Auswertung Archivdokumente zu Kunstobjekten

Anwendung von Large-Language-Models (LLMs) zur Strukturierung und Integration heterogener Daten in relationale Datenbanksysteme

Henry Rotzoll (DVZ MV GmbH)

Gliederung

1. Einführung und Fragestellung, Hintergrund
2. Datengrundlage und Umsetzung
3. Ergebnisse
4. Erkenntnisse und Empfehlungen
5. Fazit und Ausblick

Einführung und Fragestellung

Die Digitalisierung stellt Unternehmen und Organisationen vor große Herausforderungen

- sie verfügen über Archive mit großen Mengen an Dokumenten, die in eine gemeinsame Datenbank überführt werden sollen
- Dokumente wie Word- und PDF-Dateien enthalten unstrukturierte Daten, die vor der Datenintegration in ein strukturiertes Format überführt werden müssen
- Zusammenführung heterogener Daten ist ein zeitintensiver und fehleranfälliger Prozess, der mit hohem manuellen Aufwand verbunden ist

Kann der Prozess der Datenstrukturierung und -integration mithilfe von generativer KI automatisiert werden?

Hintergrund

- Masterarbeit von Henrik Bongertmann (Universität Rostock)
- In Zusammenarbeit mit dem Sachgebiet GEO in der Zeit von April bis Oktober 2024
- Bestehendes Projekt mit der Nordkirche: Kunstguterfassung der Evangelisch-Lutherischen Kirche in Norddeutschland
  - Ziel: erfasste Kunstgüter in eine relationale Datenbank zu integrieren
  - Zweck: die in den Kirchengemeinden vorhandenen Kunstobjekte über eine interaktive Karte zugänglich zu machen
- 4 verschiedene Open Source Sprachmodelle verglichen
  - 30 Testdokumente verwendet
  - Gewinner: Llama-3.1-8B
- Laptop mit Intel Core i7, 32 GB Arbeitsspeicher und Nvidia GeForce GTX 1070 mit 8 GB VRAM für die Prozessierung verwendet

Datengrundlage

- 6509 Dokumente in den Dateiformaten PDF, DOCX und DOC
  - Dokumente enthalten Objektbeschreibungen aus der Kunstguterfassung

Propstei Stralsund
Kunstguterfassung Ahrenshagen

Ahrenshagen

Kanzelaltar Mitte 18. Jh. (Reiner Nr. 1)
gestiftet vermutlich von Hans Albrecht von Plüschow (Plüskow) d. J.

Datierung: um 1745/50
Material: Holz, farbig gefasst und teilvergoldet,
Fassung: Fassung im 20. Jh. (1960er oder 70er Jahre?) erneuert, Architektur in verschiedenen hellen Grautönen mit vergoldeten Leisten, die Säulenschäfte blau mit hellen Marmorierung und vergoldeten Kapitellen, Ornamentik teilvergoldet, Wolken auf Versilberung blau gefasst, Figuren monochrom weiß, darunter Reste einer älteren Fassung[1]
Maße: H 6,75 m (lt. Reiner), Tugendfigur links 125 cm, rechts 127 cm
Inschriften: in Inschriftkartusche am Schalldeckel „Heilige nicht/ in deiner / Wahrheit", Putto links: „Du sollst Gott deinen Herrn lieben von ganzem Herzen", Putto rechts: „Du sollst deinen nächsten lieben als dich selbst"
Wappen: Predellenzone links von Plüschow (Plüskow): das Wappen ist durch die Neufassung des Retabels egalisierend überstrichen, eigentlich durch eine linke silberne Querspitze von Rot und Gold geteilt, auf dem Helm mit rechts rot-silbernen und links rot-goldenen Helmdecken ist ein brauner Stab, um den sich eine silberne Schlange zwischen zwei goldenen Hirschstangen windet.[2], rechts von Gneden
Beschreibung/Bemerkungen: dreigeschossiger architektonischer Aufbau, der mit dem Gottesauge, Rocaille- und Blüten-Ornament geschmückte Kanzelkorb im Mittelfeld wird von jeweils zwei Säulen zu den Seiten gerahmt, welche ein stark profiliertes verkröpftes Gebälk tragen, zwischen den Säulen steht jeweils eine weibliche Tugendallegorie als Freiplastik. Die Attribute beider sind verloren. **Vermutlich handelt es sich um Fides und Spes (Glaube und Hoffnung).** Das Obergeschoß wird von einer Wolkenglorie mit dem Auge Gottes mit dem hebräischen Gottesnamen bekrönt, welches von zwei auf dem Gebälk sitzenden Engeln angebetet wird. Außen neben den Engeln stehen reich mit Rocaille-Ornament verzierte Vasen auf dem Gebälk. Auf dem Schalldeckel sitzen zwei Putten mit den Gebotstafeln, in der Predella ein Abendmahlsgemälde. Christus und die Jünger sitzen an einer ungewöhnlich langen, etwas schräg stehenden Tafel, einzig Judas, mit dem Geldbeutel in der Hand, steht links vor dem Tisch, die anderen Apostel sind zu Zweier- bzw. Dreiergruppen verteilt, Johannes sitzt auf dem Schoß Christi, Christus reicht Judas den Bissen, an den Seiten sind schwer drapierte

