Jenseits des Vektors
Warum Ingenieurkunst im Wissensmanagement der wahre Motor für KI in der Energiewirtschaft ist

Computerlinguistik für die Energiewirtschaft

Herausgegeben von der STROMDAO GmbH September 2025

Die digitale Transformation und die zunehmende Komplexität regulativer Vorgaben im deutschen Energiemarkt stellen die Marktkommunikation vor enorme Herausforderungen. Insbesondere die „bilaterale Klärung“ von Ausnahmefällen, die das Gros der manuellen Bearbeitungsprozesse ausmacht, führt zu ineffizienten Abläufen, hohen Fehlerraten und einer erheblichen Belastung der Mitarbeiter. Herkömmliche IT-Systeme und naive KI-Ansätze, die auf der bloßen Vektorisierung von Dokumenten basieren, scheitern an der strukturellen und semantischen Tiefe dieser Problemstellung und verschärfen die Krise im Informationsmanagement, anstatt sie zu lösen. Dieses Whitepaper beleuchtet die Kernursachen dieser Ineffizienzen und zeigt auf, warum ein ingenieurgetriebener Ansatz im Wissensmanagement für die nachhaltige Bewältigung dieser Komplexität unerlässlich ist.

Executive Summary für IT-Führungskräfte

Dieses Whitepaper legt die technische Notwendigkeit dar, den Wert von Künstlicher Intelligenz im deutschen Energiemarkt neu zu bewerten. Der Erfolg von KI resultiert nicht aus der rohen Rechenleistung eines Large Language Models (LLM), sondern aus der präzisen, ingenieurtechnischen Kuratierung seiner Wissensbasis. Der in der IT-Branche weit verbreitete „PDF-zu-Vektor“-Ansatz der Retrieval-Augmented Generation (RAG) ist eine technologisch unzureichende Abstraktion, die an der strukturellen und semantischen Komplexität regulatorischer Dokumente deterministisch scheitert.

Die operative Herausforderung der bilateralen Klärung – das „5%-Problem“, wie es im grundlegenden Whitepaper „Exzellenz Bilaterale Klärung in der Energiewirtschaft“ definiert wird – ist lediglich ein Symptom einer tiefer liegenden Krise im Informationsmanagement. Eine Krise, die durch naive KI-Implementierungen nur noch verschärft wird. Standard-ERP-Systeme, konzipiert als „Systems of Record“ für die Massenverarbeitung, stoßen bei der Bearbeitung wissensintensiver, unstrukturierter Ausnahmefälle an ihre architektonischen Grenzen. Der Versuch, diese Lücke mit einer simplen RAG-Lösung zu schließen, wiederholt denselben fundamentalen Designfehler auf einer neuen technologischen Ebene.

Als Antwort auf diese Herausforderung hat die STROMDAO GmbH, getragen von jahrelanger Expertise in der Marktkommunikation, die Willi-Mako-Architektur entwickelt. Willi-Mako ist kein einfacher Chatbot, sondern ein als „System of Engagement“ konzipiertes Werkzeug, das auf einem sorgfältig konstruierten Wissenskern aufbaut. Dieses System transformiert unstrukturierte regulatorische Texte in einen abfragbaren, relationalen Wissensgraphen – eine Ingenieursleistung, die den entscheidenden Unterschied macht. Dieser Ansatz reduziert nachweislich Fehlerraten, beschleunigt Lösungszeiten und schafft ein sich selbst verstärkendes strategisches Gut in Form von organisationaler Intelligenz.

Dieses Dokument dekonstruiert die technischen Gründe für das Scheitern naiver RAG-Ansätze, erläutert die Prinzipien des „Knowledge Engineering“-Ansatzes der STROMDAO und liefert eine praxisnahe Demonstration anhand des BDEW PRICAT Anwendungshandbuchs. Damit wird die Überlegenheit dieses ingenieurgetriebenen Ansatzes in einem realen, geschäftskritischen Szenario bewiesen und ein neuer Standard für den Einsatz von KI in der Energiewirtschaft definiert.

Das RAG-Dilemma

Die trügerische Eleganz des Standardansatzes in komplexen regulatorischen Umgebungen

1.1 Die Verlockung der einfachen Lösung: RAG als Allheilmittel?

Die Einführung von Retrieval-Augmented Generation (RAG) hat in den IT-Abteilungen eine Welle des Optimismus ausgelöst. Die Technologie verspricht, eine der hartnäckigsten Herausforderungen zu lösen: den Zugriff auf das immense, aber unstrukturierte Wissen, das in Dokumenten, PDFs und internen Wikis schlummert. Der Reiz von RAG liegt in seiner scheinbaren Einfachheit. Anstatt Large Language Models (LLMs) in einem kostspieligen und technisch aufwendigen Prozess des Fine-Tunings auf eigene Daten zu trainieren, bietet RAG einen eleganteren Weg. Es verbindet ein vortrainiertes LLM mit einer externen Wissensdatenbank und ermöglicht dem Modell, zur Laufzeit relevante Informationen abzurufen, um eine Anfrage zu beantworten.

Dieser Ansatz scheint eine direkte Lösung für das im ersten Whitepaper identifizierte Problem zu sein, dass wertvolles Wissen in den Köpfen der Mitarbeiter oder auf unübersichtlichen Netzlaufwerken verteilt ist. Der typische, oft propagierte RAG-Workflow verstärkt diesen Eindruck der Einfachheit:

1. **Ingestion:** Eine Sammlung von Dokumenten (z. B. alle relevanten PDFs des BDEW) wird in das System geladen.

