwortbaren Fragen (*was kann dieser Abschnitt beantworten?*)
- internen Verweisen (*was können andere Dokumentenstellen beantworten*)

**Ergebnis: strukturierte Wissensseinheiten** (*statt reiner Text -Vektorisierte Token*)

# 3. Hybride Speicherung – Semantik trifft Struktur

(Vector DB + SQL DB)

**Vector DB** (*Embeddings*) → **Semantik**

semantische Ähnlichkeit  
unscharfe Suche  
thematische Nähe  
ähnliche Fragestellungen

**Postgres** (*Relationen*) → **Struktur**

Dokumentstruktur  
Kapitelhierarchie  
Referenzen zwischen Abschnitten  
Positionsdaten & Provenance  
deterministische Filter

# 4. Multi-Step Retrieval – von Treffern zu Evidenz

1. **Anfrage** - semantische Suche über Embeddings
2. **Initiale Treffer** - relevante Dokumentsegmente (Top-k)
3. **Erweiterung (Multi-Hop)** - Verweise zu anderen Stellen, Kapitelzusammenhänge, Tabellenbezüge, Bedingungen / Geltungsbereiche
4. **Aufbau eines Evidenzsets**

**Gezielte Zusammenstellung aller relevanten Informationen** *(nicht nur einzelne Treffer)*

**Evidenz = die Gesamtheit der belegbaren, relevanten Informationsfragmente, auf deren Basis eine Antwort gebildet wird**

# 5. Antwortgenerierung – kontrolliert, nachvollziehbar, lokal

## **Generierung mit Qwen 3.5** *(lokal, Open Weights)*

Ausschließlich auf Basis der aufgebauten Evidenz *(keine freie Generierung)*

### **Inklusive**

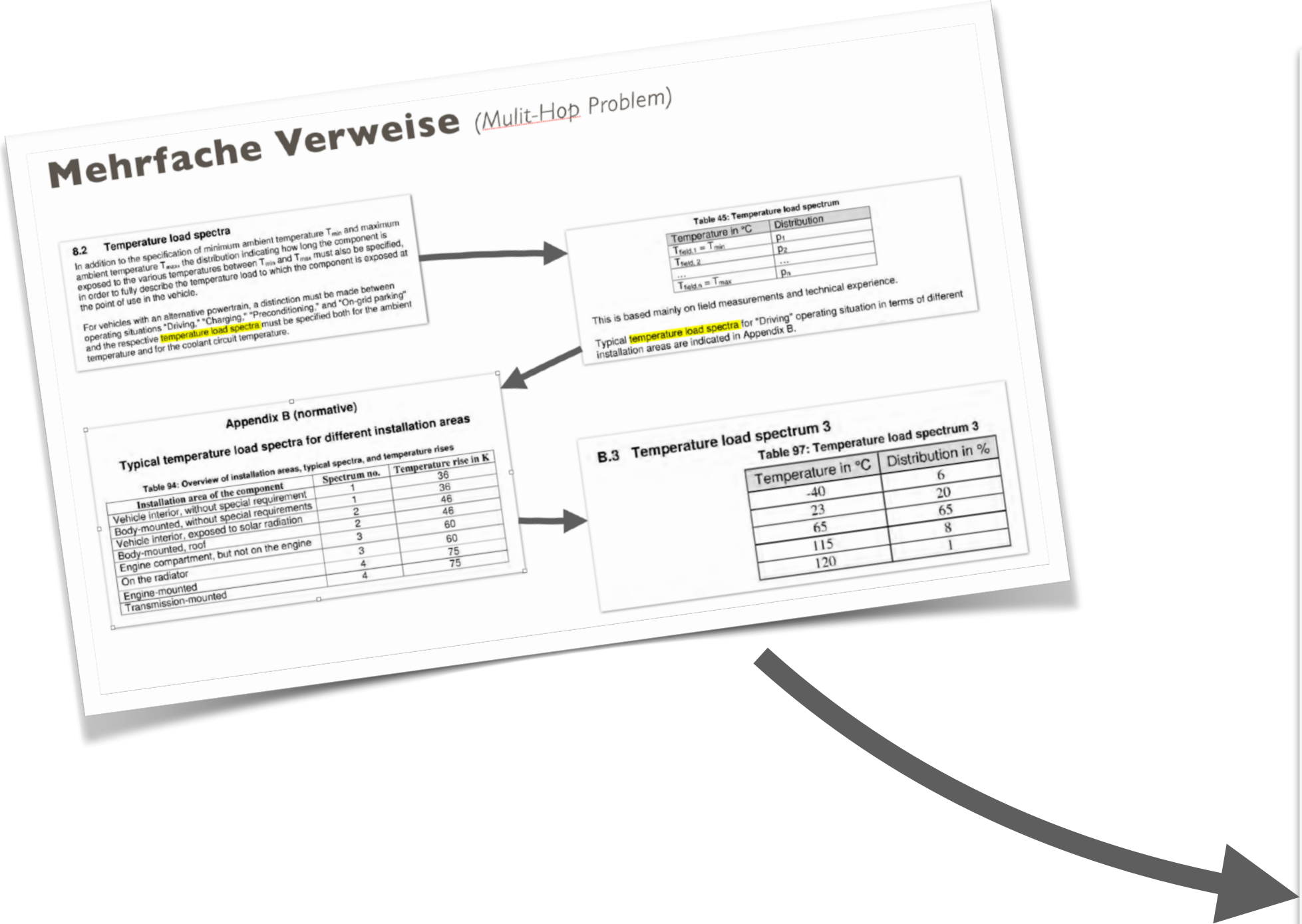
Quellenangaben *(Kapitel, Seite, Tabelle)*