[1] ANELLING 2007, S. 86f. stellte die „barocke Farbenpracht" des Kanzelaltars als selten heraus.
[2] Siehe https://de.wikipedia.org/wiki/Pl%C3%BCschow_(Adelsgeschlecht) (Abruf 02.10.2017)

Kirchenkreis Dithmarschen, Hemme, St. Marien

Beichtstuhl/Pastorenstuhl
Standort: Kirche
Maße: H ca. 229 cm, B ca. 250 cm, T ca.137,5 cm
Szenen der Wandfüllung, Rückwand: H ca. 98,2 cm; B ca. 45,8 - 46 cm
Tugenden (Front): H ca. 80 - 86,5 cm; B ca. 46,7 - 48 cm
Tugenden (Tür): H ca. 85,5 cm; B ca. 50,5 cm
Material: Holz, Eisen, Farbfassung
Datierung: 1622/1639 (angeben der Kunsttopographie), 1709 (Außenseite Zugang)
Inschriften
Überschriften der Tugenden: Der Glaube / die Liebe / die Hoffnung / die geduld / die demuth / die Vorsichtigkeith
Subscriptio (Innenraum): GENES. XLI. / GENES. XXXXIX. / GENES. XXXVII. / GENES. XXXII.
Inschriften: Der Glaube / die Liebe / die Hoffnung / die geduld / die demuth / die Vorsichtigkeith
Signatur: —
Werkstatt/Künstler: unbekannt
Beschreibung/Bemerkungen:
Das einfache Kastengestühl mit vergittertem Sichtschutz wurde auf der Evangelienseite im Chorbogen im Chorbereich aufgestellt. Ein architektonischer Aufbau mit kassettierten Feldern und Abschlussgesims bestimmt die optische Erscheinung. Das Gesims ist mit Zahnschnitt und Diamantbossen rhythmisiert. In die Brüstungsfelder sind farbige schlichte Personifikationen der Tugenden auf das Trägermaterial gemalt worden. Die christlichen Tugenden wurden mit zwei der Kardinaltugenden erweitert: Glaube, Liebe, Hoffnung, Geduld, Demut und im Innenraum die Vorsicht.
Der obere Abschnitt des stammlosen Stuhls ist zur Gemeinde, an der Längsseite wie zum Altar vergittert. Die vertikal und horizontal angeführten Stäbe sind gefast. Die vergitterte Längsseite ist den Brüstungsfeldern entsprechend vergeteilt, wobei die beiden mittleren über eine Scharnierführung zu öffnen sind; ähnlich verhält es sich mit den beiden zur Gemeinde ausgerichteten Knaggartenseiten. Die zum Altar ausgerichtete Tür ist über einen Hebel-Bügel-Verschluss zu betätigen. Ursprünglich war ein Schloss angebracht.
Die Rückwand des Innenraums bindet vier Felder der Wandfüllung mit ein, und thematisiert den Kampf Jacobs mit dem Engel - Josef wird von seinen Brüdern in den Brunnen gestossen - Josef soll

Kunst- und Kulturgut im Bereich des Kirchenkreises Mecklenburg

Ort / Örtliche Kirche
Goldberg

Ortskennzahl
407

Kirchengemeinde

Bezeichnung
ausgebautes Grabplattenfragment

Anzahl

Inventarnummer
060

Verweis auf zugehörige Objekte
059

Aufbewahrungsort / Standort
auf dem oberen Absatz der Treppe des Nordanbaus auf dem Fußboden lagernd