2. **Chunking:** Die Dokumente werden in kleinere, handhabbare Textabschnitte, sogenannte „Chunks“, zerlegt.

3. **Embedding:** Diese Chunks werden mithilfe eines Embedding-Modells in numerische Repräsentationen (Vektoren) umgewandelt, die ihre semantische Bedeutung erfassen.

4. **Speicherung:** Die Vektoren werden in einer spezialisierten Vektordatenbank indiziert.

5. **Retrieval:** Bei einer Benutzeranfrage wird diese ebenfalls in einen Vektor umgewandelt. Das System führt eine Ähnlichkeitssuche (semantische Suche) durch, um die Vektoren – und damit die Text-Chunks – zu finden, die der Anfrage am nächsten kommen.

6. **Augmentation & Generation:** Die gefundenen Chunks werden zusammen mit der ursprünglichen Anfrage als Kontext an das LLM übergeben, das auf dieser Basis eine fundierte Antwort generiert.

Diese Prozesskette ist technologisch verlockend und hat in vielen Anwendungsfällen ihre Berechtigung. In der hochkomplexen und streng regulierten deutschen Energiewirtschaft erweist sich diese Simplizität jedoch nicht als Stärke, sondern als fundamentale Schwäche – ein Versäumnis, das erfahrene IT-Architekten sofort erkennen.

1.2 Der grundlegende Fehler

„Garbage In, Garbage Out“ wird potenziert

Das Prinzip „Garbage In, Garbage Out“ ist ein Grundpfeiler der Informatik, der im KI-Zeitalter oft fahrlässig ignoriert wird. Bei RAG-Systemen erfährt es eine neue, subtilere Dimension. Die Qualität der Ausgabe eines RAG-Systems ist nicht nur von der Richtigkeit der Quelldaten abhängig, sondern in noch größerem Maße von der Qualität des abgerufenen Kontexts. Wenn der Input ein komplexes logisches Regelwerk wie ein BDEW-Anwendungshandbuch ist, ist der „Müll“, der in das System gelangt, nicht der Inhalt selbst – dieser ist hochgradig präzise und wertvoll. Der „Müll“ ist der katastrophale Verlust von Struktur, Kontext und relationalen Beziehungen, der während des naiven Ingestion-Prozesses entsteht.

Hier manifestiert sich ein fundamentales Automatisierungsparadoxon. Das erste Whitepaper beschreibt eindrücklich, wie das Streben nach einer 95%-igen „Dunkelverarbeitung“ in den Massenprozessen der Marktkommunikation (MaKo) die gesamte verbleibende Komplexität auf die restlichen 5% der manuellen Klärfälle konzentriert. Standard-ERPs scheitern an diesem „Unhappy Path“, weil ihre Architektur auf die effiziente Abwicklung von Standards ausgelegt ist, nicht auf die flexible, wissensintensive Bearbeitung von Ausnahmen. Eine naive RAG-Implementierung ist die konsequente Fortsetzung dieses Denkfehlers. Sie versucht, eine Brute-Force-Automatisierung (die unkritische Vektorisierung von allem) auf ein Problem anzuwenden, das im Kern von Nuancen, Kontext und Struktur geprägt ist. So wie das Standard-ERP am Prozess-Ausnahmefall scheitert, scheitert die naive RAG-Architektur am Wissens-Ausnahmefall – dem komplexen, strukturierten Dokument. Dies ist keine rein technische Limitierung; es ist ein strategischer Fehler, der die Versäumnisse der Vergangenheit wiederholt, anstatt sie zu beheben.

1.3 Kritischer Fehlerfall 1

Strukturelle Amnesie und der Albtraum des Chunking

Der Prozess des „Chunking“, also das Zerlegen von Dokumenten in kleinere Teile, ist für RAG-Systeme notwendig, da LLMs nur eine begrenzte Menge an Kontext auf einmal verarbeiten können (das sogenannte „Context Window“). Die gängigste Methode ist das „Fixed-Size Chunking“, bei dem Dokumente in Abschnitte mit einer festen Anzahl von Wörtern oder Tokens zerlegt werden. Diese Methode ist einfach zu implementieren, führt aber bei der Verarbeitung von normativen Dokumenten zu einem deterministischen Informationsverlust.

Regulatorische Texte sind keine fortlaufende Prosa. Ihre Bedeutung ergibt sich aus der präzisen Anordnung von Definitionen, Regeln, Bedingungen und Ausnahmen. Ein willkürlicher Schnitt mitten in einem Satz, einer Liste oder – am schlimmsten – einer Tabelle kann den Sinn vollständig zerstören. Betrachten wir als einfaches Beispiel die Legende auf Seite 5 des PRICAT Anwendungshandbuchs (AHB). Diese Legende erklärt die Bedeutung der verschiedenen Pfeile und farbigen Balken in der darauffolgenden Grafik. Ein Fixed-Size-Chunker könnte die Legende von der Grafik trennen oder sogar die Legende

selbst in zwei Teile zerreißen. Das Ergebnis: Beide Chunks wären für sich genommen bedeutungslos. Das LLM erhielte entweder eine Grafik ohne Erklärung oder eine Erklärung ohne Grafik. In beiden Fällen wäre eine korrekte Interpretation unmöglich.

