



Fakultät Informatik

**Vergleichsanalyse von multimodalen Large
Language Models und einer
OCR/YOLO-Pipeline zur
Dokumentenklassifikation für die
Bundesagentur für Arbeit**

Bachelorarbeit im Studiengang Informatik

vorgelegt von

Lukas Müller

Matrikelnummer 3698673

Erstgutachter: Prof. Dr. Natalie Kiesler

Zweitgutachter: Prof. Dr. Korbinian Riedhammer

© 2026

Dieses Werk einschließlich seiner Teile ist **urheberrechtlich geschützt**. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Autors unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen sowie die Einspeicherung und Verarbeitung in elektronischen Systemen.

Kurzdarstellung

Kurze Zusammenfassung der Arbeit, höchstens halbe Seite. Deutsche Fassung auch nötig, wenn die Arbeit auf Englisch angefertigt wird.

Inhaltsverzeichnis

1	Einleitung	1
1.1	Motivation	1
1.2	Problemstellung	2
1.3	Zielsetzung	2
2	Theoretische Grundlagen	3
2.1	Vision-Language Models (VLMs)	3
2.1.1	Vom Sprachmodell zum multimodalen Modell	3
2.1.2	Architekturkomponenten	3
2.2	Vorstellung der Modelle	3
2.2.1	Qwen-2.5-VL	3
2.2.2	Pixtral-12B	3
2.3	Fine-Tuning	3
2.3.1	Parameter-Efficient Fine-Tuning (PEFT)	3
2.4	Information Extraction und Metriken	3
2.4.1	Information Extraction	3
2.4.2	Metriken	3
2.5	Verwandte Arbeiten	3
2.6	Rahmenbedingungen und Infrastruktur	3
3	Datenbasis	5
3.1	Dokumentenarten	5
3.1.1	KG5b	6
3.1.2	Ausbildungsvertrag	8
3.1.3	Sonstige Dokumente	9
3.2	Datenschutz	9
4	Methode und Implementierung	11
5	Evaluation	13
6	Diskussion	15
7	Zusammenfassung	17

Abbildungsverzeichnis	19
Tabellenverzeichnis	21
List of Listings	23
Glossar	25

Kapitel 1

Einleitung

1.1 Motivation

Die öffentliche Verwaltung in Deutschland steht vor einer der größten Entwicklungen ihrer Geschichte. Durch den demografischen Wandel verliert der öffentliche Sektor in den kommenden Jahren eine signifikante Anzahl an erfahrenen Sachbearbeitern, während die Erwartungshaltung der Bürger an digitale und schnell verfügbare Dienstleistungen steigt. Wegen des Personalmangels und gleichzeitig wachsendem Leistungsanspruch rückt die automatisierte oder teilautomatisierte Bearbeitung von Verwaltungsprozessen in den Vordergrund. Diese Unterstützung für die Sachbearbeiter ist keine Option, sondern eine Notwendigkeit um in den kommenden Jahre Handlungsfähig zu bleiben.

Ein Beispiel hierfür ist die Bearbeitung von Kindergeldanträgen bei der Familienkasse der Bundesagentur für Arbeit. Mit einem jährlichen Aufkommen von mehreren Millionen Anträgen und starken saisonale Spitzen, etwa mit Beginn der Ausbildungen im Herbst, stößt die Familienkasse an ihre Kapazitätsgrenze.

Um sich dieser Herausforderung zu stellen, wurde bereits im Oktober 2025 ein erstes teilautomatisiertes System produktiv gesetzt. Dieses System basiert zu einem großen Teil auf einer Kombination aus OCR-Erkennung und Objektklassifizierung mittels YOLO-Modellen.

Obwohl dieses System einen wichtigen Schritt macht, zeigen sich während dem operativen Einsatz und der Entwicklung schon erste Grenzen. Besonders die Erkennung der handschriftlich ausgefüllten Dokumente stellen für diese Pipeline eine enorme Herausforderung dar.

