

# Digitale Bauleitplanung – Wie KI Systeme Bebauungspläne verstehen

Masterarbeit im Studiengang Informatik M.Sc.

Vertiefung Data Science

Fakultät Informatik

Technische Hochschule Augsburg

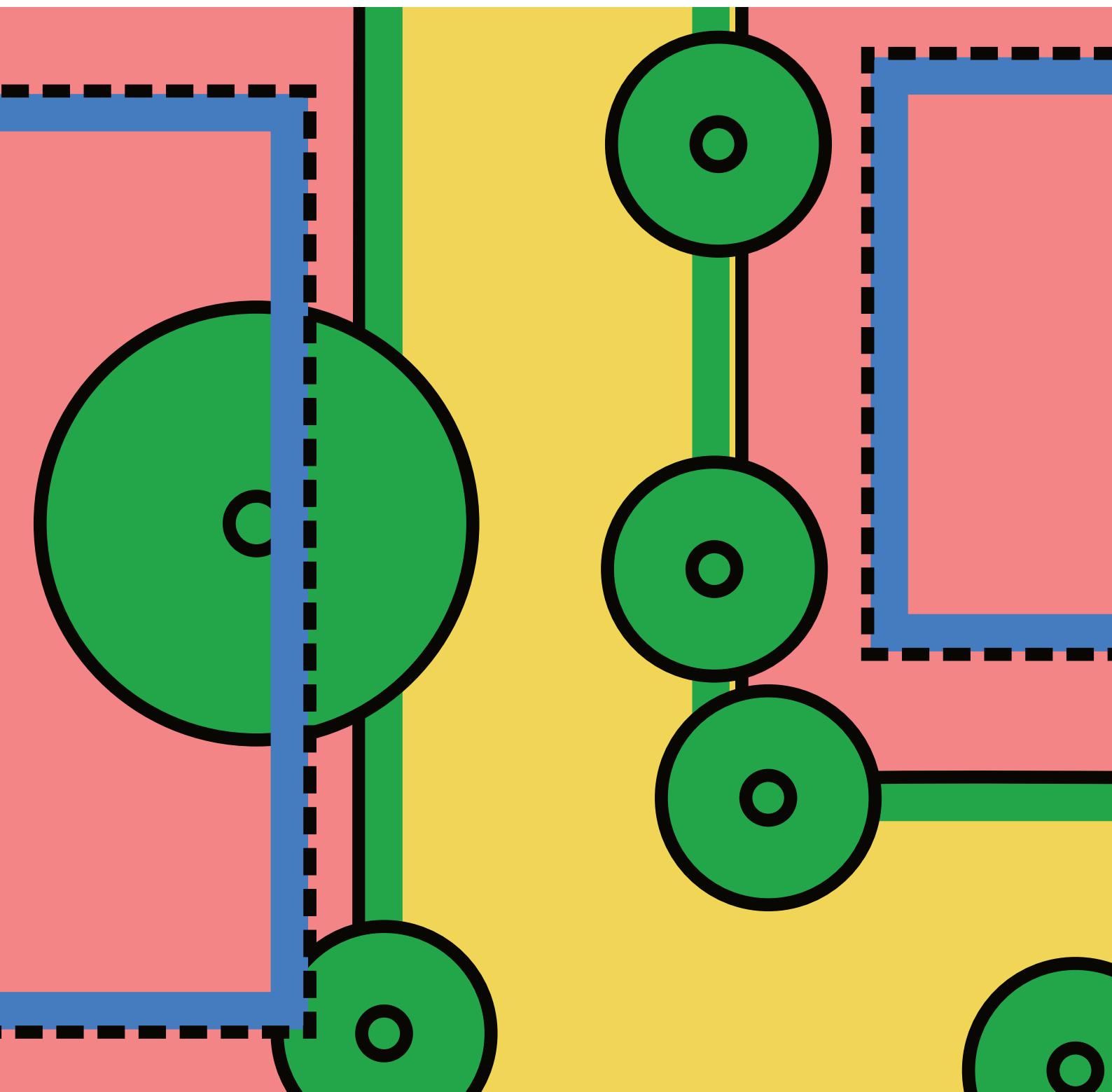
Vorgelegt von Michael Schwarz

Betreuer:innen: Prof. Dr. Kratsch, Prof. Dr.-phil. Zarcone

In Zusammenarbeit mit der Firma credium GmbH



6. September 2024



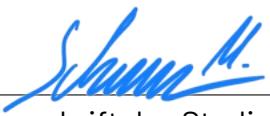
### **Erklärung zur Abschlussarbeit**

Hiermit versichere ich, die eingereichte Abschlussarbeit selbständig verfasst und keine andere als die von mir angegebenen Quellen und Hilfsmittel benutzt zu haben. Wörtlich oder inhaltlich verwendete Quellen wurden entsprechend den anerkannten Regeln wissenschaftlichen Arbeitens zitiert.

Ich erkläre weiterhin, dass die vorliegende Arbeit noch nicht anderweitig als Abschlussarbeit eingereicht wurde.

Das Merkblatt zum Täuschungsverbot im Prüfungsverfahren der Technischen Hochschule Augsburg habe ich gelesen und zur Kenntnis genommen. Ich versichere, dass die von mir abgegebene Arbeit keinerlei Plagiate, Texte oder Bilder umfasst, die durch von mir beauftragte Dritte erstellt wurden.

Augsburg, 06.09.24  
Ort, Datum

  
\_\_\_\_\_  
Unterschrift des Studierenden

## **Danksagungen**

Vielen Dank an meine Betreuer:innen Prof. Dr. Wolfgang Kratsch und Prof. Dr.-phil. Alessandra Zarcone. Einen großen Dank auch an Dr. Timm Tränkler, Dr. Lars Wederhake, Lukas Geirhos und Thomas Malchers von der Firma credium GmbH, sowie Klaus Kaufmann und Günter Hascher von der Stadt Laichingen und den vielen weiteren freiwilligen Expert:innen.

## **Abstract**

Bebauungspläne (B-Pläne) sind ein wichtiges Werkzeug in der Bauleitplanung, welche Regelungen über die mögliche Nutzung und Bebauung von genau definierten Grundstücken enthalten und damit eine geordnete städtebauliche Entwicklung gewährleisten. Aktuelle Bauplanungs- und Baugenehmigungsprozesse gestalten sich jedoch sehr langwierig, weshalb eine effizientere Prozessgestaltung durch Digitalisierung angestrebt wird. Hierzu stellt sich die Forschungsfrage „*Inwiefern sind Multimodel Large Language Models (MLLMs) in der Lage B-Pläne zu verstehen?*“ [RQ1], um Prozesse in der Bauleitplanung zu unterstützen.

Diesbezüglich untersuchte ich anhand qualitativer Forschungsmethoden mögliche Anwendungsfälle für MLLMs und bewertete das Textverständnis von GPT-4o. Via multiperspektivischer Expert:innen-Interviews wurde unter anderem sichtbar, dass große Mehrwerte in der Automatisierung von Arbeitsschritten, einem effizienten Zugriff auf Daten und einer datenbasierten Entscheidungsfindung liegen. Für alle diese Fälle sind maschinenlesbare Daten eine wesentliche Voraussetzung. Hierzu wurden Experimente durchgeführt, dessen Ergebnisse zeigen, dass GPT-4o grundlegende Textinformationen in ausreichender Qualität extrahieren und bezüglich der Prüfschritte im Baugenehmigungsprozess relevante Inhalte korrekt verknüpfen und logisch aufbereiten konnte. Die Qualität der Ergebnisse war jedoch stark von der Vorverarbeitung und der Struktur der B-Pläne abhängig.

In Bezug auf die Forschungsfrage „*Können MLLMs die Inhalte von maschinenlesbaren Bebauungsplänen mindestens genauso gut verstehen, wie Menschen dies können?*“ [RQ2], zeigt sich, dass GPT-4o insbesondere in der Kombination und Analyse von Textinhalten Stärken hat, ein grundlegendes Verständnis jedoch noch nicht erreichen konnte. Eine Zusammenarbeit zwischen Mensch und KI im Rahmen einer RAG Anwendung erscheint allerdings vielversprechend, was in einer weiterführenden Forschungsarbeit erprobt werden könnte.

# Inhaltsverzeichnis

<b>1 Einleitung</b>	<b>1</b>
1.1 Motivation und Thema . . . . .	1
1.2 Ziel und Beitrag . . . . .	1
1.3 Inhaltlicher Überblick . . . . .	2
<b>2 Domänenwissen Bauleitplanung</b>	<b>3</b>
2.1 Vorgehen und Methodik . . . . .	3
2.2 Ergebnisse . . . . .	5
2.2.1 Grundwissen Bauleitplanung . . . . .	6
2.2.1.1 Einordnung der Bauleitplanung . . . . .	6
2.2.1.2 Basiswissen über Bauleitpläne . . . . .	9
2.2.1.3 Grundlagen des Bauplanungs- und Genehmigungsvorprozess . . . . .	12
2.2.2 Insights und Leitfragen . . . . .	17
2.2.2.1 Strategien . . . . .	17
2.2.2.2 Prozesse . . . . .	19
2.2.2.3 Daten . . . . .	21
2.2.3 Use Case zur prototypischen Umsetzung . . . . .	23
<b>3 Multimodal Large Language Models</b>	<b>25</b>
3.1 Vorgehen und Methodik . . . . .	25
3.2 Ergebnisse . . . . .	26
3.2.1 (Text-)Transformer-Architektur . . . . .	26
3.2.1.1 Input-Embeddings Layer . . . . .	28
3.2.1.2 Positional-Encoding Layer . . . . .	30
3.2.1.3 Attention Layer . . . . .	31
3.2.2 Vision-Transformer-Architektur . . . . .	34
3.2.3 Zusammenführung von Text und Vision . . . . .	35
3.2.3.1 CLIP (Contrastive Language-Image Pre-Training) . . . . .	36
3.2.3.2 LLaVA (Large Language and Vision Assistant) . . . . .	36
<b>4 Experimente und Evaluation</b>	<b>39</b>
4.1 Vorgehen und Methodik . . . . .	39
4.2 Ergebnisse . . . . .	41
4.2.1 Aufbau der Experimente . . . . .	41
4.2.1.1 MLLM Verständnislevel . . . . .	41
4.2.1.2 Datensatz . . . . .	42
4.2.2 Durchführung der Experimente . . . . .	43
4.2.2.1 Phase 1: Daten aufbereiten . . . . .	44
4.2.2.2 Phase 2: Daten extrahieren . . . . .	44
4.2.2.3 Phase 3: Daten kombinieren . . . . .	48
4.2.3 Evaluation der Ergebnisse . . . . .	48
4.2.3.1 Zeichenerklärung . . . . .	49
4.2.3.2 Planzeichnung . . . . .	50
4.2.3.3 Textteil . . . . .	51
4.2.3.4 Prüfungsschema . . . . .	53

4.2.4 Interpretation der Ergebnisse . . . . .	55
<b>5 Fazit und Ausblick</b>	<b>58</b>
5.1 Beantwortung der Forschungsfragen . . . . .	58
5.2 Anmerkung zu den Forschungsergebnissen . . . . .	59
5.3 Weiterführende Forschung . . . . .	59
5.4 Schlusswort . . . . .	59
<b>6 Anhang</b>	<b>60</b>
6.1 Expert:innen-Interviews . . . . .	60
6.2 Experimente und Evaluation . . . . .	60
6.3 Präsentationen . . . . .	60
6.4 Sonstiges . . . . .	60
<b>Glossar</b>	<b>60</b>
<b>Abkürzungsverzeichnis</b>	<b>65</b>
<b>Abbildungsverzeichnis</b>	<b>66</b>
<b>Tabellenverzeichnis</b>	<b>68</b>
<b>Literaturverzeichnis</b>	<b>68</b>

# 1 Einleitung

## 1.1 Motivation und Thema

Laut einer Studie von 2016 der Firma McKinsey befand sich die Baubranche auf dem letzten Platz des Digitalisierungsindizes (McKinsey 2016). Fast zehn Jahre nach Veröffentlichung der Studie wurde der Digitalisierungsstand in der Baubranche erneut von dem Fraunhofer-Institut für experimentelles Software Engineering (=IESE) untersucht. Diese Studie zeigte, dass vor allem Inkompatibilitäten beim Datenaustausch, eine mangelhafte Datenqualität und ein großer Datenbestand zu Herausforderungen bei der Digitalisierung des Baubereichs führen (Feth et al. 2023).