Dieser „Chunking Nightmare“ ist in der Fachliteratur gut dokumentiert. Zu kleine Chunks führen zu einem Verlust des übergeordneten Kontexts, während zu große Chunks zu viel „Rauschen“ (irrelevante Informationen) in den Kontext einbringen. Beides erhöht das Risiko von Halluzinationen, bei denen das LLM plausible, aber sachlich falsche Informationen erfindet. Für die Energiewirtschaft, in der es auf die exakte Auslegung von Fristen, Formaten und Prozessschritten ankommt, ist dieses Risiko für jede verantwortungsbewusste IT-Abteilung inakzeptabel.

1.4 Kritischer Fehlerfall 2

Die katastrophale Fehlinterpretation von Tabellen

Der schwerwiegendste Fehler naiver RAG-Ansätze liegt in ihrem Umgang mit Tabellen. In Dokumenten wie dem PRICAT AHB sind Tabellen keine bloßen Illustrationen oder Zusammenfassungen – sie *sind* die Spezifikation. Sie definieren die Struktur von EDIFACT-Nachrichten, die Bedingungen für einzelne Datenfelder und die Logik von Geschäftsprozessen. Die gängige Praxis in einfachen RAG-Pipelines, diese Tabellen zu „flatten“, d.h. sie in einen linearen Textstrom umzuwandeln, ist gleichbedeutend mit der Zerstörung ihrer logischen Integrität.

Umfangreiche Forschung belegt, dass dieser Prozess die intrinsischen Zeilen-Spalten-Beziehungen vernichtet und zu einem massiven Informationsverlust führt, der eine präzise Datenabfrage unmöglich macht. Ein LLM kann eine Frage wie „Was ist die Bedingung für das Feld BGM 1001 im Anwendungsfall 27002?“ nicht zuverlässig beantworten, wenn die logische Verknüpfung zwischen der Zeile für

`BGM 1001` und der Spalte `Bedingung` in der Tabelle auf Seite 7 des PRICAT AHB bei der Vektorisierung verloren gegangen ist. Das Modell erhält nur eine lose Abfolge von Wörtern und Zahlen, aus der es die ursprüngliche, rigide Struktur erraten muss – eine Einladung zu Fehlern.

Dies führt zu einer entscheidenden Erkenntnis: Die Tabellen in BDEW-Regelwerken müssen als Code, nicht als Text behandelt werden. Eine Zeile in der Tabelle auf Seite 7 des PRICAT AHB ist nicht einfach nur Text; sie ist eine ausführbare logische Regel. Sie hat eine definierte Struktur mit Komponenten wie

`EDIFACT Struktur`, `Beschreibung`, `Prüfidentifikator` und `Bedingung`. Dies ist analog zu einer Codezeile in einem Computerprogramm. Den Text dieser „Codezeile“ einfach zu vektorisieren, ist so, als würde man einen Screenshot von einem Programm machen und erwarten, dass ein Compiler ihn versteht. Es ignoriert die Syntax, die Beziehungen zwischen den Elementen und die operative Semantik. Jedes System, das den Anspruch hat, diese Dokumente zu „verstehen“, muss Tabellen als strukturierte Daten behandeln, die geparst und modelliert werden müssen, nicht als unstrukturierten Text, der

lediglich eingebettet wird. Diese Neubewertung des Problems verschiebt den Fokus von einer reinen Herausforderung der natürlichen Sprachverarbeitung (NLP) hin zu einer Aufgabe des Data Engineering und der Wissensmodellierung – einer Kerndisziplin der STROMDAO GmbH.

1.5 Kritischer Fehlerfall 3

Semantische Leerräume und domänenspezifische Sprache

Der dritte kritische Fehlerpunkt betrifft die semantische Ebene. Generische Embedding-Modelle, die auf dem allgemeinen Sprachgebrauch des Internets trainiert wurden, haben erhebliche Schwierigkeiten mit der hochspezialisierten Terminologie der Energiewirtschaft. Begriffe wie „Positionsnummer“, „Regelzone“, „Bilanzierungsdatum“, „Prüfidentifikator“ oder „MaLo/MeLo-IDs“ haben im Kontext der Marktkommunikation eine extrem präzise und oft kontraintuitive Bedeutung, die ein allgemeines Modell nicht erfassen kann.

Dies führt zu einem Phänomen, das als „Retrieval Irrelevance“ bekannt ist. Das System findet zwar Dokumente, die die eingegebenen Schlüsselwörter enthalten, verfehlt aber die eigentliche Absicht des Nutzers, weil es die kontextuelle Bedeutung nicht versteht. Ein weiteres Problem ist die „Synonym Blindness“. Das System erkennt möglicherweise nicht, dass ein Sachbearbeiter, der nach „MSB“ fragt, Informationen über den „Messstellenbetrieb“ sucht, oder dass „LF“ und „Lieferant“ austauschbar sind.

Im Ergebnis wird der Nutzer gezwungen, seine Anfrage so zu formulieren, dass sie exakt den Wortlaut im Dokument trifft. Er muss das „magische Schlüsselwort“ erraten, anstatt seine Frage in natürlicher Sprache stellen zu können. Damit geht der Hauptvorteil eines konversationellen KI-Systems verloren, und die Interaktion wird zu einer frustrierenden, ineffizienten Stichwortsuche, die kaum besser ist als die `Strg+F`-Funktion in einem PDF-Reader.