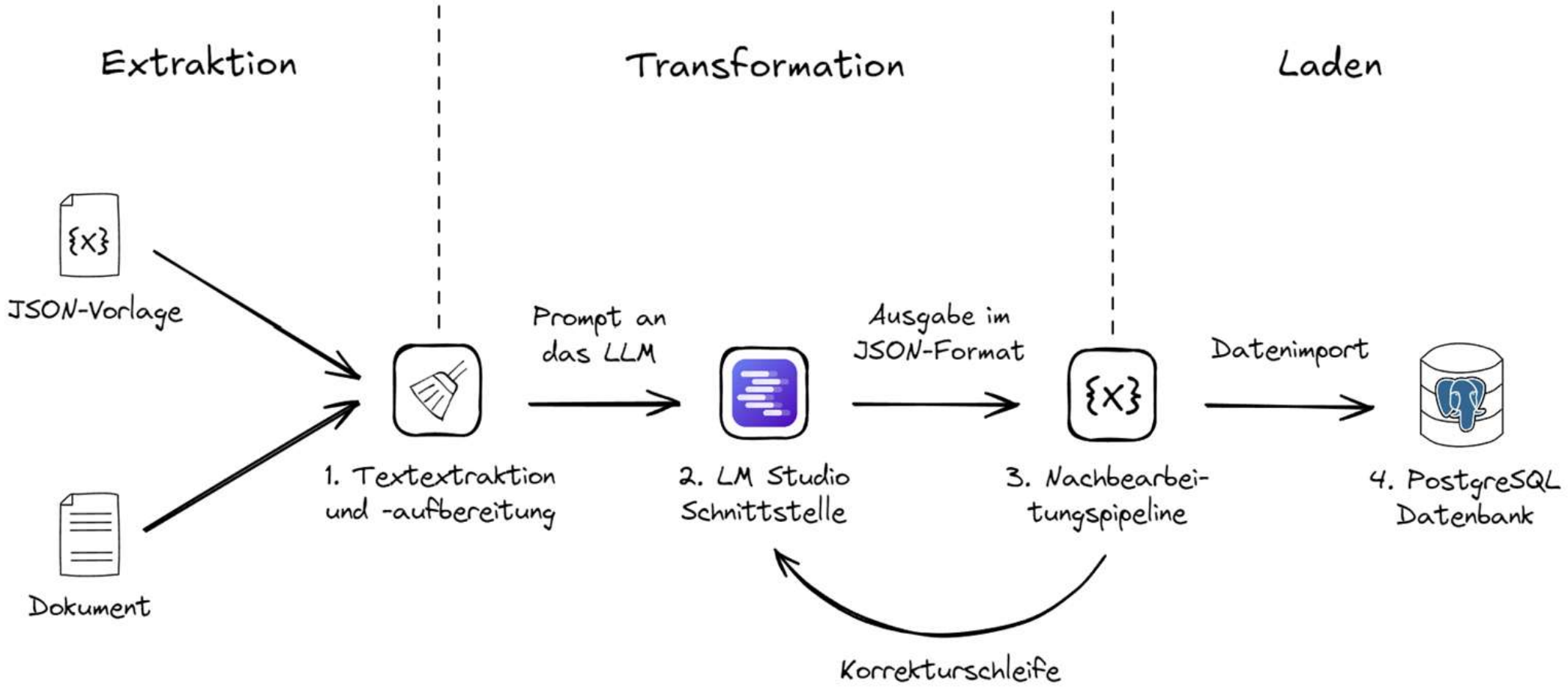
Abweichender Herkunftsort
vorher eingelassen auf der Empore in die Nordwand östlich des Eingangs zum Treppenaufgang