Besonders deutlich werden diese Herausforderungen in der Bauleitplanung sichtbar. Sie ist das wichtigste Planungswerkzeug zur Lenkung und Ordnung der städtebaulichen Entwicklung einer Kommune in Deutschland (Wikipedia 2024d). Im Wesentlichen dient hierzu der sogenannte B-Plan, der die Einzelheiten der Flächennutzung und die Genehmigung von städtebaulichen Maßnahmen rechtsverbindlich regelt; ob und wie ein Grundstück bebaut werden darf. Bauanträge weisen jedoch oft lange Verarbeitungszeiten auf, bis eine Baugenehmigung erteilt wird. Das ist unter anderem auf viele bestehende analoge Prozesse zurückzuführen. Vor diesem Hintergrund wandte sich die Stadt Laichingen an die Firma cedium GmbH, um einen Einblick zu erhalten, wie deren Baugenehmigungsprozesse mithilfe von Künstliche Intelligenz (KI)-Methoden automatisiert und unterstützt werden könnte.

## 1.2 Ziel und Beitrag

Das Ziel meiner Masterarbeit ist es zu untersuchen, inwiefern MLLMs B-Pläne verarbeiten können. Hierzu versuche ich folgende Forschungsfrage zu beantworten: „*Inwiefern sind MLLMs in der Lage B-Pläne zu verstehen?*“ [RQ1]. Der Fokus der Untersuchungen liegt auf dem Textverständnis von MLLMs und nicht auf den grafischen Aspekten eines B-Plans. Als Leitlinien der Forschung dienen folgende Unterfragen:

1. *Wie gut ist die Qualität der extrahierten Daten?* [SQ1.1] Zum Beispiel in Bezug auf Korrektheit und Konsistenz.
2. *Inwiefern kann ein Verständnis zu den extrahierten Informationen geschaffen werden?* [SQ1.2] Zum Beispiel bei Fragen zur Baugenehmigung.

Im Rahmen eines Baugenehmigungsverfahrens soll anhand eines prototypischen KI-Systems abschließend die Forschungsfrage „*Können MLLMs die Inhalte von maschinenlesbaren Bebauungsplänen mindestens genauso gut verstehen, wie Menschen dies können?*“ [RQ2] geklärt werden.

Das Forschungsgebiet der generativen Sprachmodelle ist äußerst dynamisch, sodass aktuell nur wenige bis gar keine Vergleichsstudien bzgl. des Verständnisses von MLLMs zu B-Pläne existieren. Meine Forschungsergebnisse bieten einen grundlegenden Einblick zum aktuellen Stand eines state-of-the-art MLLM wie GPT-4o in Bezug auf die Verarbeitungsfähigkeiten von B-Pläne. Dadurch wird sichtbar, welche Arbeitsschritte MLLMs besonders gut lösen können und welche Forschungsrichtungen sich als entsprechend vielversprechend erweisen.

### **1.3 Inhaltlicher Überblick**

Zunächst wird in Kapitel 2 relevantes Domänenwissen vermittelt und Anwendungsfälle vorgestellt. Im Anschluss werden die technischen Grundlagen zu MLLMs in Kapitel 3 detailliert erläutert, um ein Verständnis für das eingesetzte Modell GPT-4o zu schaffen. Abschließend wird in Kapitel 4 der angewandte Forschungsteil beschrieben, was das Vorgehen der durchgeführten Experimente und die Evaluation der erzielten Ergebnisse umfasst.

## 2 Domänenwissen Bauleitplanung

Im ersten Schritt arbeitete ich mich in die Domäne Bauleitplanung ein. Die Methodik und Ergebnisse werden auf den folgenden Seiten dargelegt.

### 2.1 Vorgehen und Methodik

Abbildung 1 veranschaulicht den Prozess der Einarbeitung in die Domäne, sowie der Erarbeitung und Auswahl eines Use Case zur prototypischen Umsetzung. Der Prozess ist stark an das Design Thinking Framework des Double Diamond angelehnt, der aus vier Phasen besteht: Entdecken, Definieren, Entwickeln und Liefern. Der Double Diamond-Prozess legt großen Wert auf Nutzer:innenzentrierung, sowie ein realistisches Verständnis der Bedürfnisse und Probleme der Stakeholder:innen. Ein weiterer Aspekt des Double Diamond ist der Wechsel zwischen konvergenten und divergenten Vorgehensweisen, wodurch ein Thema umfassend erforscht und anschließend die gesammelten Informationen verdichtet werden. Das iterative Vorgehen ermöglichte Fehlannahmen und -entscheidungen frühzeitig erkennen und korrigieren zu können.

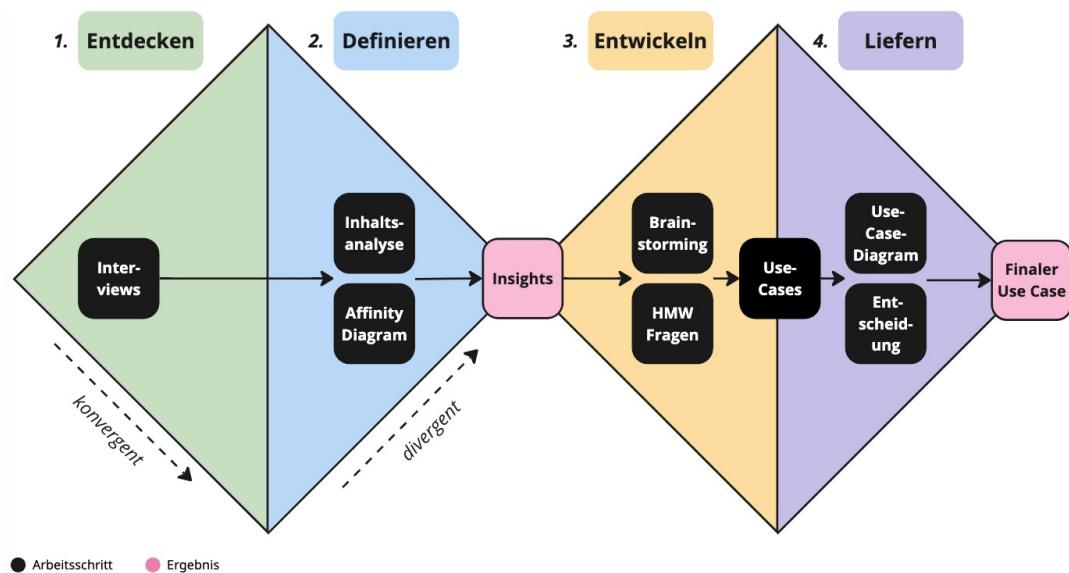


Abbildung 1: Einarbeitung in die Domäne anhand der Double Diamond Methode (Eigene Darstellung. Angelehnt an Council (2023))

Die Phase **Entdecken** hatte das Ziel möglichst viele Informationen über die Domäne Bauleitplanung zu den thematischen Schwerpunkten B-Plan und Baugenehmigungsprozess zu sammeln. Dies erfolgte anhand von Interviews mit Expert:innen aus dem Fachbereich. Zusätzlich wurde der Bauplanungsprozess betrachtet, da die Unterlagen des Bauantrags einen starken Einfluss auf den Baugenehmigungsprozess nehmen. Unterschiedliche Perspektiven und praktische Einblicke bildeten die Datenbasis meiner Forschung, um Komplexität und Mehrwerte hinter potenziellen Use Cases verstehen zu können. Als Qualitative Forschungsmethode wurden unstrukturierte Expert:innen-Interviews durchgeführt.

Die Spontanität, welche unstrukturierte Interviews bieten, hatte mehrere Vorteile: Hierdurch war es möglich, auf Interviewergebnisse von bereits durchgeföhrten Interviews Bezug zu nehmen, um beispielsweise Aussagen aus einer weiteren Perspektive beleuchten zu können. Außerdem konnten nach Bedarf individuelle thematische Interviewschwerpunkte gesetzt werden, was essenziell dazu beigetragen hat, eine hochwertige Datenbasis zu generieren (Tegan 2024). Die teilnehmenden Personen wurden im Hinblick auf ihre berufliche Rolle und Expertise ausgewählt. Themen wie Prozesse/Arbeitsschritte, Regularien, Herausforderungen, Stakeholder:innen, Daten/Dokumente und KI Potenziale bildeten einen flexiblen Rahmen für den thematischen Aufbau eines Interviews. Abbildung 2 veranschaulicht die unterschiedlichen beruflichen Bereiche der teilnehmenden Personen und die jeweiligen Berufsrichtungen.

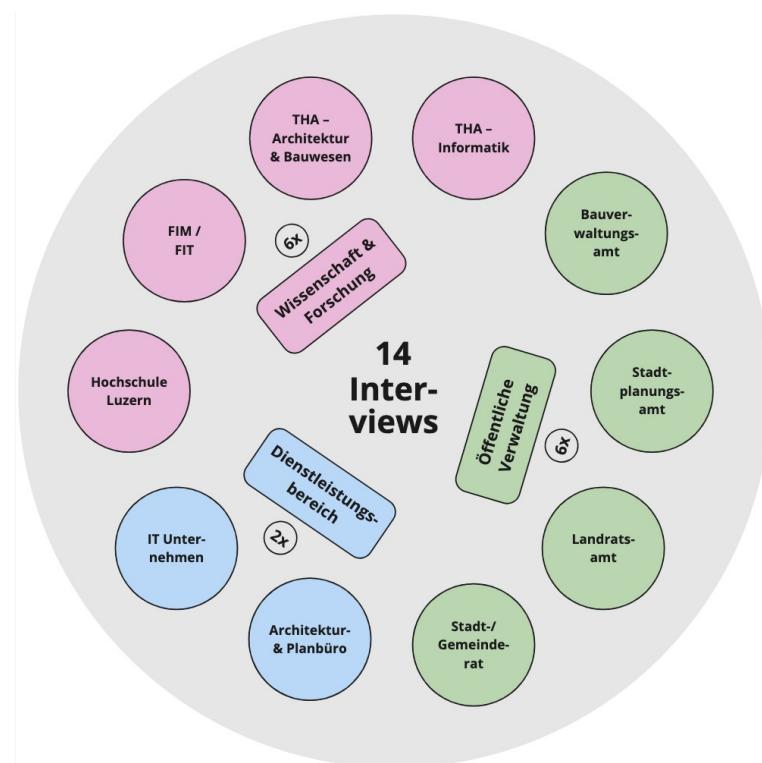


Abbildung 2: Überblick der durchgeföhrten Interviews und eingenommenen Perspektiven (Eigene Darstellung)

Im Laufe der Forschung stellte sich heraus, dass nach 14 Interviews ausreichende Ergebnisse und eine multiperspektivische Betrachtung erzielt wurde. Die größten Bereiche mit jeweils sechs Interviews waren die öffentlichen Verwaltungsbereiche und Wissenschaft und Forschung, was auf den Forschungsfokus der Masterarbeit zurückzuföhren ist. Die Interviews wurden digital via Zoom<sup>1</sup>/MS Teams<sup>2</sup> als Video-Call-Formate durchgeführt. Die Interviews dauerten zwischen 30 und 90 Minuten. Die besprochenen Inhalte wurden direkt während dem Video-Call gemeinsam in der Whiteboard-Anwendung Miro<sup>3</sup> festgehalten, um die Antworten der Teilnehmenden zu erfassen. Zusätzlich nutze ich zum Erstellen

<sup>1</sup><https://zoom.us>

<sup>2</sup><https://teams.microsoft.com>

<sup>3</sup><https://miro.com>

schriftlicher Notizen die App Notability<sup>4</sup>, die Mitschriebe arbeitete ich im Nachgang in das zugehörige Miro Board ein. Die Interviews wurden nicht gefilmt oder aufgenommen, so dass diese nur mithilfe der Notizen in Miro nachvollzogen werden können.