Genau an diesem Punkt setzt die Thesis an. Durch den rasanten Fortschritt im Bereich der Generativen künstlichen Intelligenz öffnen sich neue Lösungen für Bearbeitung der Anträge. Insbesondere VLMs, die visuelle Informationen und Texte simultan verarbeiten, stellen eine vielversprechende Möglichkeit dar, die YOLO-OCR-Pipeline zu ersetzen. Sie können flexibel auf unbekannte Layouts reagieren, handschriftlichen Text extrahieren und den Kontext interpretieren.

Mit der Entwicklung der VLMs ergibt sich die Chance auf eine weitere Verbesserung für die Sachbearbeiter, jedoch muss getestet werden, ob diese Technologie bereits reif genug ist, um im öffentlichen Sektor eingesetzt zu werden.

1.2 Problemstellung

1.3 Zielsetzung

- Forschungsfragen

Kapitel 2

Theoretische Grundlagen

2.1 Vision-Language Models (VLMs)

2.1.1 Vom Sprachmodell zum multimodalen Modell

2.1.2 Architekturkomponenten

2.2 Vorstellung der Modelle

2.2.1 Qwen-2.5-VL

2.2.2 Pixtral-12B

2.3 Fine-Tuning

2.3.1 Parameter-Efficient Fine-Tuning (PEFT)

2.4 Information Extraction und Metriken

2.4.1 Information Extraction

2.4.2 Metriken

2.5 Verwandte Arbeiten

2.6 Rahmenbedingungen und Infrastruktur

Kapitel 3

Datenbasis

3.1 Dokumentenarten

Im Rahmen der Kindergeldbeantragung sind verschiedene Nachweise gültig. Zu den anerkannten Dokumententypen zählen der offizielle Vordruck der Bundesagentur für Arbeit (KG5b)[?] sowie Ausbildungsverträge. Zusätzlich laden Kunden häufig weitere Unterlagen, wie beispielsweise Schulbescheinigungen, im Portal hoch. Da diese für den Kindergeldantrag nicht im Fokus stehen, werden sie im Folgenden unter der Kategorie „Sonstiges“ zusammengefasst.

Ziel der Arbeit ist die Extraktion spezifischer Informationen aus den genannten Dokumentenklassen. Um die gewonnenen Informationen aus dem Dokument bereitzustellen, wird ein flaches JSON-Schema verwendet. Die Wahl dieses Formats wird dadurch begünstigt, dass moderne LLMs durch ihr Training bereits eine hohe Zuverlässigkeit in der Generierung valider JSON-Strukturen aufweisen.

Die Qualität der Dokumente, die im Onlineportal hochgeladen werden, ist sehr unterschiedlich. Neben Scans mit guter Belichtung und hoher Auflösung enthält der Datensatz auch Fotos, die aus verschiedenen Winkeln und Entfernungen aufgenommen wurden. Eine zusätzliche Herausforderung ist, dass die Dokumente häufig handschriftlich ausgefüllt sind.

In den Test- und Trainingsdatensätzen kommen diese Probleme in unterschiedlichen Konstellationen vor. Der Testdatensatz umfasst insgesamt 60 Dokumente, die sich gleichmäßig auf 20 Ausbildungsverträge, 20 KG5b-Formulare und 20 sonstige Dokumente verteilen. Der Trainingsdatensatz setzt sich aus 165 KG5b-Formularen, 227 Ausbildungsverträgen und 218 sonstigen Dokumenten zusammen.

Die Erstellung Ground Truth erfolgte in einem zweistufigen Verfahren. Zunächst wurden die JSON-Strukturen für den Testdatensatz vollständig manuell erstellt. Dieser Testdatensatz diente anschließend dazu, den System-Prompt zu optimieren, bis die Ergebnisse des Basis-Modells eine zufriedenstellende Qualität erreichten. Um den Annotationsaufwand für den Trainingsdatensatz zu reduzieren, wurde dieser optimierte Prompt für ein sogenanntes Pre-Labeling genutzt. Das Modell generierte dabei erste Vorschläge für die JSON-Schemas,

welche im Anschluss manuell validiert und korrigiert wurden. Dieser modellgestützte Annotationsprozess (Model-Assisted Labeling) ermöglichte eine effiziente Erstellung der Trainingsdaten bei gleichbleibend hoher Datenqualität.