Die Willi-Mako-Doktrin

Die Ingenieurskunst des Wissens im Kern der KI

2.1 Der architektonische Leitstern

Von der Vektordatenbank zum Wissensgraphen

Die Architektur von Willi-Mako ist nicht nur eine Softwarelösung; sie ist die Manifestation einer Ingenieursdoktrin, die bei der STROMDAO GmbH im Zentrum jeder Entwicklung für die Marktkommunikation steht. Der entscheidende konzeptionelle Wandel besteht darin, die zugrundeliegende Qdrant-Collection nicht als einen simplen Speicher für Text-Chunks zu betrachten, sondern als eine vorkompilierte, strukturierte Repräsentation des regulatorischen Wissens. Das Ziel ist nicht, dem LLM Prosa zum Lesen vorzulegen, sondern ihm strukturierte *Fakten* zu liefern, aus denen es eine präzise und verifizierbare Antwort synthetisieren kann.

Diese strategische Ausrichtung harmoniert perfekt mit der Entscheidung, RAG anstelle von Fine-Tuning einzusetzen. Die Wahl zwischen diesen beiden Methoden ist eine der fundamentalsten in der Entwicklung von Enterprise-KI. Fine-Tuning eignet sich hervorragend, um einem Modell eine neue *Fähigkeit* oder einen bestimmten *Stil* beizubringen. RAG hingegen ist die Methode der Wahl, um einem Modell neues, sich änderndes

Wissen zugänglich zu machen. Die Regelwerke des BDEW sind ein dynamischer Wissenskorpus. Ein Fine-Tuning des Modells bei jeder neuen Version des PRICAT AHB wäre unpraktikabel und extrem kostspielig. Mit dem RAG-Ansatz von Willi-Mako muss lediglich die Wissensbasis – die Qdrant-Collection – aktualisiert werden. Dies schafft ein wartbares, skalierbares und zukunftssicheres System – Attribute, die jede IT-Führungskraft fordert.

2.2 Prinzip 1

Intelligente Ingestion und semantische Strukturierung

Anstatt dem naiven Paradigma der blinden Dokumentenzerlegung zu folgen, basiert der Ingestion-Prozess von Willi-Mako auf einem fundamentalen Prinzip des Software-Engineerings: Verstehe deine Datenstruktur, bevor du sie verarbeitest.

Semantisches Chunking: Willi-Mako lehnt das willkürliche Fixed-Size-Chunking ab und setzt stattdessen auf eine Methode, die die inhärente Struktur des Dokuments respektiert. Die Segmentierung erfolgt entlang natürlicher semantischer Grenzen: Absätze, Überschriften, Listenelemente und – von entscheidender Bedeutung – einzelne Tabellenzeilen. Jeder so erzeugte Chunk repräsentiert eine in sich geschlossene logische Einheit. Dies stellt sicher, dass der Kontext innerhalb einer Informationseinheit vollständig erhalten bleibt.

Tabellen-Dekonstruktion und -Rekonstruktion: Hier liegt das Herzstück des Systems und die Kernkompetenz der STROMDAO-Ingenieure. Tabellen aus Dokumenten wie dem PRICAT AHB werden nicht in Text umgewandelt. Stattdessen werden sie mithilfe fortschrittlicher Parsing-Techniken, analog zu den in der Forschung beschriebenen Methoden , dekonstruiert. Jede einzelne Zeile der Tabelle wird als ein strukturiertes Datenobjekt behandelt, das die Spalten-Wert-Beziehungen explizit bewahrt.

Beispielsweise wird die Zeile für `BGM 1001` aus der Tabelle auf Seite 7 des PRICAT AHB nicht als der String „BGM 1001 Ausgleichsenergiepreis Z04 X (A)“ gespeichert. Stattdessen wird sie in ein JSON-ähnliches Objekt transformiert, das die Logik der Tabelle abbildet:

JSON

```
None
{
  "segment_group": "BGM",
  "data_element": "1001",
  "description": "Ausgleichsenergiepreis",
  "code": "Z04",
  "process_id_ref": "27001",
  "condition_ref": ",,",
  "source_page": 7
}
```

Durch diese Transformation wird die implizite Logik der Tabelle in eine explizite, maschinenlesbare Form überführt. Die Information ist nicht länger nur Text, sondern ein abfragbarer Datensatz. Dies macht die Logik des Regelwerks für das System rechentechnisch zugänglich und ist die Voraussetzung für die präzisen Antworten, die einfache RAG-Systeme nicht liefern können.

2.3 Prinzip 2

Facettenreiche Metadaten als kontextueller Klebstoff

Jeder einzelne Chunk in der Qdrant-Collection von Willi-Mako wird mit einer reichhaltigen Schicht an Metadaten angereichert. Diese Metadaten sind der Schlüssel zur Überwindung

von Retrieval-Irrelevanz und Rauschen, da sie eine präzise Filterung des Suchraums ermöglichen, bevor die semantische Suche überhaupt beginnt.

Für einen Chunk aus dem PRICAT AHB könnten diese Metadaten beispielsweise umfassen:

- `document_source: 'PRICAT_AHB_2_0e'`
- `publication_date: '2024-06-19'`
- `section_id: '5.1'`
- `section_title: 'Preisblätter Ausgleichsenergiepreis und MSB-Leistungen'`
- `process_id: '27002'`
- `data_type: 'table_row'`
- `relevant_role:`
- `valid_from: '2024-01-01'`

Diese Metadaten agieren als kontextueller Klebstoff, der die einzelnen Wissensbausteine miteinander verbindet. Wenn ein Benutzer eine Frage zum MSB-Preisblatt stellt, kann das System eine Vorfilterung durchführen, die den gesamten Vektorraum auf Chunks beschränkt, bei denen beispielsweise `process_id == '27002'` ist. Dies reduziert das Rauschen dramatisch und stellt sicher, dass nur hochrelevante Informationen in den Kontext des LLM gelangen. Es ist der Unterschied zwischen der Suche nach einer Nadel in einem Heuhaufen und der Suche nach einer Nadel in einer Schachtel mit Nadeln.