In der zweiten Phase **Definieren** wurden die gesammelten Informationen aus den Interviews analysiert und zu Insights verdichtet. Die verdichteten Informationen bildeten so eine solide Grundlage für die anschließende Entwicklungsphase. Die Qualitative Inhaltsanalyse wurde mithilfe eines Affinity Diagram durchgeführt, um aus der Entdecken-Phase gesammelten Informationen klar definierte Insights und Themenfelder zu synthetisieren; z.B. durch thematische Zusammenhänge zwischen den einzelnen Aussagen. Die Daten ließen sich auf 66 Insights und acht Themenfelder komprimieren, was die Komplexität der Datenmenge reduzierte und die Übersichtlichkeit erhöhte. Diese Insights wurden entsprechend auf die Einhaltung von Gütekriterien wie Transparenz, Intersubjektivität und Reichweite geprüft, worauf ich im Kapitel 5 genauer eingehen werde.

In der dritten Phase **Entwickeln** wurden auf Basis der 66 Insights Use Cases entwickelt, aus denen später in der vierten Phase ein Use Case für den angewandten Teil meiner Masterarbeit zur prototypischen Umsetzung ausgewählt wurde. Zur lösungsorientierten Entwicklung von Use Cases wurde die Design-Methode How-Might-We-Fragen (=HMW) eingesetzt. HMW-Fragen bieten einen strukturierten Ansatz, um Brainstorming zu lenken und Stakeholder:innenbedürfnisse sicherzustellen. Insgesamt formulierte ich 22 HMW-Fragen und entwickelte basierend auf diesen 31 Use Cases.

In der vierten Phase **Liefern** wurde in einem Entscheidungsmeeting zusammen mit der Firma credium und Prof. Dr. Kratsch ein finaler Use Case ausgewählt. Als Entscheidungsgrundlage diente das Use Case Diagramm in Abbildung 13 (Kapitel 2 Abschnitt 2.2.3), womit die Zusammenhänge zwischen den Use Cases und Stakeholder:innen in Bezug auf ein KI-System visualisiert wurden. Dazu wurden die 31 Use Cases nochmals auf 14 Use Cases komprimiert, da sich die Use Cases inhaltlich aus den unterschiedlichen Themenfeldern überschnitten. Diese wurden anschließend den Stakeholder:innen aus den Bereichen Bauverwaltungsaamt, Bauausschuss, Stadt-/Gemeinderat und Architekturbüro zugeordnet, welche besonders stark im Baugenehmigungsprozess auf kommunaler Ebene involviert sind. Ebenfalls wurde die Firma credium als Stakeholder:in betrachtet, was ausschließlich einen informativen Zweck erfüllte.

Die Auswahl des Use Case wurde anhand der Kriterien Mehrwert, Komplexität und persönliches Weiterentwicklungsinteresse getroffen. Entsprechend sollte der Use Case für möglichst viele Stakeholder:innen einen Mehrwert bieten, als auch realistisch im Rahmen meiner Masterarbeit umsetzbar sein.

Im Anhang (Kapitel 6) stehen die Miro-Boards zu den *Interviews* und dem *Interview-Clustering* zur Einsicht zur Verfügung.

## 2.2 Ergebnisse

Im Folgenden werden die gewonnenen Erkenntnisse und Ergebnisse aus den Phasen des Double Diamond-Prozess thematisch zusammengefasst vorgestellt.

---

<sup>4</sup><https://notability.com>

## 2.2.1 Grundwissen Bauleitplanung

### 2.2.1.1 Einordnung der Bauleitplanung

Um B-Pläne verstehen zu können ist ein Blick in die Bauleitplanung nötig, welche in Abbildungen 3 aus der Vogelperspektive in das föderale System eingeordnet wurde und aus drei wesentlichen Säulen besteht: Raumordnung, *Baurecht* und *Verwaltung*.

Das Baurecht klärt WAS und die Raumplanung WO und WIE bebaut werden darf (Hascher 2024; Fina 2023). Hierzu geben Bund und Länder die rechtlichen Rahmenbedingungen vor und übertragen die Verwaltungsaufgaben der Bauleitplanung an die entsprechenden Kommunen, wodurch eine eigenverantwortliche Gestaltung der Stadtentwicklung ermöglicht wird (Geirhos 2024).

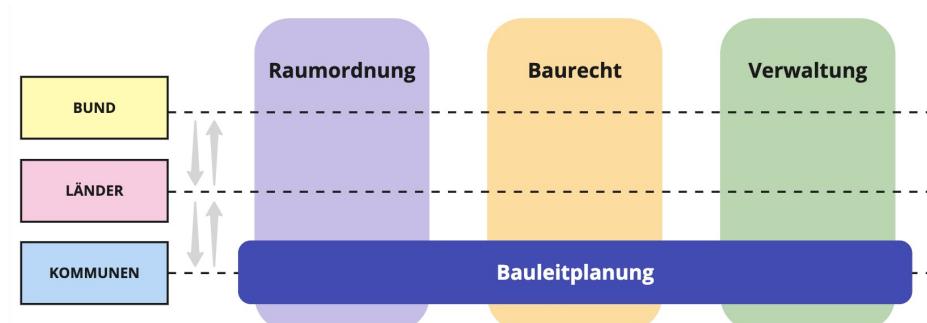


Abbildung 3: Einordnung der Bauleitplanung in das föderale System (Eigene Darstellung)

Das öffentliche Baurecht, siehe Abbildung 4, dient dem Interessenausgleich zwischen der geschützten Baufreiheit der Grundstückseigentümer:innen und dem häufig davon abweichendem Interesse der Allgemeinheit an einer möglichst sinnvollen Nutzung des nur begrenzt vorhandenen Baugeländes (vgl. Art. 14 Abs. 1 Grundgesetz (=GG) und Art. 2 Abs. 1 GG) und lässt sich grob in Bauplanungs- und Bauordnungsrecht unterteilen (Hascher 2024; Fina 2023). Bauordnungsrecht ist objektbezogenes Recht und enthält formelle Vorschriften für das bauaufsichtliche Verfahren und materielle Regelungen im Hinblick auf die Errichtung, Erhaltung, Änderung, Nutzung und den Abbruch baulicher Anlagen Hascher (2024); Fina (2023). Das Bauplanungsrecht, auch Städtebaurecht genannt, bezieht sich auf die Flächennutzung und betrachtet einzelne Bauvorhaben in einem größeren städtebaulichen Zusammenhang (Hascher 2024; Fina 2023). Das Zusammenspiel von Bauplanungs- und Bauordnungsrecht zeigt sich unter anderem am Beispiel von Abstandsregelungen von Gebäuden. So finden sich Regelungen zum Abstand von Gebäuden sowohl im Bauplanungsrecht als auch im Bauordnungsrecht durch das Abstandsflächenrecht (§ 5, 6 Landesbauordnung (=LBO)) (Fina 2023). Zu den von der Baurechtsbehörde gemäß § 58 Abs. 1 S. 1 LBO zu prüfenden Vorschriften, zur Erteilung einer Baugenehmigung, zählen die bauplanungsrechtlichen Vorschriften des Baugesetzbuchs (=BauGB) und der Baunutzungsverordnung (=BauNVO) sowie die bauordnungsrechtlichen Vorschriften der LBO (Fina 2023). B-Pläne fallen speziell in die Kategorie Bebauungsrecht, was im Abschnitt 2.2.1.2 genauer erklärt wird.

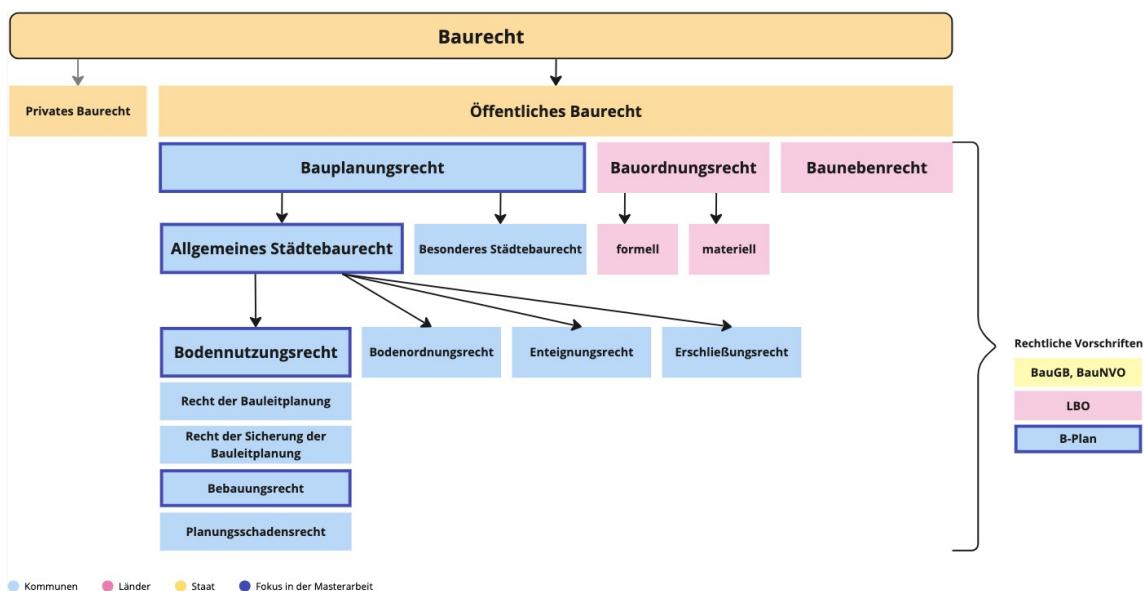


Abbildung 4: Einordnung der Bauleitplanung in das öffentliche Baurecht (Eigene Darstellung. Angelehnt an Hascher(2024); Fina(2023))

Kommunen haben im Rahmen des Bauplanungsrechts das Recht, Festsetzungen zu gestalten, welche flächenbezogene Anforderungen an ein Bauvorhaben regeln und eine geordnete städtebauliche Entwicklung sicherstellt (§§ 1–38 BauGB)(Hascher 2024). Zentrales Element hierfür ist die Bauleitplanung: „**Aufgabe der Bauleitplanung ist es, bauliche und sonstige Nutzung von Grundstücken einer Gemeinde vorzubereiten und zu leiten**“(§ 1 Abs. 1 BauGB)(Hascher 2024). In der Praxis erfolgt dies unter anderem durch die Entwicklung und Einhaltung der Bauleitpläne FNP und Bebauungsplan (B-Plan), die im Abschnitt 2.2.1.2 genauer erklärt werden. Gemäß § 1 Abs. 4 BauGB sind die Bauleitpläne den Zielen der Raumordnung bzw. überörtlichen Raumplanung anzupassen (Hascher 2024). Die Raumordnung, siehe Abbildung 5, umfasst die planmäßige Ordnung, Entwicklung und Sicherung von größeren Gebietseinheiten Deutschlands, im Hinblick auf eine ausgewogene Aufteilung von sozialen, wirtschaftlichen und ökologischen Funktionen vorhandener Landfläche. Diese erfolgt auf Bundesebene durch das Raumordnungsgesetz (=ROG), auf Landesebene durch das Landesplanungsgesetz (=LpG), Landesentwicklungsprogramme (=LEPROG), Landesentwicklungspläne (=LEP) und Regionalpläne (=RP).