Nachvollziehbare Herleitung der Antwort

### **Kernidee**

Keine Blackbox – jede Antwort ist belegbar

# Multi Hop Ergebnis



## Prompt:

Welche Temperatúra-  
Anforderungen ergeben sich  
an ein Steuergerät, das im  
Motorraum verbaut wird?

## 1. Temperaturklassen und -bereiche (Temperaturlastspektren)

Je nach genauer Einbaulage im Motorraum werden zwei Spektren unterschieden:

| Einbauort                                               | Spektrum   | Temperaturanstieg | Max. Temperatur |
|---------------------------------------------------------|------------|-------------------|-----------------|
| Motorraum, aber <b>nicht am Motor</b> (z. B. am Kühler) | Spektrum 3 | +60 K             | 120 °C          |
| <b>Am Motor montiert</b> (Engine-mounted)               | Spektrum 4 | +75 K             | 140 °C          |

Temperaturverteilung Spektrum 3 (nicht am Motor):

| Temperatur | Anteil |
|------------|--------|
| -40 °C     | 6 %    |
| 23 °C      | 20 %   |
| 65 °C      | 65 %   |
| 115 °C     | 8 %    |
| 120 °C     | 1 %    |

Temperaturverteilung Spektrum 4 (am Motor):

| Temperatur | Anteil |
|------------|--------|
| -40 °C     | 6 %    |
| 23 °C      | 20 %   |
| 85 °C      | 65 %   |
| 135 °C     | 8 %    |
| 140 °C     | 1 %    |

# Reale Antwort: Multi Hop Ergebnis

documents.jaims.ai

Hallo wie kann ich Ihnen helfen?

Welche Temperaturanforderungen ergeben sich für ein Steuergerät das im Motorraum direkt am Motor verbaut wird?

ANSWER

Für ein Steuergerät, das **direkt am Motor im Motorraum verbaut wird**, ergeben sich folgende Temperaturanforderungen gemäß **VW 80000:2017-10**:

**1. Temperaturbelastungsspektrum (Temperaturlastspektrum)**

Der Einbauort "**Engine-mounted**" (am Motor montiert) ist dem **Temperaturlastspektrum 4** (Spectrum no. 4) zugeordnet mit einem **Temperaturanstieg von 75 K**.