2.4 Prinzip 3
Hybride Suche für Präzision und Vollständigkeit

Die Suche in regulatorischen Texten erfordert zwei unterschiedliche Herangehensweisen: das exakte Finden von Codes und Identifikatoren und das inhaltliche Verstehen natürlichsprachlicher Fragen. Eine rein semantische Suche ist für Ersteres ungeeignet, eine reine Schlüsselwortsuche für Letzteres.

Eine semantische Suche versteht, dass „Kosten für den Zählerwechsel“ und „Preise für den Austausch von Messeinrichtungen“ konzeptionell ähnlich sind. Eine Schlüsselwortsuche hingegen ist darauf ausgelegt, exakte Zeichenketten wie „BGM+Z32“ oder den Prüfidentifikator „27002“ zu finden.

Willi-Mako implementiert daher einen hybriden Suchansatz, der die Stärken beider Methoden kombiniert. Das System analysiert die Benutzeranfrage und entscheidet intelligent, wann eine exakte Übereinstimmung erforderlich ist und wann eine semantische Ähnlichkeitssuche zielführender ist. Oft werden beide Methoden parallel angewendet und ihre Ergebnisse gewichtet, um eine optimale Balance zwischen Präzision und Vollständigkeit zu erreichen. Dieser ingenieurtechnische Ansatz stellt sicher, dass sowohl der Fachexperte, der nach einem bestimmten Code sucht, als auch der neue Mitarbeiter, der eine allgemeine Frage stellt, die bestmögliche Antwort erhalten.

Tabelle 1
Eine vergleichende Analyse der Wissensmanagement-Architekturen

Die folgende Tabelle fasst die fundamentalen Unterschiede zwischen dem naiven RAG-Ansatz und der von STROMDAO entwickelten, kuratierten Wissensarchitektur zusammen.

Feature-Dimension	Naiver RAG (Der „PDF-zu-Vektor“-Ansatz)	Der Engineering-Ansatz der STROMDAO (Willi-Mako)
Daten-Ingestion	Einfache Textextraktion aus PDFs.	Intelligentes Parsen der Dokumentstruktur (Überschriften, Listen, Tabellen).
Chunking-Strategie	Willkürliche, auf fester Größe basierende Token-Splits.	Semantisches Chunking basierend auf natürlichen Dokumentgrenzen (Absätze, Tabellenzeilen).
Umgang mit Tabellen	Tabellen werden zu unstrukturiertem Text verflacht, was die relationale Integrität zerstört.	Tabellen werden in strukturierte Objekte (z. B. JSON) dekonstruiert, wodurch Zeilen-Spalten-Beziehungen für präzise Abfragen erhalten bleiben.

Kontextuelle Integrität	Hohes Risiko der „Kontextfragmentierung“, bei der zusammengehörige Informationen über Chunk-Grenzen hinweg getrennt werden.	Der Kontext wird durch semantisches Chunking bewahrt und durch umfangreiche, abfragbare Metadaten angereichert.
Retrieval-Mechanismus	Verlässt sich typischerweise ausschließlich auf die semantische (Vektor-)Suche.	Hybride Suche, die semantische Suche für die Absicht mit Schlüsselwortsuche für die Präzision bei Codes und Identifikatoren kombiniert.
Abfragegenauigkeit	Anfällig für Irrelevanz, Rauschen und Halluzinationen aufgrund von mehrdeutigem Kontext.	Hohe Präzision durch Metadaten-Vorfilterung und hybride Suche, die dem LLM sauberen, deterministischen Kontext liefert.
Verifizierbarkeit/ Nachvollziehbarkeit	Antworten sind schwer auf eine spezifische Quellpassage zurückzuführen, insbesondere bei großen, verrauschten Chunks.	Antworten werden aus diskreten Fakten synthetisiert, die jeweils exakt auf ihren Quell-Chunk und ihr Dokument zurückführbar sind.

Wartungsaufwand	Geringe anfängliche Einrichtungskosten, aber hohe versteckte Kosten durch schlechte Leistung und ständiges Prompt-Engineering.	Höhere Anfangsinvestition in Data Engineering, führt aber zu einem robusten, zuverlässigen und leicht aktualisierbaren Wissenskern.

Ein praktischer Einblick
Vom PRICAT-Labyrinth zur handlungsfähigen Klarheit

3.1 Das Szenario
Eine komplexe Preisanfrage

Um die theoretischen Argumente greifbar zu machen, wird ein realistisches, geschäftskritisches Szenario durchgespielt, das jeder IT-Leiter aus dem Support-Alltag kennt. Wir versetzen uns in die Lage eines erfahrenen Sachbearbeiters in der Abteilung für Marktkommunikation eines Stadtwerks.

Die Aufgabe: „Unser Abrechnungssystem hat eine PRICAT-Nachricht von einem neuen Messstellenbetreiber (MSB) für ein intelligentes Messsystem (iMSys) zurückgewiesen. Die Installation fand am 15. Januar 2024 statt, und das Preisblatt soll ab diesem Datum gelten. Die übermittelte Nachricht enthält eine Artikel-ID im Format `n13-n2`. Der vom System generierte Fehler lautet ‚Ungültiges Artikel-ID-Format für das angegebene Gültigkeitsdatum'. Ich muss nun überprüfen, ob diese Ablehnung gemäß dem neuesten BDEW PRICAT Anwendungshandbuch korrekt ist, und die exakte Regel finden, die unsere Entscheidung rechtfertigt, um dies dem Marktpartner klar kommunizieren zu können."