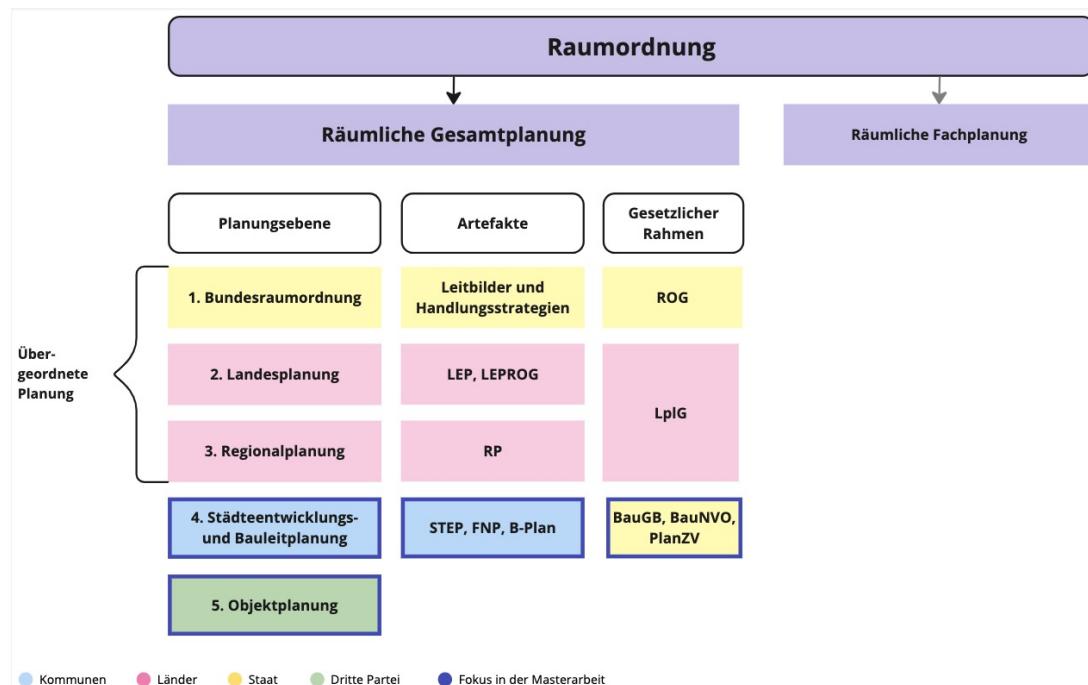


Abbildung 5: Einordnung der Bauleitplanung in die räumliche Gesamtplanung (Eigene Darstellung. Angelehnt an Hascher (2024); Wikipedia (2024e))

In der Abbildung 6 sind die öffentliche Verwaltung mit ihren beteiligen Personengruppen dargestellt, welche relevante Rollen in der Bauleitplanung einnehmen. Zusätzlich gibt es die Personengruppe dritte Parteien, welche Tätigkeiten übernehmen (z.B. Zeichnen eines B-Plans) oder durch Meinungen, Investitions- und Bauvorhaben wesentlich in die Prozesse der Bauleitplanung eingreifen (Kaiser 2024). Während der Forschung haben sich vor allem die Gruppen Planungs-/Architekturbüros, Bauverwaltungsamt und Stadt-/Gemeinderat als wichtige Stakeholder:innen für die bauplanungsrechtliche Prüfung eines Bauantrags herausgestellt.

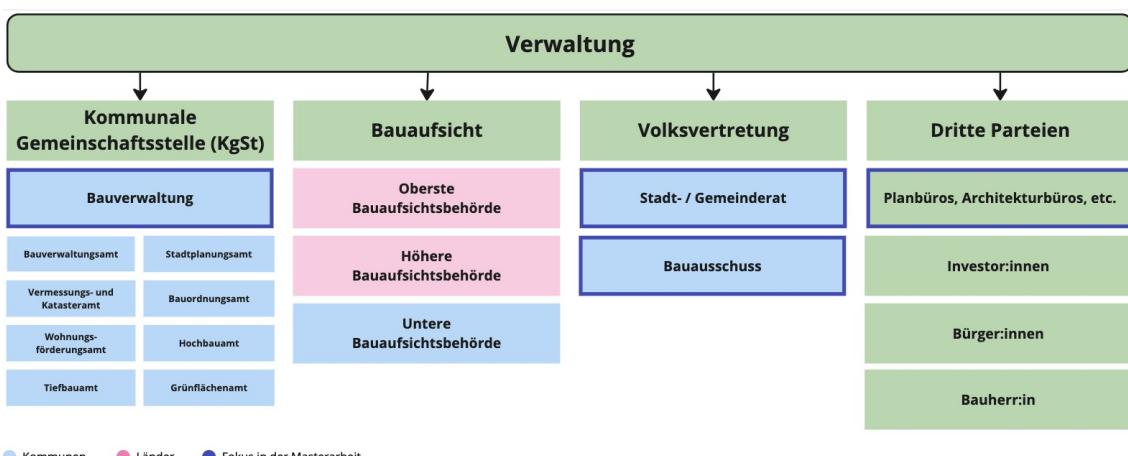


Abbildung 6: Stakeholder:innen in der Bauleitplanung (Eigene Darstellung. Angelehnt an Kaiser (2024))

### 2.2.1.2 Basiswissen über Bauleitpläne

Im Folgenden wird erklärt, welche Informationen ein B-Plan beinhaltet und auf welche Quellen sich diese Informationen referenzieren.

Die wichtigsten Werkzeuge der Bauleitplanung sind die sogenannten Bauleitpläne, welche sowohl zeichnerische als auch textliche Regelungen zur Flächennutzung eines Gemeindegebiets beinhalten und aus einer Planzeichnung und textlichen Begründungen bestehen (Fina 2023). Das Baugesetzbuch sieht zwei Arten von Bauleitplänen vor: der *Flächennutzungsplan (FNP)* und der *Bebauungsplan (B-Plan)* (Fina 2023).

Mit dem FNP formuliert die Kommune ein städtebauliches Entwicklungskonzept für das gesamte Gemeindegebiet für einen Zeithorizont von 10 bis 15 Jahren – man spricht dabei von einer vorbereitenden Bauleitplanung (Fina 2023). Dabei wird die zukünftig beabsichtigte Flächennutzung in Grundzügen dargestellt (Fina 2023). Die Inhalte von FNP werden als Darstellungen bezeichnet und haben den Charakter verwaltungsinterner Regelungen (Fina 2023). Zum Beispiel werden einem Teilgebiet eine oder mehrere Kategorien zugeordnet, wie z.B. Bauflächen, Art der Baugebiete, Verkehrsflächen, Grünflächen, Wasserflächen oder Flächen für die Landwirtschaft (vgl. § 5 Abs. 2 a BauGB) (Fina 2023). Wichtig ist dabei, dass die Darstellungen keine abschließende Regelungen zur Genehmigung von Bauvorhaben darstellen (Fina 2023). Abbildung 7 zeigt einen Ausschnitt eines FNP der Stadt München (Fina 2023).

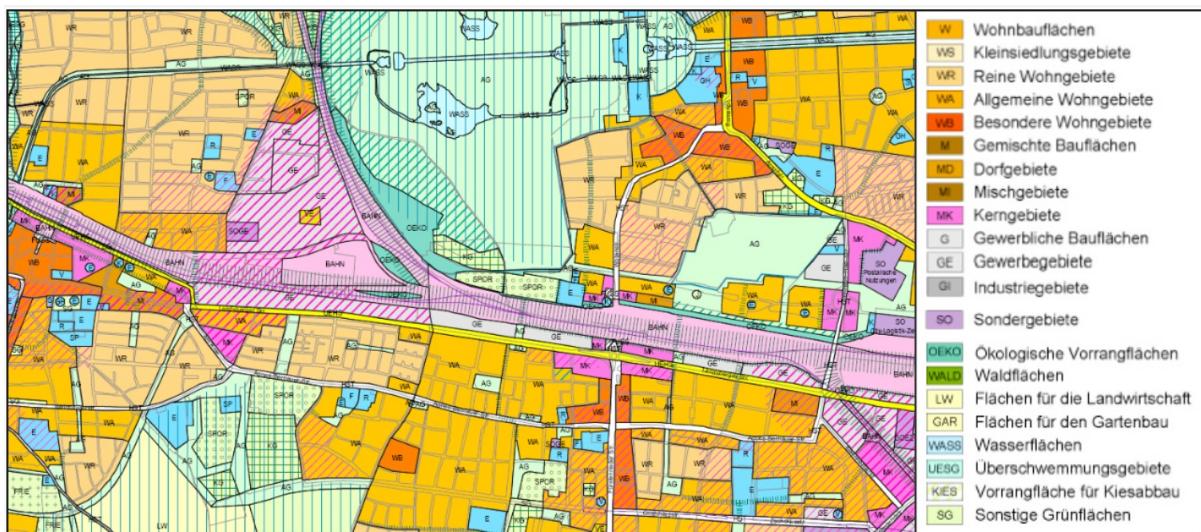


Abbildung 7: Ausschnitt eines FNP der Stadt München, Fina (2023)

Der B-Plan ist grundsätzlich aus dem FNP zu entwickeln und darf den Darstellungen eines FNP nicht widersprechen (Hascher 2024). Zusätzlich werden im B-Plan Festsetzungen zur baulichen Nutzung bzgl. eines Teilgebiets festgelegt (Hascher 2024). Dieser regelt, ob und wie ein Grundstück bebaut werden darf (vgl. § 10 Abs. 1 BauGB), weshalb von einer verbindlichen Planung gesprochen wird, dessen Geltungsdauer nicht begrenzt ist (Hascher 2024). Wichtig dabei zu beachten ist, dass die jeweilige Baugesetzesfassung zum Zeitpunkt der Veröffentlichung des B-Plan gilt – das sogenannte statische Baurecht

(Hascher 2024).

Es wird zwischen zwei Arten von B-Plänen unterschieden: Einfacher und Qualifizierter B-Plan (Fina 2023). Die Mindestinhalte eines qualifizierten B-Plans sind in § 30 Abs. 1 BauGB definiert und in der BauNVO jeweils im Detail spezifiziert (Fina 2023):

1. **Art der baulichen Nutzung** legt fest, wie das Grundstück benutzt werden darf, z.B. Wohngebiet (WA) oder Gewerbegebiet (GE)
2. **Maß der baulichen Nutzung** regelt im weitestgehenden Sinn das mögliche Bauvolumen auf einem Grundstück, wie z.B. Grundflächenzahl (GRZ) oder Geschossflächenzahl (GFZ)
3. **Überbaubare Grundstücksfläche** definiert auf welchem Grundstücksteil ein Gebäude errichtet werden darf, wie z.B. durch Baugrenzen (siehe blaue Linien im B-Plan)
4. **Bauweise** regelt das Verhältnis eines Gebäudes zu den seitlichen Grundstücksgrenzen und das sich daraus ergebende Straßen- bzw. Quartierbild, wie z.B. Einzelhäuser (offen) oder Blockbebauungen (geschlossen)

Diese Kerninformationen zu Gebäuden werden in einer sogenannten Nutzungsschablone (siehe Tabellenerklärung in Abbildung 8) in der Planzeichnung zusammengefasst (Fina 2023). Weitere mögliche Festsetzungen sind durch § 9 BauGB festgelegt (Hascher 2024). Abbildung 8 zeigt einen Ausschnitt eines B-Plans inklusive einer Tabellenerklärung der Nutzungsschablone.