3.1.1 KG5b

Der Vordruck KG5b ist, wie bereits erwähnt, ein offizielles Dokument der Bundesagentur für Arbeit, welches als Bescheinigung der Ausbildungsstätte dient. Volljährige Kinder weisen damit gegenüber der Familienkasse den Fortbestand ihrer Ausbildung nach, was die Voraussetzung für den weiteren Kindergeldbezug ist.

Da es sich um ein standardisiertes Formular handelt, ist das Layout gleichbleibend. Theoretisch vereinfacht dies die Extraktion der Informationen, da relevante Felder an denselben Positionen zu erwarten sind. In der Praxis ergibt sich jedoch eine hohe Varianz durch den Ausfüllprozess. Obwohl der Vordruck digital ausfüllbar angeboten wird, erfolgt die Bearbeitung in der Regel handschriftlich. Zudem erzwingen Unterschriften und Firmenstempeln einen doppelten Medienbruch. Das Dokument wird ausgedruckt, bearbeitet und anschließend wieder digitalisiert.

Abbildung 3.1 zeigt ein exemplarisch ausgefülltes KG5b-Formular mit Markierung der zu extrahierenden Felder.

Das Zielformat für die Extraktion wird durch das folgende JSON-Schema definiert:

```

1 {
2   "file_name": {
3     "type": "kg5b",
4     "name_child": "name, vorname",
5     "birthday_child": "DD.MM.YYYY",
6     "start_date_apprenticeship": "DD.MM.YYYY",
7     "end_date_apprenticeship": "DD.MM.YYYY",
8     "date_document": "DD.MM.YYYY",
9     "stamp_company": true,
10    "signature_company": true,
11    "signature_child": true,
12    "signature_legal_guardian": true,
13    "apprenticeship_ended": true,
14    "exam_month": "MM"
15  }
16 }
```


Listing 3.1: JSON-Schema für KG5b

FamName und Vorname der Kindergebührenberechtigten Person


Mustermann, Mathias

KindergebHr:

0283475234



K



Familienkasse

Telefonische Rückfrage tagtäglich unter Nr.:

Erklärung zum Ausbildungsverhältnis

Angaben zum Kind

FamName
Mustermann

name

Vorname
Max

birthday_child

Geburtsdatum
01.03.2002

Angaben zum Ausbildungsverhältnis

Berufsausbildung als
Apotheker

start_date_apprenticeship

Beginn der Ausbildung
01.10.2025

apprenticeship_ended

Die Ausbildung/beruf ist noch nicht beendet. (Bitte Nachweise beifügen!)

Die Ausbildung dauert voraussichtlich bis
01.10.2028

end_date_apprenticeship

Das Ausbildungsverhältnis ist beendet. (Bitte Nachweise beifügen!)

a) Hat der/die Auszubildende die Abschlussprüfung bestanden? ☐ ja ☐ nein

Falls ja:
Letzter Prüfungstag war der:
Die offizielle schriftliche Unterrichtung über das Gesamtergebnis der Abschlussprüfung (Ausbildung oder Zugang der Bescheinigung bzw. vorläufige Bescheinigung des Prüfungsausschusses über das Bestehen der Prüfung oder Ähnliches) erfolgte am:

b) Ist die Abschlussprüfung nach Ablauf der vertraglichen Ausbildungszeit abgelegt worden? ☐ ja ☐ nein

Falls ja:
Wurde das Ausbildungsverhältnis bis zur offiziellen schriftlichen Unterrichtung über das Gesamtergebnis der Abschlussprüfung verlängert? ☐ ja ☐ nein

c) Wurde das Ausbildungsverhältnis aus anderen Gründen beendet? ☐ ja ☐ nein

Falls ja:
Das Ausbildungsverhältnis endete am:


d) Wurde der/die Auszubildende vom Ausbildungsbetrieb als Arbeitnehmer(in) übernommen? ☐ ja ☐ nein

Falls ja:
Beginn des Arbeitsverhältnisses:

Bestätigung des Ausbildungsbetriebes stamp_company

Die obigen Angaben zum Ausbildungsverhältnis sind richtig und vollständig

Apotheker Max Marktstr.1 66666 Markt Tel.: 66 666 6666 66



Unterchrift

Unterchrift

Firmenstempel des Ausbildungsbetriebes

Hinweis an den Kindergebührenberechtigten: Bitte füllen Sie Punkt e) erst aus, nachdem der Ausbildungsbetrieb die Angaben zum Ausbildungsverhältnis bestätigt hat!

e) Hat das Kind bei einem anderen Arbeitgeber ein Arbeitsverhältnis begonnen oder wird es demnächst ein Arbeitsverhältnis bei einem anderen Arbeitgeber beginnen? ☐ ja ☐ nein

Falls ja:
(voraussichtlicher) Beginn des Arbeitsverhältnisses:

Wir versichern, dass unsere Angaben vollständig sind und der Wahrheit entsprechen. Uns ist bekannt, dass wir alle Änderungen, die für den Anspruch auf Kindergeld von Bedeutung sind, unverzüglich der Familienkasse mitzuteilen haben. Den Inhalt des Merkblattes Kindergeld (zu finden unter www.bzst.de oder www.familienkasse.de) haben wir zur Kenntnis genommen.

Hinweis zum Datenschutz: Ihre Daten werden gemäß der §§ 31, 62 bis 78 Einkommensteuergesetz und der Regelungen der Abgabenordnung bzw. aufgrund des Bundeskindergeldgesetzes und des Sozialgesetzbuches verarbeitet. Zweck der Verarbeitung der Daten ist die Prüfung Ihres Anspruchs auf Kindergeld. Nähere Informationen über die Verarbeitung Ihrer Daten durch die Familienkasse und um Ihren Rechten nach Artikel 13 bis 22 der Datenschutz-Grundverordnung erhalten Sie im Internet auf der Seite Ihrer Familienkasse (zu finden unter www.arbeitsagentur.de/nachrichte/familienkasse), auf der auch die Kontaktdaten der/des Datenschutzbeauftragten bereitgestellt sind. Kindergeldangaben in der Regel nach dem Ende der Kindergeldzahlung noch für 6 Jahre aufbewahrt.

Unterchrift der kindergebührenberechtigten Person bzw. der gesetzlichen Vertretung

Mustermann

signature_legal_adult

Unterchrift des volljährigen Kindes

Max Mustermann

signature_child

Alle Eingaben löschen

Drucken

Speichern

(a) Erste Seite

(b) Zweite Seite

Abbildung 3.1: KG5b-Formular

Eine besondere Herausforderung bei der Verarbeitung stellen die binären Informationen dar, wie etwa die Checkbox für das Feld `apprenticeship_ended`. Hier muss das Modell visuell unterscheiden, ob ein Kasten leer, angekreuzt oder durchgestrichen ist. Ebenso kritisch ist die Detektion von `signature_company` und `stamp_company`. Im Gegensatz zu Textfeldern ist hier nicht der textliche Inhalt des Stempels relevant, sondern lediglich dessen Vorhandensein. Die Stempel sind oft blass oder eingefärbt, was besonders bei Schwarz-Weiß Scans zu einer schlechten Qualität führt. Zudem liegen die Felder `date_document` und `signature_company` nah beieinander. Unterschriften sind regelmäßig größer als ihr vorgesehener Platz, wodurch das Feld `date_document` überdeckt wird. Zusätzlich werden die Stempel häufig zusammen mit der Unterschrift auf der linken Seite platziert.

3.1.2 Ausbildungsvertrag

Im Gegensatz zu den standardisierten KG5b-Formularen weisen die Ausbildungsverträge eine signifikant höhere Varianz auf. Dies ist auf die Vielzahl unterschiedlicher zuständiger Stellen (z.,B. Industrie- und Handelskammern, Handwerkskammern, Ärztekammern) und Firmen zurückzuführen, die jeweils ein eigenes Layout definieren. Die Vielfalt der Dokumentenstruktur reicht dabei von formularbasierten Layouts bis hin zu unstrukturierten Fließtexten.