Material/Technik
Stein

Maße
H: 67 cm
B: 102,5 cm
T: cm
L: cm
Gewicht: g
Künstler/Werkstatt

Umsetzung

- Die Schritte im Implementierungsprozess



Prompt

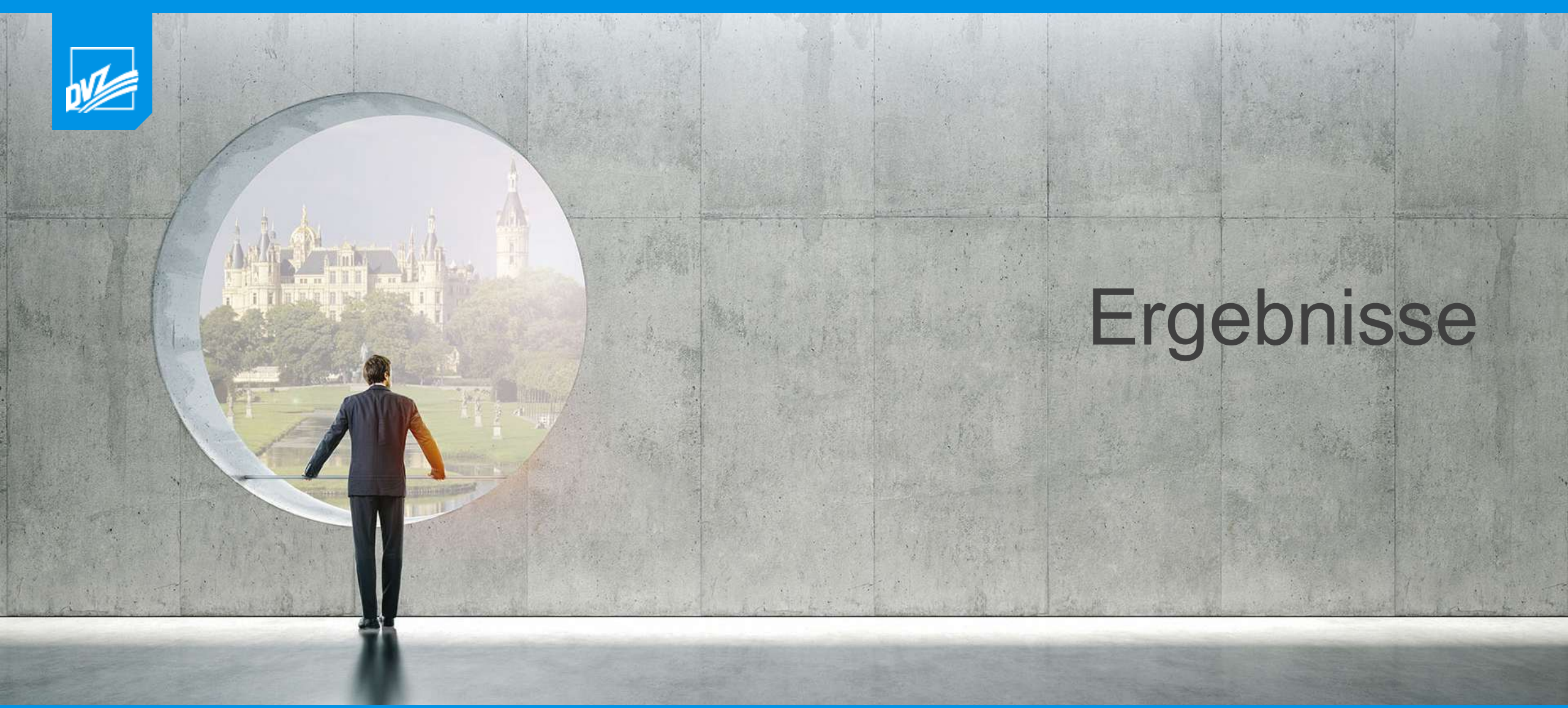
Prompt zur Strukturierung von Dokumenten in ein gültiges JSON-Format

```
In der folgenden Objektbeschreibung wurden Informationen über ein
bestimmtes Kunstgut erfasst:
{document_text}

Aufgabe: Übertrage alle Informationen aus der Objektbeschreibung in
ein gültiges JSON-Objekt anhand der folgenden Vorlage:
{json_template}

Hinweis: Die Informationen sollen ungekürzt und unverändert übertragen
werden. Übertrage für jedes Merkmal den gesamten Text bis das nächste
Merkmal beginnt. Es dürfen keine Informationen verloren gehen.
Verwende die vorgegebenen Schlüssel aus der Vorlage für das JSON-Objekt
in der Antwort. Formatiere alle Werte im JSON-Objekt als Text-String.
Setze die Schlüssel und Werte in doppelte Anführungszeichen, um
sicherzustellen, dass die Antwort der korrekten JSON-Syntax entspricht.
Die Antwort soll ausschließlich das JSON-Objekt und keinen zusätzlichen
Text enthalten.
```