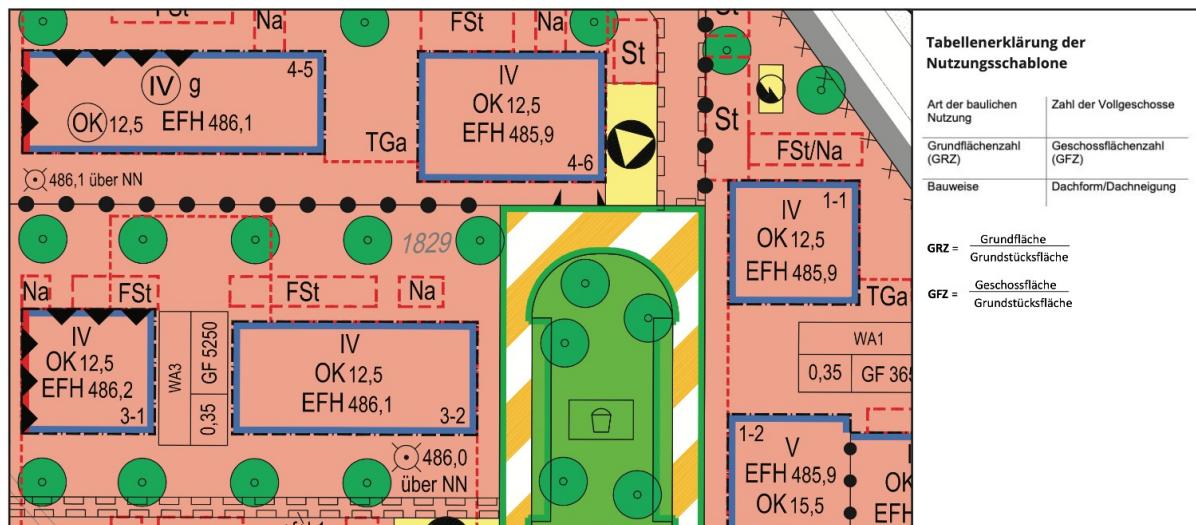


Abbildung 8: Ausschnitt eines B-Plan der Stadt Augsburg und Erläuterung der Nutzungs schablone (Fina 2023)

Zusätzlich zu den Festsetzungen der B-Pläne definiert § 9 BauGB, dass eine Begründung (z.B. Beschreibung und Ziele der Planung, Umweltbericht) und Kennzeichnungen (z.B. belastender Boden, besondere äußere Einwirkungen) beizulegen sind, dessen Informatio-

nen besonders relevant sind, um die Absichten und Rahmenbedingungen hinter den Festsetzungen verstehen zu können (Fina 2024). Zusätzlich müssen für die Planzeichnung genormte Planzeichen eingehalten werden; siehe Farbflächen und verwendete Symbole in Abbildung 8 (Fina 2023). Der einfache B-Plan unterscheidet sich von dem qualifizierten B-Plan ausschließlich dadurch, dass der einfache B-Plan die Mindestinhalte nicht einhalten muss (Kaufmann & Hascher 2024).

B-Pläne sind aufzustellen, sobald und soweit die städtebauliche Entwicklung und Ordnung dies erfordert (§ 1 Abs. 3 BauGB), was im Ermessen der Kommune liegt (Fina 2023). Entsprechend gibt es auch unbeplante Innen- und Außenbereiche ohne B-Plan in denen Bauvorhaben sich in die nähere Umgebung einfügen müssen, was im Rahmen der Masterarbeit nicht weiter betrachtet wird (Hascher 2024). Ein Überblick zur Rechtsnatur von Bauleitplänen veranschaulicht Abbildung 9. Darauf werden die wesentlichen Gemeindegebiete, Bauleitpläne und deren Zusammenwirken, sowie die B-Plantypen und deren Regelungsinhalte visuell veranschaulicht.

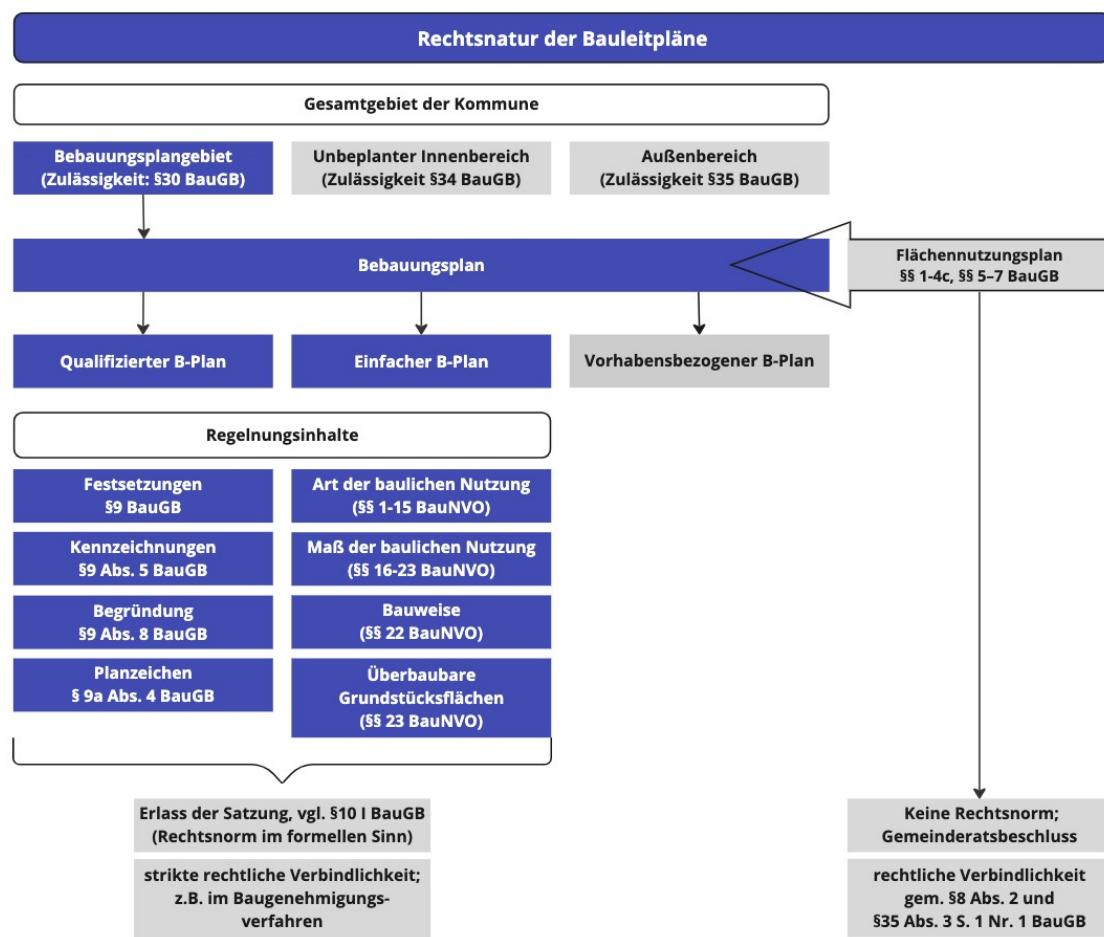


Abbildung 9: Rechtsnatur von Bauleitplänen im beplanten Gebiet (Eigene Darstellung. Angelehnt an Hascher (2024))

### 2.2.1.3 Grundlagen des Bauplanungs- und Genehmigungsprozess

Die wichtigsten drei Prozesse, die in der Bauleitplanung eng mit B-Pläne verbunden sind, sind der Bauplanungsprozess, Baugenehmigungsprozess und das Aufstellungsverfahren eines B-Plan. Das Aufstellungsverfahren konnte im Rahmen meiner Masterarbeit aus zeitlichen Gründen nicht betrachtet werden. Im Folgenden werden die relevantesten Arbeitsschritte im Bauplanungs- und Genehmigungsprozess vorgestellt. Durch den Einblick in die Prozesse konnten wichtige Arbeitsschritte kennengelernt und nachvollzogen, wie auch Herausforderungen innerhalb dieser Prozesse, sowie beteiligte Stakeholder:innen identifiziert werden.

#### Bauplanungsprozess

Der Bauplanungsprozess startet in der Regel damit, dass ein:e Bauherr:in ein Architekturbüro zu einem Bauvorhaben beauftragt und richtet sich grob nach den Leistungsphasen der Honorarordnung für Architekt:innen und Ingenieur:innen (=HOAI) Bruch & Bruch (2024). Die Leistungsphasen, siehe Abbildung 10, beschreiben die einzelnen Planungsabschnitte der Gesamtleistung von Architekt:innen und Ingenieur:innen bei der Planung und Realisierung von Bauvorhaben(Wikipedia 2023f). Zusammen mit Bruch & Bruch(2024) wurden drei von neun Leistungsphasen der HOAI ausgewählt und näher beleuchtet, die im engen Zusammenhang mit B-Plänen und Genehmigungsprozessen stehen. Diese sind Vorplanung, Entwurfsplanung und Genehmigungsplanung, welche in Abbildung 10 farblich hervorgehoben sind.

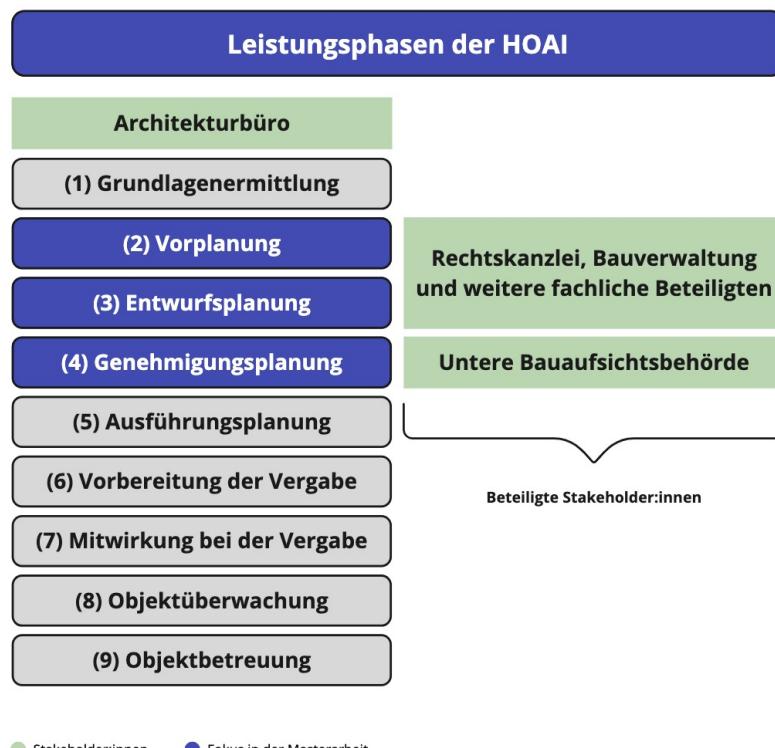


Abbildung 10: Bauplanungsprozess anhand der HOAI Leistungsphasen (Eigene Darstellung. Angelehnt an Architektenkammer (2021))

Nachdem Anforderungen und Rahmenbedingungen mit dem:der Bauherr:in geklärt sind, entsteht in der Vorplanungsphase ein erster konkreter Entwurf des Bauvorhabens, das sogenannte Planungskonzept (Bruch & Bruch 2024). Dieses umfasst unter anderem eine grundlegende Gestaltung und Funktion des Gebäudes, was in der Vorverhandlung mit dem Bauverwaltungsamt hinsichtlich der Genehmigungsfähigkeit besprochen wird (Bruch & Bruch 2024). Das ist nötig, da z.B. veraltete B-Pläne modernen Bauvorhaben entgegenstehen oder fehlende Informationen die Planung angreifbar machen Bruch & Bruch (2024); Maile(2024). In diesem Schritt wird ebenfalls Rechtsberatung hinzugezogen, um baurechtlich möglichst viel Klarheit zu schaffen, da sowohl das Baurecht als auch die Festsetzungen im B-Plan Auslegungen bedarfen (Bruch & Bruch 2024; Maile 2024).

In der Entwurfsphase erfolgen weitere Ausarbeitungen des Vorentwurfs, sodass ein vollständiges Planungskonzept entsteht (Bruch & Bruch 2024). Hierbei fließen alle städtebaulichen, gestalterischen, sozialen und rechtlichen Aspekte in das Planungskonzept ein (Bruch & Bruch 2024). Zusätzlich wird auch in dieser Phase der aktuelle Stand des Planungskonzepts mit dem Bauverwaltungsamt abgestimmt und Rechtsberatung hinzugezogen (Bruch & Bruch 2024).