Dieses Szenario ist exemplarisch für die täglichen Herausforderungen in der bilateralen Klärung: Es erfordert ein tiefes Verständnis für komplexe, voneinander abhängige Regeln, die über ein dichtes technisches Dokument verstreut sind.

3.2 Teil A
Der manuelle Kampf – Die Navigation durch das PRICAT-PDF

Der Sachbearbeiter öffnet die PDF-Datei des „PRICAT Anwendungshandbuch 2.0e" und beginnt die typische, fehleranfällige Vorgehensweise, die in vielen Abteilungen noch Standard ist.

Schritt 1: Den richtigen Abschnitt finden. Der Sachbearbeiter weiß, dass es um die Preisgestaltung des MSB geht. Ein Blick in das Inhaltsverzeichnis auf Seite 3 führt ihn schnell zu Kapitel 5.1, „Preisblätter Ausgleichsenergiepreis und MSB-Leistungen".

Schritt 2: Den relevanten Prozess identifizieren. Auf Seite 7 beginnt die zentrale Tabelle, die die EDIFACT-Struktur beschreibt. Er sieht zwei Spalten für unterschiedliche Prozesse: 27001 (Ausgleichsenergiepreise) und 27002 (Messstellenbetrieb). Die Beschreibung für 27002, „Preisblatt Messstellenbetrieb / Konfigurationen MSB an LF/NB", bestätigt ihm, dass dies der korrekte Kontext für seine Anfrage ist.

Schritt 3: Das Feld für die Artikel-ID lokalisieren. Nun muss er die spezifische Regel für die Artikel-ID finden. Er scannt die linke Spalte der Tabelle nach dem Begriff „Artikel". Auf

Seite 10, innerhalb der Segmentgruppe SG36, wird er fündig: Das Feld `LIN 7140` wird als „Produkt-/Leistungsnummer“ beschrieben.

Schritt 4: Die Bedingungen entschlüsseln. Dies ist der kognitiv anspruchsvollste Teil. Die Spalte „Bedingung“ für dieses Feld enthält einen dichten Block von Verweisen und Logik. Der Sachbearbeiter sieht Hinweise auf verschiedene Formate: `Format: Artikelnummer`, `Format: n1-n2-n1-n3` und `Format: n13-n2`. Das Format in seiner fehlerhaften Nachricht, `n13-n2`, entspricht der Regel ``. Die zugehörige Bedingung lautet `( A )`.

Schritt 5: Die entscheidende Verbindung herstellen. Der Sachbearbeiter muss nun deduktiv vorgehen. Er untersucht die Bedingungen für die *anderen* Formate. Für das Format `n1-n2-n1-n3` (Regel `) lautet die Bedingung `(A A)`. Er sucht nun nach der Definition der Bedingung` in der äußerst rechten Spalte der Tabelle. Dort findet er den entscheidenden Satz: „wenn der Zeitpunkt im DTM+157 DE2380 ≥ 01.01.2024 00:00 Uhr gesetzlicher deutscher Zeit“. Sein Gültigkeitsdatum ist der 15. Januar 2024, was eindeutig *nach* diesem Stichtag liegt.

Schritt 6: Der „Aha-Moment“ und die Gegenprobe. Um seine Hypothese zu bestätigen, sucht er nach einer Regel, die das `n13`-Format explizit an ein Datum *vor* dem Stichtag bindet. Er geht zurück zu Seite 8, zur Zeile für das `DTM 2380` Segment (Gültigkeitsbeginn). Dort findet er die Bedingung ``: „Wenn BGM DE1001 = Z32 und LIN DE7140 im Format n13, dann muss der hier genannte Zeitpunkt < 01.01.2024 00:00 Uhr gesetzlicher deutscher Zeit sein“. Dies bestätigt unmissverständlich, dass Formate, die mit `n13` beginnen, für Zeiträume *vor* dem 1. Januar 2024 vorgesehen waren. Die Systemablehnung war korrekt.

Zusammenfassung des manuellen Prozesses: Dieser Prozess hat selbst für einen Experten zwischen fünf und zehn Minuten gedauert. Er erforderte das Springen zwischen den Seiten 7, 8 und 10, das Halten mehrerer voneinander abhängiger Bedingungen im Arbeitsgedächtnis und ein hohes Maß an Abstraktionsvermögen. Das Risiko, eine entscheidende Bedingung zu übersehen oder falsch zu interpretieren, ist erheblich.

3.3 Teil B

Die Willi-Mako-Erfahrung – Von der Anfrage zur verifizierbaren Antwort

Der Sachbearbeiter wendet sich nun an die Chat-Schnittstelle von Willi-Mako.

Benutzeranfrage: „Ein MSB hat uns ein PRICAT Preisblatt für ein iMSys mit Gültigkeit ab 15.01.2024 geschickt. Die Artikel-ID hat das Format n13-n2. Ist das korrekt?“

Die deterministische Prozesskette in Willi-Mako:

1. **Anfrage-Dekonstruktion:** Die NLP-Schicht des Systems zerlegt die Anfrage und identifiziert die Schlüsselentitäten: `PRICAT Preisblatt`, `MSB`, `iMSys`, `Gültigkeit ab 15.01.2024`, `Artikel-ID Format n13-n2`

2. **Metadaten-Filterung:** Das System initiiert eine Suche in der Qdrant-Collection, wendet aber zunächst einen strikten Vorfilter an. Basierend auf den Entitäten `PRICAT Preisblatt` und `MSB` filtert es den Suchraum auf alle Chunks, die die Metadaten `document_source: 'PRICAT_AHB_2_0e'` UND `process_id: '27002'` enthalten.