Datenmodell

- Vorlage im JSON Format (29 Merkmale)

```
json_template = {
    "Ortsname:",
    "Kirche": "",
    "Ortskennzahl": "",
    "Kirchengemeinde": "",
    "Bezeichnung": "",
    "Anzahl": "",
    "Inventarnummer": "",
    "Verweis auf zugehörige Objekte": "",
    "Aufbewahrungsort/Standort": "",
    "Material/Technik": "",
    "Maße": "",
    "Gewicht": "",
    "Künstler/Werkstatt": "",
    "Entstehungsort/Herkunft": "",
    "Datierung": "",
    "Beschreibung/Bemerkungen": "",
    "Bildthema/Bildprogramm": "",
    "Inschriften": "",
    "Metallmarken und Auflösung": "",
    "Zustand/Schaden": "",
    "Fassung": "",
    "Restaurierung": "",
    "Leihvertrag": "",
    "Literatur": "",
    "Historische Bildquellen/Archivalische Quellen/Biografische Verweise": "",
    "Stifter": "",
    "Foto Nummern": "",
    "Erfasser": "",
    "Erfassungsdatum/Letzte Änderung": ""
}
```



Ergebnisse

Finetuning Modell-Parameter

Temperatur-Parameter bestimmt die Modellkreativität bei der Antwortgenerierung

- Kreativität zwischen 0 und 2, wobei 0 möglichst deterministische Antworten erzeugt
- Testdurchlauf mit 30 Dokumenten ohne Nachbearbeitung der Ausgabe
- Niedrige Temperatur-Parameter (0,1 - 0,4) konnten höhere Erfolgsrate aufweisen

Parameter	Erfolgsrate	Antwortzeit	Antwortlänge	Token/sec
temp = 1.0	17 von 30 (57%)	72 s	1134 Token	16 t/s
temp = 0.7	18 von 30 (60%)	75 s	1187 Token	16 t/s
temp = 0.4	22 von 30 (73%)	75 s	1181 Token	16 t/s
temp = 0.1	21 von 30 (70%)	75 s	1175 Token	16 t/s

Tabelle 5: Vergleich der Erfolgsraten bei der Strukturierung von 30 Dokumenten in das Zielformat JSON mit unterschiedlichen Anpassungen der Modell-Kreativität

Nachbearbeitungspipeline

Vergleich der Erfolgsraten mit und ohne Nachbearbeitung der Modellantworten

- Bereinigung der generierten Ausgabe soll gültiges JSON-Format sicherstellen
- Korrekturschleife bei fehlerhaften Modell-Ausgaben, die nicht korrigiert werden können
- Effektive Nachbearbeitung verbessert Erfolgsrate signifikant

Performance/ Strukturierung	Erfolgsrate Gesamt (in %)	Korrektur-läufe	Ungültige Antworten	Laufzeit gesamt
ohne Nachbearbeitung	22 von 30 (73%)	-	8 von 30 (27%)	37 min
mit Korrekturschleife	23 von 30 (77%)	8 (27%)	15 von 38 (40%)	44 min
mit Nachbearbeitung	29 von 30 (97%)	-	1 von 30 (3%)	38 min
beides zusammen	30 von 30 (100%)	1 (3%)	1 von 31 (3%)	39 min

Tabelle 6: Vergleich der Erfolgsraten bei der Strukturierung von 30 Dokumenten in das Zielformat JSON mit Korrekturschleife und Nachbearbeitung der Ausgabe

Ergebnisse der Fallstudie

5443 Dokumente konnten mit einer Erfolgsrate von 96% importiert werden

Quell-Ordner	Anzahl Dokumente	Erfolgreich importiert	Korrektur-läufe	Antwort-länge	Antwort-zeit	Laufzeit gesamt
A	842	829 (98%)	55 (7%)	698 Token	38 s	9h 46min
B	1729	1642 (95%)	79 (5%)	651 Token	36 s	ca. 18h
C	1199	1124 (94%)	69 (6%)	786 Token	48 s	ca. 17h
D	1900	1848 (97%)	85 (5%)	817 Token	49 s	ca. 27h
Gesamt	5670	5443 (96%)	288 (5%)	742 Token	43s	ca. 71h
Schnitt	100	96 (96%)	5 (5%)	742 Token	43s	1h 15min

Tabelle 8: Performance des LLM-basierten ETL-Prozesses für verschiedene Datenquellen