In der Genehmigungsphase bereiten Architekt:innen den Entwurf zu genehmigungsfähigen Plänen (z.B. Bauzeichnungen und Lageplan) auf, die den Anforderungen des Bauverwaltungsamts entsprechen, und reichen den Antrag mit allen benötigten Unterlagen bei der Untere Bauaufsichtsbehörde ein (Bruch & Bruch 2024). Bei Rückfragen oder einer Ablehnung von der Behörde kümmert sich der/die Architekt:in um nötige Anpassungen des Plankonzepts oder reicht entsprechend Befreiungs- bzw. Ausnahmeanträge nach (Bruch & Bruch 2024). Dieser Vorgang wiederholt sich, bis eine Baugenehmigung vorliegt (Bruch & Bruch 2024). Laut Bruch & Bruch (2024) könnte die Genehmigung eines Bauvorhabens theoretisch schnell über die Bühne gehen, was jedoch durch regelmäßig benötigte Ausnahme- und Befreiungsanträge zu langwierigen Prozessen mutieren kann.

## **Baugenehmigungsprozess**

In der Regel startet ein Genehmigungsprozess mit der Einreichung des Bauantrags in der Untere Bauaufsichtsbehörde. Hierzu stellen die meisten Bundesländer bereits Online-Plattformen zur Verfügung, um die Dokumente digital im PDF Format hochzuladen (Kaiser 2024; Kaufmann & Hascher 2024). Anschließend werden alle beteiligten Behörden über den Eingang des Bauantrags informiert und können über die einheitliche Plattform auf die eingereichten Unterlagen zugreifen (Kaiser 2024).

Eine Kommune führt je nach Bauvorhaben unterschiedliche bauplanungsrechtliche Prüfungsverfahren durch. Diese sind das **Kenntnisgabe-, reguläre** und **vereinfachte** Prüfungsverfahren (Kaiser 2024; Kaufmann & Hascher 2024). Das reguläre Prüfungsverfahren ist das ausführlichste Programm und beinhaltet die meisten Arbeitsschritte aufseiten der Behörden(Kaiser 2024). Das Kenntnisgabe- und vereinfachte Prüfungsverfahren bestehen aus weniger Arbeitsschritten, wodurch die Bearbeitung eines Bauvorhabens schneller durchlaufen werden kann (Kaiser 2024; Kaufmann & Hascher 2024). Im Rahmen der Masterarbeit wurde ein Bauvorhaben im beplanten Innenbereich betrachtet, für welches ein B-Plan vorliegt und das reguläre Prüfungsverfahren durchläuft; siehe Abbildung 11.

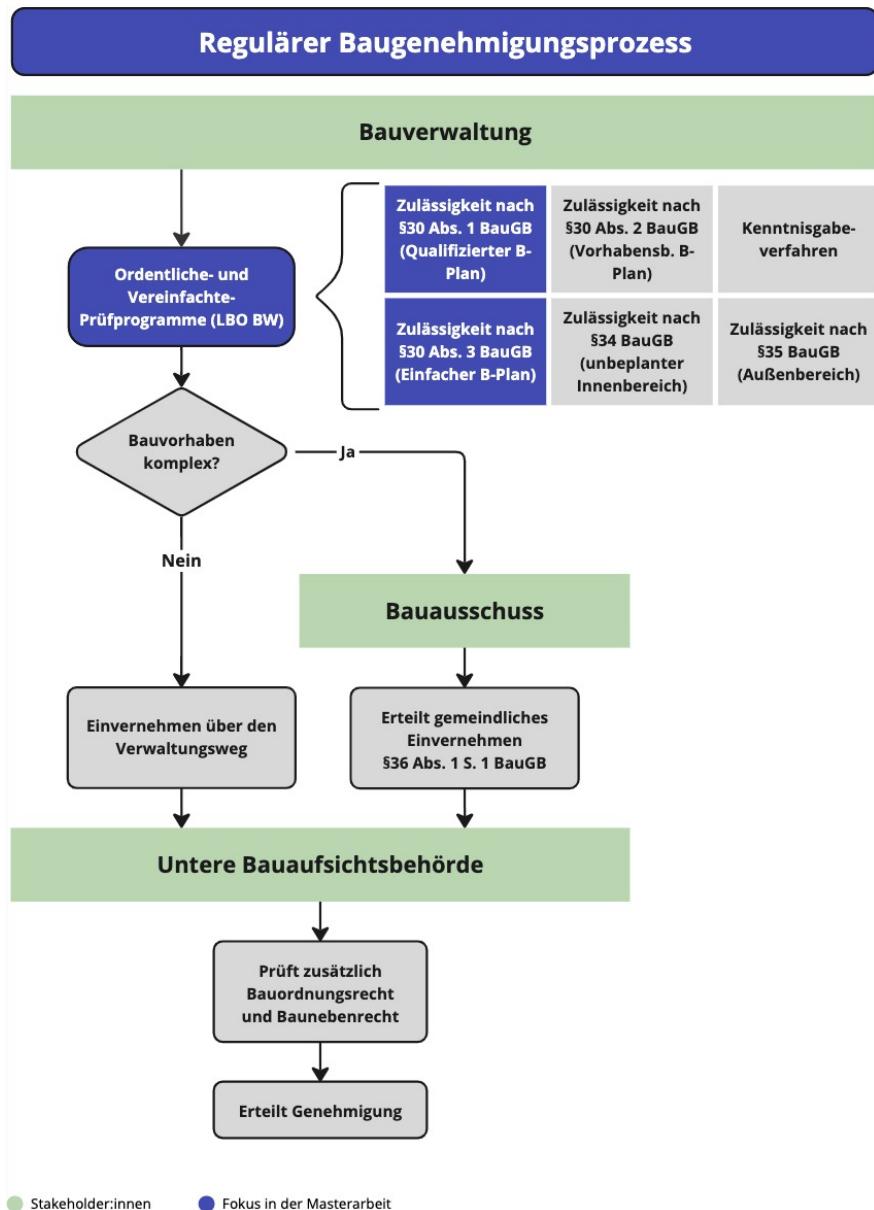


Abbildung 11: Überblick eines regulären Baugenehmigungsprozesses (Eigene Darstellung. Angelehnt an Hascher (2024))

Das Bauverwaltungsamt übernimmt federführend die rechtliche Prüfung des Bauvorhabens, die je nach Bauvorhaben unterschiedlich ausgeprägt ist (Kaiser 2024). In diesem Prüfungsverfahren ist die Zulässigkeit nach § 30 Abs. 1 oder 3 BauGB zu prüfen, da ein qualifizierter oder einfacher B-Plan für das Teilgebiet, in der das Bauvorhaben geplant ist, vorliegt. Genaueres zum Prüfungsprogramm wird in Abschnitt 2.2.1.3 erläutert. Im nächsten Schritt schätzt das Bauverwaltungsamt die Komplexität des Bauvorhabens nach eigenem Ermessen ein (Kaiser 2024). Handelt es sich z.B. um einen Anbau einer Garage oder einer Gartenhütte, wird das gemeindliche Einvernehmen über den Verwaltungsweg an die Untere Bauaufsichtsbehörde weitergeleitet (Kaiser 2024). Handelt es sich jedoch um ein komplexes Bauvorhaben, wie z.B. einem Neubau, wird eine ausführliche Sitzungsvorlage für den Bauausschuss erstellt (Kaiser 2024). Die Sitzungsvorlage beinhaltet neben der Prüfungsergebnisse unter anderem den Sachverhalt, Begründungen und Emp-

fehlungen zu Ermessensentscheidungen und einen Beschlussvorschlag (Kaiser 2024). Im Bauausschusses wird anhand der Sitzungsvorlage das Bauvorhaben aus mehreren Perspektiven ausführlich beleuchtet. Hierbei wird unter anderem die Einhaltung des Bauvorhabens nach dem Ermessen des Bauausschusses entschieden und die Ergebnisse in einem sogenannten Sitzungsdokument festgehalten (Kaiser 2024). Zusätzlich ist die Sitzung öffentlich für Bürger:innen zugänglich, sodass die Entscheidung bzgl. des Einvernehmens möglichst transparent und nachvollziehbar ist (Kaiser 2024). Das Sitzungsdokument wird anschließend an die Untere Bauaufsichtsbehörde weitergeleitet, die das Bauvorhaben zusätzlich auf Einhaltung des Bauordnungs- und Baubebenrecht prüft und abschließend die offizielle Baugenehmigung erteilt (Kaiser 2024).

### **Prüfungsprogramm zur Zulässigkeit nach §30 BauGB**

Im Folgenden werden die Prüfungsschritte des Prüfungsprogramms der Zulässigkeit §30 Abs. 1 BauGB vorgestellt. Der Einblick hilft zu verstehen, welche Informationen aus dem B-Plan für die Kernprüfung besonders relevant sind. Zusätzlich können aus den einzelnen Prüfungsschritten später konkrete Arbeitsschritte für das geplante prototypische Kl-System abgeleitet werden.

Im ersten Schritt wird vom Bauverwaltungsamt geprüft, ob das Grundstück erschlossen ist, was die Grundvoraussetzung dafür ist, ob auf dem Grundstück ein Gebäude errichtet werden darf (Kaufmann & Hascher 2024). Dazu zählen unter anderem ein Anschluss an das öffentliche Straßen-, Wege- und Versorgungsnetz wie z.B. ein Wasser- und Elektrizitätsanschluss (Kaufmann & Hascher 2024).

Sind diese Voraussetzungen erfüllt, werden im zweiten Schritt die Festsetzungen aus dem B-Plan mit dem Bauvorhaben verglichen, um sicherzustellen, dass das Bauvorhaben die Festsetzungen aus dem B-Plan einhält (Kaufmann & Hascher 2024). Im Interview mit Kaufmann & Hascher (2024) konnte ich anhand eines bereits abgeschlossenen Bauvorhabens aus dem Jahr 2011 in der Stadt Laichingen die entsprechenden Prüfungsschritte am B-Plan L04 praktisch nachvollziehen (Bauplan und B-Plan stehen im Anhang (Kapitel 6) zur Einsicht bereit). Als Hilfsmittel diente hierzu eine Tabelle, welche alle zu prüfenden Festsetzungen beinhaltete (Kaufmann & Hascher 2024); siehe Tabelle 1.