**Temperaturbelastungsspektrum 4** definiert folgende Temperaturverteilung:

| Temperatur | Zeitanteil |
|------------|------------|
| -40 °C     | 6%         |
| 23 °C      | 20%        |
| 85 °C      | 65%        |
| 135 °C     | 8%         |
| 140 °C     | 1%         |

**Maximale Betriebstemperatur: 140 °C** (für 1% der Betriebszeit)

# **Wir sind einen grossen Schritt weiter aber noch nicht am Ende**

## **Typische Fehler sind:**

- Bedingungen werden als allgemeine Aussagen extrahiert
- Ausnahmen werden übersehen
- Tabellen werden unvollständig interpretiert
- „shall“, „may“, „must“, „typically“ werden nicht sauber unterschieden
- normative vs. informative Inhalte werden vermischt
- Verweise auf andere Kapitel werden nicht als harte Abhängigkeit erkannt



# FRAGEN ?

**REConf 2026**

Nevres Balic, FORVIA Hella

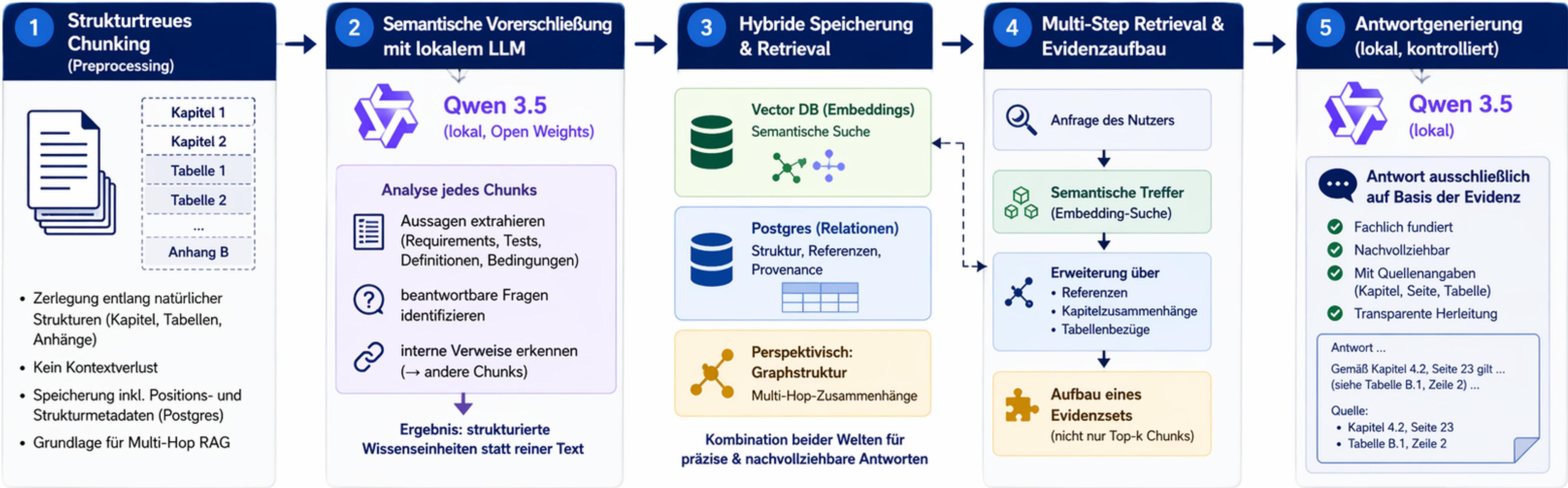
Andreas Willert, SodusWillert

# Lokale KI-Analyse vertraulicher OEM-Dokumente (Mitgeltende Unterlagen)

Hochwertige, nachvollziehbare Antworten bei voller Datenhoheit – mit kleinen, lokal betreibbaren LLMs



100% Datenhoheit  
On-Premise  
Keine Cloud



Technologiestack (lokal)



LLM (Analyse & Generierung):  
Qwen 3.5 (lokal, Open Weights)



Embedding Modelle:  
Qwen 3.5 Embedding (lokal)



Datenbanken:  
Vector DB (z. B. Qdrant / Milvus) + Postgres (SQL)

Warum dieser Ansatz?



Vertrauliche OEM-Dokumente bleiben vollständig im Haus



Hohe Antwortqualität durch strukturorientiertes RAG & Multi-Hop Retrieval



Nachvollziehbarkeit durch Provenance & Quellenangaben



Effizient & ressourcenschonend durch lokale, kompakte LLMs (Qwen 3.5)

Unsere Philosophie

“

Nicht „größeres Modell“, sondern  
bessere Struktur + bessere Evidenz +  
kontrollierte Generierung

”

Maximale Qualität. Minimale Abhängigkeiten.  
Volle Kontrolle.