3. **Hybride Suche:**
   - Das System führt eine Schlüsselwortsuche nach den Termen `Artikel-ID`, `LIN`, `7140` und den Format-Strings durch.
   - Gleichzeitig führt es eine semantische Vektorsuche nach dem Konzept „Regeln für Artikel-ID-Formate in Abhängigkeit vom Datum“ durch.

4. **Kontext-Zusammenstellung:** Der Suchprozess liefert eine kleine Anzahl hochrelevanter, strukturierter Fakten-Chunks. Darunter befinden sich die entscheidenden, als strukturierte Objekte gespeicherten Tabellenzeilen:
   - Der Fakt, der das Format `n13` mit der Bedingung `` und dem Datum `< 01.01.2024` verknüpft.
   - Der Fakt, der das Format `n1-n2-n1-n3` mit der Bedingung `` und dem Datum `≥ 01.01.2024` verknüpft.

5. **Synthese und Generierung:** Das LLM erhält diesen sauberen, unzweideutigen und bereits vorstrukturierten Kontext. Es muss kein komplexes Dokument interpretieren; es erhält eine Reihe klarer logischer Aussagen. Auf dieser Grundlage synthetisiert es eine kohärente und präzise Antwort.

Die Antwort von Willi-Mako (Generierte Ausgabe):

> Nein, das ist nicht korrekt.
>
> Gemäß dem PRICAT Anwendungshandbuch (Version 2.0e) gilt für Preisblätter des Messstellenbetriebs (Prüfidentifikator 27002) mit einem Gültigkeitsbeginn (DTM+157) am oder nach dem 01.01.2024 eine geänderte Regelung für die Produkt-/Leistungsnummer (Artikel-ID).
>
> - **Für Zeitpunkte ≥ 01.01.2024** ist das Format `n1-n2-n1-n3` zu verwenden (Bedingung).
>
> - **Für Zeitpunkte < 01.01.2024** war das Format `n13` (und verwandte Formate) gültig (Bedingung).
>
> Da der Gültigkeitsbeginn der 15.01.2024 ist, hätte die Artikel-ID im Format `n1-n2-n1-n3` übermittelt werden müssen. Die Ablehnung ist somit korrekt.
>
> **Quellen:**

- *PRICAT AHB 2.0e, Seite 10, SG36 LIN, DE7140, Bedingung *
- *PRICAT AHB 2.0e, Seite 8, DTM 2380, Bedingung *

Zusammenfassung des KI-gestützten Prozesses: Die Vorteile für eine IT-Abteilung sind offensichtlich: Die Antwort wird in Sekunden statt in Minuten geliefert. Die Genauigkeit ist extrem hoch, da sie auf der maschinellen Verarbeitung strukturierter Fakten beruht. Am wichtigsten ist die Verifizierbarkeit: Das System liefert nicht nur eine Antwort, sondern auch die exakten Quellenangaben, was für die Compliance und die Kommunikation mit Marktpartnern unerlässlich ist.

Die strategische Dividende

Konservierung von Domänenwissen als strategisches IT-Asset

4.1 Eindämmung der „Experten-Erosion“ und Schaffung einer lebendigen Wissensbasis

Der manuelle Prozess zur Klärung der PRICAT-Anfrage ist vollständig vom impliziten Wissen eines einzelnen Mitarbeiters abhängig. Dieses Wissen ist ein fragiles Gut, ausgesetzt dem Risiko der „Experten-Erosion“, wenn erfahrene Mitarbeiter die Organisation verlassen.

Willi-Mako wirkt diesem Risiko direkt entgegen, indem es dieses implizite Expertenwissen externalisiert und als strategisches IT-Asset kodifiziert. Die Fähigkeit, die komplexen Abhängigkeiten im PRICAT AHB zu navigieren, ist nicht länger an eine Person gebunden, sondern wird zu einer reproduzierbaren, systemimmanenten Fähigkeit. Die kuratierte Wissensbasis fungiert als dauerhaftes, zugängliches institutionelles Gedächtnis.

Mehr noch, dieses Gedächtnis ist eine „lebendige“ Ressource. Wenn der BDEW die Version 2.0f des PRICAT AHB veröffentlicht, muss nicht das gesamte System neu trainiert werden. Es müssen lediglich die betroffenen strukturierten Fakten in der Qdrant-Collection aktualisiert werden. Dieser gezielte Eingriff bringt das gesamte Wissen des Systems augenblicklich auf den neuesten Stand – ein entscheidender Vorteil gegenüber dem starren Wissen eines feinabgestimmten Modells.

4.2 Revolutionierung von Einarbeitung und Training

Das ursprüngliche Whitepaper stellt die Behauptung auf, dass Willi-Mako die Einarbeitungs- und Schulungszeit für Sachbearbeiter um bis zu 50% reduzieren kann. Das in Abschnitt 3 durchgespielte Szenario liefert den konkreten Beweis dafür.

Ein neuer Mitarbeiter wäre im manuellen Prozess völlig überfordert. Mit Willi-Mako kann derselbe Mitarbeiter die Frage in natürlicher Sprache stellen und erhält in Sekundenschnelle eine korrekte, begründete und mit Quellen belegte Antwort.