Festsetzungen	Bauantrag	Bebauungsplan	entspricht / widerspricht	Befreiung / Ausnahme	Erteilt
Art der baulichen Nutzungen	<ul style="list-style-type: none"> <li>Bauantrag und Bauzeichnung</li> <li>Wohnhaus</li> </ul>	<ul style="list-style-type: none"> <li>Nr. 1.1.1 i.V. m. § 1 Abs. 3 BauNVO Allgemeines Wohngebiet (WA) im zeichnerischen Teil des B-Plan</li> <li>§ 4 Abs. 2 Nr. 1 BauNVO Wohngebäude allgemein zulässig</li> </ul>	<input checked="" type="checkbox"/>		
Maß der baulichen Nutzungen	<ul style="list-style-type: none"> <li>Lageplan (zeichnerischer und schriftlicher Teil)</li> <li>GRZ 0,17 / GFZ 0,35</li> <li>Zahl der Vollgeschosse II</li> </ul>	<ul style="list-style-type: none"> <li>Nr. 1.1.1 i.V.m. § 1 Abs. 3 BauNVO GRZ 0,4 / GFZ 0,8</li> <li>Nr. 1.1.3 i.V.m. zeichnerischem Teil Vollgeschosse I</li> </ul>	<input checked="" type="checkbox"/> <input type="checkbox"/>	Befreiung §§ 36, 31 Abs. 2 BauGB	Entscheidung LRA 18.05.2011
Bauweise	<ul style="list-style-type: none"> <li>Lageplan (zeichnerischer Teil)</li> <li>Mit seitlichen Abstandsfächern</li> </ul>	<ul style="list-style-type: none"> <li>Nr. 1.2 i.V.m. zeichnerischem Teil</li> <li>Offene Bauweise</li> </ul>	<input checked="" type="checkbox"/>		
Überbaubare Grundstücksflächen	<ul style="list-style-type: none"> <li>Lageplan (zeichnerischer Teil)</li> <li>Gebäude liegt im Osten und Norden außerhalb der Baugrenzen</li> </ul>	<ul style="list-style-type: none"> <li>Nr. 1.2 i.V.m. zeichnerischem Teil</li> <li>Überschreitung Baugrenzen</li> </ul>	<input type="checkbox"/>	Befreiung §§ 36, 31 Abs. 2 BauGB	Entscheidung LRA 18.05.2011
Stellung der baulichen Anlage	<ul style="list-style-type: none"> <li>Lageplan (zeichnerischer Teil)</li> <li>West-Ost Richtung</li> </ul>	<ul style="list-style-type: none"> <li>Nr. 1.3 i.V.m. zeichnerischem Teil</li> <li>West-Ost Richtung</li> </ul>	<input checked="" type="checkbox"/>		
Lage der Garagen	<ul style="list-style-type: none"> <li>Lageplan (zeichnerischer Teil)</li> <li>Garagen sollen nicht im Haus und auch nicht an den dafür vorgesehenen Stellen errichtet werden</li> </ul>	<ul style="list-style-type: none"> <li>Nr. 1.4 i.V.m. zeichnerischem Teil</li> <li>Garagen sind im Haus oder an den dafür ausgewiesenen Stellen mit ebenem Dach zu erstellen</li> </ul>	<input type="checkbox"/>	Befreiung §§ 36, 31 Abs. 2 BauGB	Entscheidung LRA 18.05.2011
Höhenlage der Gebäude	<ul style="list-style-type: none"> <li>Lageplan (zeichnerischer Teil)</li> <li>EFH: 774,05 m üNN</li> </ul>	<ul style="list-style-type: none"> <li>Nr. 1.5 i.V.m. Beilageplan und zeichnerischem Teil</li> </ul>	<input checked="" type="checkbox"/>		
Freizuhaltenden Sichtfelder	<ul style="list-style-type: none"> <li>Lageplan (zeichnerischer Teil)</li> <li>Baugrundstück nicht betroffen</li> </ul>	<ul style="list-style-type: none"> <li>Nr. 1.6 i.V.m. zeichnerischem Teil</li> </ul>	<input checked="" type="checkbox"/>		

Tabelle 1: Prüfschema zur bauplanungsrechtlichen Prüfung der Zulässigkeit nach §30 Abs. 1 und 3 BauGB, Qualifizierter B-Plan (Eigene Darstellung. Angelehnt an Kaufmann & Hascher (2024))

Schritt für Schritt wurden in den Spalten **Bauantrag** und **Bebauungsplan** die jeweiligen Daten bzgl. der jeweiligen **Festsetzungen** der aus den vorliegenden Dokumenten, B-Plan und Bauantrag, übertragen (Kaufmann & Hascher 2024). Im nächsten Schritt wurden diese Werte miteinander verglichen und in der Spalte **widerspricht (×) entspricht (✓)** jeweils ein Häkchen, falls das Bauvorhaben die Festsetzung erfüllt oder ein Kreuz bei einem Widerspruch eingetragen (Kaufmann & Hascher 2024). Bei einem Widerspruch wurde entsprechend geprüft, ob bereits eine Ausnahme oder Befreiung vorliegt und ob diese bereits erteilt wurde (Kaufmann & Hascher 2024). Die Information wurden in den Spalten **Ausnahme**, **Befreiung** und **Erteilt** dokumentiert (Kaufmann & Hascher 2024). Zum Beispiel verdeutlicht der Eintrag zu **Zahl der Vollgeschosse** in Kategorie **Maß der baulichen Nutzung**, dass im Bauvorhaben zwei Vollgeschosse geplant wurden und im B-Plan nur ein Vollgeschoss erlaubt war, was zu einem Widerspruch führt (×). Diesbezüglich wurde eine Befreiung eingereicht, welche am 18.05.2011 erteilt wurde, sodass trotz Widerspruch zwei Vollgeschosse zugelassen werden können.

Zusätzlich zu den Festsetzungen aus dem B-Plan werden im nächsten Schritt noch weitere Kriterien geprüft wie z.B. das Rücksichtnahmegerbot und die Gebietsverträglichkeit, dessen Einhaltung in der Regel durch die Analyse zum „**Einfügen des Gebäudes in die nähere Umgebung vgl. § 34 BauGB**“ geprüft wird (Kaufmann & Hascher 2024). Werden alle Festsetzungen eingehalten und im Falle von Regelverstößen alle Befreiungen/Ausnahmen erteilt, kann einem einheitlichen Einvernehmen der Kommune in der Regel nichts mehr entgegenstehen (Kaufmann & Hascher 2024).

## 2.2.2 Insights und Leitfragen

In diesem Abschnitt werden die zentralen Insights aus der Analyse der Expert:innen-Interviews präsentiert. Hierzu wurden die Insights thematisch zusammengefasst und einer Hauptkategorie (*Strategie, Prozesse und Daten*) zugeordnet. Zu jedem Themenschwerpunkt wurde anschließend eine Leitfrage formuliert.

### 2.2.2.1 Strategien

Die Kategorie Strategien stellt zwei wesentliche Einflussfaktoren für die Weiterentwicklung von B-Pläne vor: *Digitale Transformation* und *Nachhaltige Stadtentwicklung*.

#### Digitale Transformation

Bund und Länder arbeiten aktuell an Standards, Softwarelösungen und gesetzlichen Rahmenbedingungen für eine nachhaltige digitale Transformation der Bauleitplanung Maile (2024). Laut Maile (2024) gibt es erhebliches Potenzial für Digitalisierung im Bereich der Bauleitplanung, insbesondere durch die Integration von BIM, XPlanung und ALKIS in bestehende CAD Programme durch die Nutzung des Standards XBau. Dieser gibt vor, dass in Zukunft B-Pläne im XPlanGML Format (=XPlanung) vorliegen, welche alle Informationen eines qualitativen B-Plan (textlich und zeichnerisch) in strukturierter maschinenlesbarer Form beinhalten (Zwick & Wagner 2024). Der B-Plan und dazugehörigen Geodaten (z.B. ALKIS) werden durch den XBau Standard direkt in die IT-Systeme der Architekt:innen integriert. In Ergänzung mit BIM wird es zukünftig möglich sein, eine detailreiche und raumbezogene Prüfung der Bauzeichnung in 3D durchführen zu können (Krause 2022b). Abbildung 12 zeigt, wie dies innerhalb eines CAD Programms aussehen könnte. Zusätzlich soll das Bauverwaltungsamt durch den Zugriff auf digitale Planungsdaten, Entwürfe oder Bauanträge zukünftig automatisiert durch Algorithmen prüfen lassen können (Krause 2022b).

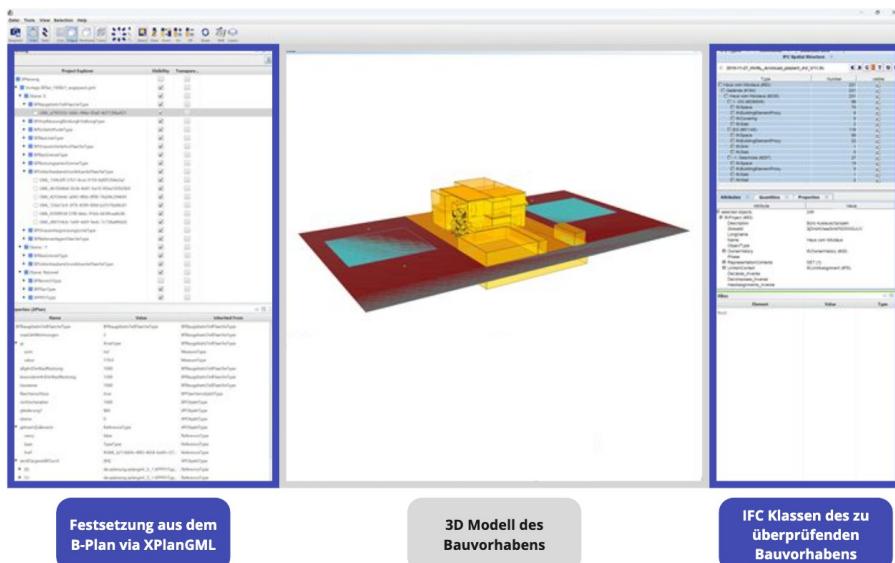


Abbildung 12: Modellbasierter Abgleich der planungsrechtlichen Festsetzungen im CAD Programm (Eigene Darstellung. Angelehnt an Krause (2022b))

OZG und INSPIRE bilden den rechtlichen Rahmen, welche Bund, Länder und Kommunen zur Digitalisierung von Dienstleistungen, Transformation von Bestandsdaten und Prozesse, sowie zur Offenlegung von unter anderem geobasierten Daten verpflichten (Wikipedia 2024r; GDI-DE 2024). Beispielsweise ist XPlanung seit 2017, BIM eingeschränkt seit 2020 und ALKIS seit 2018/2019 verbindlich anzuwenden (Krause 2022c; Wikipedia 2024a,g).

Die digitale Transformation ist eine große Herausforderung für Kommunen und zeigt sich unter anderem darin, dass die meisten Kommunen noch am Anfang der Transformation stehen Maile (2024); Zwick & Wagner (2024). Nur wenige Kommunen, wie z.B. die Stadt Hamburg, sind bereits 100% XPlan-konform und arbeiten an der Digitalisierung von Prozessen in der Bauleitplanung, wie z.B. das Projekt DiPlanung<sup>5</sup> (Maile 2024; Zwick & Wagner 2024). Zusätzlich nimmt laut Künster(2024) die Einarbeitung in Software-Programme aktuell mehr Zeit in Anspruch als die eigentliche Arbeit im Planungsbüro, was unter anderem an häufigen Updates und der schnellen Entwicklung neuer Softwarelösungen liegt. Der Paradigmenwechsel von visuellen Arbeitsmethoden hin zu Datenstrukturen führt dazu, dass bestehendes Fachpersonal im Bereich Bauleitplanung entsprechende Weiterbildungen und Schulungen in den Bereichen IT (z.B. Programme und Datenstandards) und allgemeiner Medienkompetenz benötigen (Zwick & Wagner 2024; Künster 2024). Moderne Studiengänge wie z.B. Digitaler Baumeister an der THA bilden entsprechend zukünftiges Fachpersonal bereits mit den nötigen IT-Fähigkeiten im Studium aus (Maile 2024).

Aus den Zusammenhängen lässt sich schließen, dass offene Standards und Konzepte für eine digitale Bauleitplanung heute bereits zur Verfügung stehen und zusätzliche rechtliche Rahmenbedingungen sicherstellen, dass diese in Zukunft umgesetzt bzw. genutzt werden. Damit die digitale Transformation jedoch gelingen kann, müssen alle beteiligten Personen an einem Strang ziehen, was zur folgenden Leitfrage führt: „Wie können Länder und Kommunen unterstützt werden um die digitale Transformation zu meistern?“.

## Nachhaltige Stadtentwicklung

Die Charta von Athen und die Charta von Leipzig sind zwei wichtige Leitgedanken in der Stadtplanung, die jeweils zu unterschiedlichen Zeiten entstanden sind und unterschiedliche städtebauliche Philosophien widerspiegeln Fina (2024). In den ersten Jahrzehnten des 20. Jahrhunderts verschlechterten sich die Lebensbedingungen in großen Städten erheblich durch die Industrialisierung (Wikipedia 2024i). Die Athen-Charta versuchte das Problem zu lösen, indem die Städte funktional in Zonen aufgeteilt und durch Grünflächen sowie Verkehrsachsen klar strukturiert wurden, um Wohn-, Arbeits- und Erholungsbereiche effektiv zu trennen (Wikipedia 2024i). Die Charta von Leipzig, verabschiedet 2007 und 2020, beruft sich auf die drei bekannten Dimensionen der Nachhaltigkeit und spricht sich für eine sozial gerechte, ökologisch stabile und wirtschaftlich prosperierende Stadt aus und steht im Kontext der Globalisierung, der aktuellen sozialen Herausforderungen und dem Klimawandel (Wieder et al. 2021). Für eine gemeinwohlorientierte Stadtentwicklungspolitik ist die Stärkung der Handlungsfähigkeit der Städte durch unter anderem einer aktiven strategischen Bodenpolitik und Flächennutzungsplanung eine wichtige Basis für das Erreichen dieser Ziele (Wieder et al. 2021).