Eine weitere Herausforderung stellt der Umfang der Dokumente dar. Ausbildungsverträge umfassen häufig mehrere Seiten, wobei relevante Informationen und irrelevante Informationen, wie zum Beispiel gesetzliche Rahmenbedingungen, AGBs oder Rechtsbelehrungen, gemischt auftreten. Dies erschwert die Fokussierung des Modells auf die relevanten Daten.

Analog zu den KG5b-Formularen liegt auch hier durch die notwendigen Unterschriften und Stempel ein Medienbruch vor. Ein wesentlicher Vorteil gegenüber den handschriftlich ausgefüllten Anträgen besteht jedoch darin, dass die inhaltlichen Daten der Verträge fast ausschließlich maschinenschriftlich vorliegen, was die Zeichenerkennung erleichtert.

Das Zielformat für die Extraktion wird durch das folgende JSON-Schema definiert:

```

1 {
2   "file_name": {
3     "type": "vertrag",
4     "name_child": "name, _vorname",
5     "birthday_child": "DD.MM.YYYY",
6     "start_date_apprenticeship": "DD.MM.YYYY",
7     "end_date_apprenticeship": "DD.MM.YYYY",
8     "date_document": "DD.MM.YYYY",
9     "stamp_company": true,
10    "signature_company": true,
11    "signature_child": true,
12    "signature_legal_guardian": true
13  }
14 }
```

Listing 3.2: JSON-Schema Vertrag

Die Felder stimmen weitestgehend mit denen der KG5b-Formulare überein, jedoch entfallen die Felder `exam_month` und `apprenticeship_ended`, da diese selten auf Verträgen zu finden sind.

3.1.3 Sonstige Dokumente

Die Kategorie **Sonstiges** steht als Auffangklasse für alle restlichen Dokumente bereit. Hier gehören zum Beispiel Schulbescheinigungen, Anträge auf Eintragung bei der Handelskammer oder Studienbescheinigungen dazu. Da diese Dokumente keine Relevanz für die Weiterbeantragung des Kindergeldes bei volljährigen Auszubildenden hat, werden keine Informationen benötigt. Somit ist rein die Klassifizierung wichtig, weshalb das extrahierte JSON-Schema folgendermaßen aussieht:

```
1 {  
2   "file_name": {  
3     "type": "sonstiges"  
4   }  
5 }
```

Listing 3.3: JSON-Schema Sonstiges

3.2 Datenschutz

Da es sich bei allen verwendeten Dokumenten um Echtdaten aus den Fachverfahren handelt, enthalten diese sensible personenbezogene Daten. Der Schutz dieser Daten hat höchste Priorität.

Um die Sicherheit bei der Verarbeitung der Echtdaten zu gewährleisten, wurde ausschließlich auf die Infrastruktur der Bundesagentur für Arbeit zurückgegriffen. Die Daten liegen in einem gesicherten Netzwerkbereich, auf den nur ein autorisierter Personenkreis Zugriff hat. Durch die Verwendung von Echtdaten wird die Qualität der Ergebnisse deutlich gesteigert.

Alle in dieser Arbeit abgebildeten Dokumentenbeispiele wurden jedoch synthetisch erzeugt und beinhalten keinerlei sensible Daten.

Kapitel 4

Methode und Implementierung

Kapitel 5

Evaluation

Kapitel 6

Diskussion

Kapitel 7

Zusammenfassung

Abbildungsverzeichnis

3.1 KG5b-Formular	7
-----------------------------	---

Tabellenverzeichnis

List of Listings

3.1 JSON-Schema für KG5b	6
3.2 JSON-Schema Vertrag	8
3.3 JSON-Schema Sonstiges	9

Glossar

FamKa Familienkasse der Bundesagentur für Arbeit. i

JSON JavaScript Object Notation. i, 5, 6, 8, 9

LLM Large Language Model. i, 5

MLLM Multimodal Large Language Model. i

OCR Optical Character Recognition. i, 1

PEFT Parameter-Efficient Fine-Tuning. i

VLM Vision Language Model. i, 1, 2

YOLO You Only Look Once. i, 1