Fehlerursachen

515 Fehlversuche auf Dokumente (24%) und LLM-Antworten (76%) zurückzuführen

Quell-Ordner	Durch-läufe	Fehl-versuche	Leere Datei	Token -limit	JSON ungültig	Spalte ungültig
A	897	68 (7%)	0 (0%)	2 (3%)	22 (32%)	44 (65%)
B	1808	166 (9%)	60 (36%)	2 (1%)	59 (36%)	45 (27%)
C	1268	144 (11%)	47 (33%)	3 (2%)	59 (41%)	35 (24%)
D	1985	137 (7%)	6 (5%)	3 (2%)	36 (26%)	92 (67%)
Gesamt	5958	515 (9%)	113 (22%)	10 (2%)	176 (34%)	216 (42%)

Tabelle 9: Fehlerursachen bei der LLM-basierten Strukturierung von Dokumenten

Erkenntnisse

Erkenntnisse

Die Antwortqualität von LLMs wird durch viele Faktoren beeinflusst

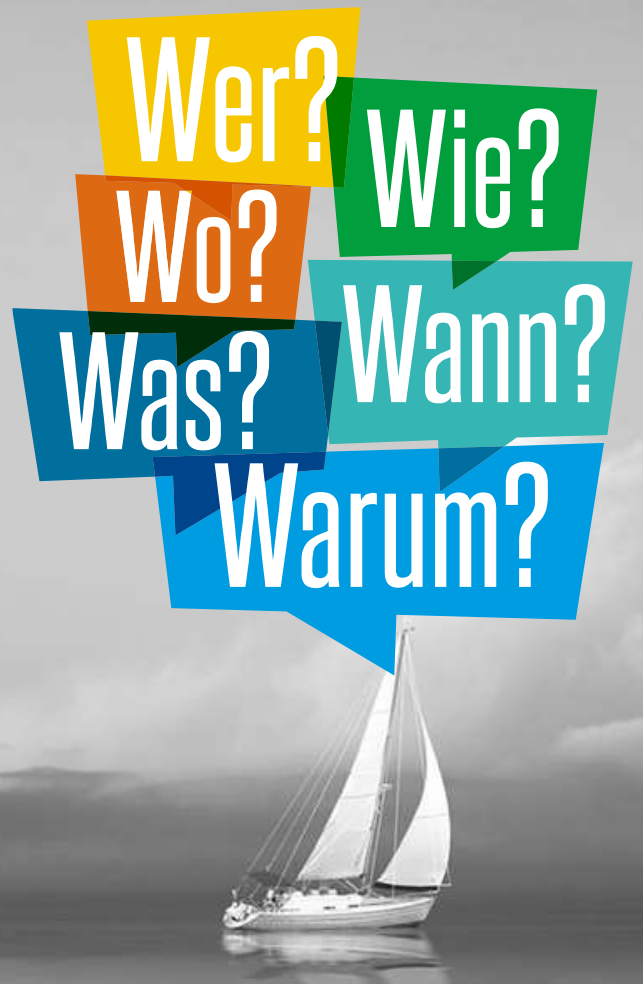
- Viele Stellschrauben, die optimiert werden können
  - Textaufbereitung, Datenmodell, Prompt-Engineering und Modell-Parameter
  - Auswirkungen sind nicht vorhersehbar, da die Funktionsweise eine Black-Box ist
  - Optimierungsprozess nach dem Trial-and-Error-Prinzip
- Nachbearbeitungspipeline konnte die Erfolgsrate signifikant erhöhen
  - LLMs sind keine Alleskönner, sondern nur ein Werkzeug
  - Eine effektive Nachbearbeitung der Modellausgabe ist daher unerlässlich
- Optimierungen müssen individuell auf den Anwendungsfall abgestimmt sein

Fazit und Ausblick

Ergebnisse zeigen Potenzial von LLMs bei der automatisierten Datenstrukturierung

- Erfolgreiche Fallstudie: Über 5000 Dokumente wurden in die Datenbank importiert
- Open-Source-LLMs können unstrukturierte Texte in strukturierte Daten überführen
- LLM-basierter ETL-Prozess kann manuellen Aufwand erheblich reduzieren
- Optimierungsmaßnahmen konnten Ergebnisqualität signifikant steigern
- Ansatz ist auf andere Anwendungsfälle übertragbar (mit kleineren Anpassungen)
- Es können Empfehlungen für ein allgemeines Vorgehen abgeleitet werden

Diskussion



Fragen?

Kontakt

Henry Rotzoll

+ 49 385 4800 532

h.rotzoll@dvz-mv.de

IHR ANSPRECHPARTNER