Dies transformiert den Lernprozess fundamental. Anstatt Monate damit zu verbringen, dichte regulatorische Dokumente auswendig zu lernen, können neue Mitarbeiter fast sofort produktiv und regelkonform arbeiten. Für die IT-Abteilung bedeutet dies eine schnellere Time-to-Productivity neuer Mitarbeiter und eine geringere Belastung für den First- und Second-Level-Support. Die Schulung verlagert sich von einem passiven, einmaligen Ereignis zu einem aktiven, kontinuierlichen „Training-on-the-Job“.

4.3 Aufbau operativer Resilienz für die Komplexität von morgen

Das Whitepaper „Exzellenz Bilaterale Klärung“ blickt zurecht auf die zukünftigen Herausforderungen des Energiemarktes: Redispatch 2.0, der flächendeckende Rollout intelligenter Messsysteme (iMSys) und die Umsetzung von §14a EnWG. Jede dieser Initiativen wird eine Flut neuer Datenformate und komplexerer Prozessketten mit sich bringen. Die Komplexität wird exponentiell zunehmen.

In diesem Umfeld offenbart der von STROMDAO entwickelte, kuratierte Wissensansatz seinen größten strategischen Vorteil: skalierbare Intelligenz. Der manuelle Ansatz des Wissensmanagements skaliert bestenfalls linear. Dieses Modell ist nicht nachhaltig und wird unweigerlich zu einem Engpass führen.

Der Ansatz von Willi-Mako hingegen bietet eine skalierbare, zukunftssichere Plattform. Der einmalige Aufwand, eine neue Verordnung intelligent zu parsen, zu strukturieren und in den Wissenskern zu integrieren, macht dieses Wissen sofort und für jeden einzelnen Nutzer in der gesamten Organisation verfügbar. Der Return on Investment der anfänglichen Data-Engineering-Anstrengung wächst mit jeder neuen Verordnung und jedem neuen Nutzer. Dies stellt einen fundamentalen Wandel in der Ökonomie von Compliance und Wissensmanagement dar. Das Unternehmen baut eine Fähigkeit auf, die es ihm ermöglicht, zukünftige Komplexität nicht nur zu bewältigen, sondern sie effizient zu absorbieren und in operatives Handeln umzusetzen.

Über die STROMDAO GmbH: Ihr Partner für die digitale Energiewende

Die STROMDAO GmbH wurde 2017 mit einer klaren Vision gegründet: die Komplexität der Energiewende durch intelligente, datengetriebene Lösungen zu meistern. Unsere Arbeit wird von den Kernwerten **Innovation, Transformation und Nachhaltigkeit** angetrieben.

Seit unserer Gründung haben wir uns als führender Experte für die digitale Infrastruktur der Energiewirtschaft etabliert. Mit Produkten wie dem **GrünstromIndex** (seit 2019) und dem KI-Assistenten **Willi Mako** (seit 2024) übersetzen wir tiefgreifende Markt- und Prozessexpertise in praxistaugliche, wertschöpfende Werkzeuge.

Wir verstehen uns nicht nur als Softwareanbieter, sondern als strategischer Partner für regionale Energieversorger und Stadtwerke. Unsere Kunden profitieren von der kollektiven Intelligenz aus über 10.000 gelösten Fällen und der Expertise von mehr als 500 Branchenexperten, die zu unserer Wissensbasis beitragen. Unsere Mission ist es, die "Blackbox MaKo" zu öffnen und sie von einem unkalkulierbaren Risiko in einen steuerbaren Wettbewerbsvorteil zu verwandeln.

Schlussfolgerung

Der neue Imperativ für Unternehmens-KI

Die Analyse zeigt unmissverständlich: Die Ära, in der Künstliche Intelligenz als eine Blackbox behandelt wurde, die man mit rohen Dokumenten füttert, ist für ernsthafte Unternehmensanwendungen vorbei. Insbesondere in hochregulierten Branchen wie der Energiewirtschaft liegt die Zukunft effektiver, zuverlässiger und vertrauenswürdiger KI in einer präzise instrumentierten Schnittstelle zwischen Fachexpertise und KI-gestützter Prozessautomatisierung.

Der „Expert-in-the-Loop“-Ansatz, der durch die kuratierte Wissensarchitektur von Willi-Mako verkörpert wird, setzt hierfür den neuen Maßstab. Er erfordert eine strategische Vorabinvestition in Data Engineering und Wissensmodellierung, anstatt sich auf die trügerische Einfachheit des „PDF-zu-Vektor“-Ansatzes zu verlassen. Der Ertrag dieser Investition ist jedoch ein System, das nicht nur Fragen beantwortet, sondern die kollektive Intelligenz der Organisation systematisch erfasst, bewahrt und verstärkt. Es transformiert passives, in Dokumenten gefangenes Wissen in aktives, abrufbares und handlungsleitendes Wissen.

Im Energiemarkt von morgen, in dem Komplexität die einzige Konstante ist, ist die Fähigkeit, regulatorisches Wissen mit Geschwindigkeit und Präzision zu navigieren, mehr als nur eine operative Effizienzsteigerung – sie ist ein entscheidender Wettbewerbsvorteil. Die STROMDAO GmbH zeigt mit Willi-Mako den Weg auf, wie die Bewältigung regulatorischer Komplexität von einer kostspieligen Belastung zu einem strategischen, von der IT gemanagten Gut umgewandelt werden kann.

Kleingedrucktes

STROMDAO GmbH
E-Mail: kontakt@stromdao.com
Telefon: +49 6226 9680090
Web: https://stromdao.de

Amtsgericht Mannheim - HR-B 728691 - Umsatzsteuer-ID: DE311820716