Dies deutet darauf hin, dass der Baubestand in den Städten stark funktional durch die Charta von Athen geprägt worden ist und nun nach dem Leitbild der Charta von Leipzig transformiert werden soll. Die Baufläche in Deutschland ist bereits verbaut und kann nur

---

<sup>5</sup><http://diplanung.de>

langsam umgestaltet werden – „*Die grüne Wiese ist vorbei*“ (Maile 2024). Bestandsgebäude stellen somit gewissen Pfadabhängigkeiten dar, die zu Nutzungskonflikte führen, welche einer modernen Entwicklung im Weg stehen (Maile 2024). Daraus lässt sich schließen, dass die Bestandsdaten von B-Pläne vor dem Jahr 2007 in einigen Aspekten den heutigen modernen Bauvorhaben entgegenstehen. Die Architekt:innen Bruch & Bruch (2024) argumentieren, dass die Inhalte von veralteten Bauleitplänen der Kern des Problems in der Bauplanung und Baugenehmigung sind. Damit Genehmigungsprozesse in Zukunft schneller ablaufen können, müssen Ausnahme- und Befreiungsanträge reduziert werden, die heutzutage jedoch die Regel sind. Laut Maile (2024) sind alte B-Pläne jedoch per se nicht schlecht, viele B-Pläne sind auch aus heutigen Gesichtspunkten noch ausreichend und erfüllen ihren Zweck. Des Weiteren spricht Fina (2024) davon, dass das Potenzial von B-Pläne im Bereich nachhaltiges Bauen aktuell noch nicht voll ausgeschöpft wird.

Daraus kann die Schlussfolgerung gezogen werden, dass die Analyse von B-Pläne auf nachhaltige Indikatoren als Zeigerwirkung für eine nachhaltige Stadtentwicklung und Weiterentwicklung von Bestandsplänen genutzt werden könnte. Dies führt zu folgender Leitfrage: „*Welche nachhaltigkeites Indikatoren gibt es in B-Pläne und wie können diese extrahiert werden?*“.

### **2.2.2.2 Prozesse**

Die Kategorie Prozesse beschreibt den Einfluss von B-Pläne in gängige Abläufe der Bauleitplanung, wie z.B. bei der Entscheidungsfindung oder Bürger:innenbeteiligung in Entwurfs- und Genehmigungsprozessen.

#### **(Teil-)Automatisierung**

Wie bereits in Abschnitt 2.2.2.1 erläutert, befindet sich die digitale Transformation der Bauleitplanung auf kommunaler Ebene überwiegend noch am Anfang. Laut Kaufmann & Hascher (2024) werden in der Stadt Laichingen seit drei Jahren neue B-Pläne im Format XPlanGML erstellt oder bei Änderungen alter B-Pläne diese zu XPlanGML transformiert. Wann und wie der Großteil der bestehenden B-Pläne XPlanGML-konform sein wird, ist noch nicht klar. Der Transformationsprozess wird auf mehrere Jahre geschätzt (Kaufmann & Hascher 2024).

Ein weiteres Beispiel ist der digitale Bauantrag, welcher die Einreichung der Unterlagen eines Bauvorhabens vereinfacht und diese über eine Schnittstelle für alle Stakeholder:innen zur Verfügung stellt (Kaiser 2024). Laut Kaiser (2024) ist das bereits eine positive Verbesserung, was bestehende analoge Verwaltungsprozesse stark vereinfacht (Kaiser 2024). Laut Maile (2024) und Kaufmann & Hascher (2024) reicht die Lösung jedoch noch nicht weit genug, da die Daten weiterhin unstrukturiert als PDF Datei vorliegen und damit die Vorteile von einer maschinellen Automatisierung nicht ausgeschöpft werden können. Entsprechend ist der digitale Workflow nach dem Einreichen der Unterlagen des Bauantrags schon wieder vorbei und muss manuell von den Behörden weiterverarbeitet werden Kaufmann & Hascher (2024).

Die aktuelle Situation weist darauf hin, dass erstmal eine strukturierte Datenbasis geschaffen werden muss, bis Algorithmen im großen Maßstab Arbeitsschritte abnehmen können. Dennoch gibt es bereits Softwarelösungen rund um die Verwaltung, die schon jetzt analoge Prozesse unterstützen und ihren Mehrwert leisten, was zur folgenden Leit-

frage führt: „*Wie können manuelle Arbeitsschritte in Entwurfs- und Genehmigungsphasen effizienter gestaltet werden?*“.

### **Effizienter Zugriff auf Daten**

B-Pläne haben viele Querverweise zu Paragraphen, Detailwissen oder Umgebungsfaktoren. Ebenfalls beinhalten B-Pläne implizite Festsetzungen, die nicht explizit aufgeführt werden müssen (=Festsetzungautomatik). Für ein ganzheitliches Bild und klare Rahmenbedingungen bzgl. der Festsetzungen müssen entsprechend alle Informationen über unterschiedlichste Quellen zusammengetragen werden, was viel Zeit in Anspruch nimmt (Kaiser 2024). Kaiser (2024) erklärt, dass Regelungen zu Feinheiten wie z.B. Dachaufbauten und Zwerchgiebel, wie auch unterschiedliche Berechnungsformeln zu Kniestöcken, unnötig viel Zeit in der Grundlagenforschung zur Prüfung von Bauvorhaben beansprucht.

Ähnlich aufwendig ist die Beschaffung aller nötigen Rahmenbedingungen in der Entwurfsphase von Bauvorhaben, was zu viel Austausch mit Behörden zu den Festsetzungen eines B-Plan führt Bruch & Bruch (2024).

Laut Schwindling (2024) wurden bereits in der Vergangenheit viele Fragestellungen und Problematiken in Planungs- und Genehmigungsprozessen gelöst und der Zugriff auf diese Informationen einen großen Mehrwert in laufenden Verfahren bringen könnten. Der Zugang zu diesen Informationen ist allerdings nur eingeschränkt möglich und mit viel Recherchearbeit verbunden (Schwindling 2024).

Durch den Aufbau von Gemeindeplattformen ist der Zugriff auf B-Pläne einfacher geworden, allerdings ist die Suche zu spezifischen Festsetzungen, Entscheidungen oder Informationen aktuell noch nicht möglich (Schwindling 2024; Bruch & Bruch 2024). Ebenfalls sind die Gemeindeportale aktuell noch nicht vernetzt, sodass eine Suche über alle Bestandsdaten in ganz Deutschland nicht möglich ist. Des Weiteren sind nicht alle Daten öffentlich über die Gemeindewebsiten zugänglich, sodass z.B. Rohdaten wie CAD-Planzeichnungen persönlich angefragt werden müssen (Schwindling 2024).

Aus den Erfahrungen lässt sich schließen, dass die Interpretation von B-Pläne zusätzliche Quellen erfordert und vergangene Entscheidungen zur Problemlösung in aktuellen Verfahren beisteuern können. Dies führt zu folgender Leitfrage: „*Wie können öffentliche Daten in der Bauleitplanung effizienter zugänglich gemacht werden?*“.

### **Datenbasierte Entscheidungen**

Die Kommune ist für die Sicherung der Bauleitplanung zuständig und muss entsprechend viele Entscheidungen im Laufe des Genehmigungsprozesses eines Bauvorhabens treffen (Kaiser 2024). Manche Entscheidungen beziehen sich auf klare Festsetzungen, die objektiv entschieden werden können, andere Entscheidungen wie z.B. zu Ausnahme- und Befreiungsanträge sind wiederum stark von der betrachtenden Perspektive abhängig (Kaiser 2024). Ebenfalls ist die Frage, ob sich das Gebäude der näheren Umgebung einfügt oder nicht, nicht immer eindeutig lösbar und erfordert in der Regel mehrere Perspektiven und Meinungen (Kaiser 2024).

Ist die Entscheidung nicht objektiv zu beantworten oder liegt im Widerspruch zu den Festsetzungen im B-Plan, wird von einer *Entscheidung nach eigenem Ermessen* gesprochen

(Geirhos 2024). Ermessensentscheidungen werden im Bauausschuss besprochen und getroffen und sind nötig, falls z.B. veraltete Festsetzungen moderne Bauvorhaben behindern würden, und zugleich problematisch, da die Entscheidungen von subjektiven Einschätzungen abhängen (Geirhos 2024).

Plant beispielsweise eine Person ein Gebäude mit PV-Anlage zu bauen, muss die Ausrichtung des Daches und die Dachform für die optimale Energieerzeugung nach den Umweltfaktoren ausgerichtet werden. Dem könnten allerdings Festsetzungen eines B-Plan widersprechen, da z.B. zum Zeitpunkt der Planung noch keine PV-Anlagen gebaut worden sind und somit nur eine ungünstige Ausrichtung des Daches erlaubt ist.

Bezüglich dieses Dilemmas könnten Daten, wie z.B. vergangene Entscheidungen zu Bauvorhaben, technische Analysen von Indikatoren in den Begründungen der Festsetzungen oder von Geodaten der näheren Umgebung, bei der Entscheidungsfindung hinzugezogen werden, um den Entscheidungsprozess zu unterstützen und objektiver und transparenter zu gestalten. Des Weiteren könnten neue B-Pläne mit einer soliden Datenbasis vorausschauender und effizienter geplant werden. Diese Betrachtungen führen zu folgender Leitfrage: „**Wie können Entscheidungen datanbasiert untermauert werden, um Fairness zu gewährleisten?**“.

### **Beteiligung und Transparenz**

Öffentliche Sitzungen erfüllen mehrere Aufgaben, wie z.B. eine breite Perspektive auf ein Bauvorhaben einnehmen, Abstimmungen zwischen den Vertreter:innen im Bauausschuss durchführen, wie auch Transparenz und Bürger:innen-Nähe herstellen (Kaiser 2024). Um als Bürger:in jedoch aktiv eigene Interessen einbringen zu können, braucht es einen Zugang zu Informationen über aktuelle Bauvorhaben und die Möglichkeit im Vorfeld Stellungnahme zu beziehen (Baur 2024). Da es bei Bauvorhaben eine Vielzahl von Akteur:innen und Interessen gibt, benötigen Verwaltungsbüros verschiedene Hilfsmittel um die Datemenge zerkleinern und gruppieren zu können, sodass die Stellungnahmen effizient weiterverarbeitet und in Entscheidungsprozess einbezogen werden können (Fina 2024; Baur 2024). Dies führt zu der Leitfrage: „**Wie können öffentliche Meinungen effizient analysiert und in Prozessen integriert werden?**“.

#### **2.2.2.3 Daten**

##### **Aufbereitung von Bestandsdaten**

Zukünftig können Informationen aus B-Pläne im XPlanung Standard mit klassischer Software ohne KI ausgelesen werden (Zwick & Wagner 2024). Stand heute liegt der Großteil aller verfügbaren B-Pläne allerdings noch unstrukturiert als PDF Format vor, was eine maschinelle Verarbeitung erschwert (Baur 2024). Die Wichtigkeit strukturierter qualitativer Daten für Anwendungen vortrainierter Sprachmodelle wurde zusätzlich am CHIASM Event<sup>6</sup>, an der THA, in mehreren Gruppendiskussionen bestätigt. In Deutschland gibt es schätzungsweise 750.000 B-Pläne, deren Neuentwicklung und Digitalisierung viel Geld und noch einige Jahre an Arbeit kosten werden (Zwick & Wagner 2024; Rehkop 2024), was sich beispielsweise in Städten wie Augsburg und vor allem in kleineren Kommunen wie Laichingen oder Gersthofen zeigt (Zwick & Wagner 2024; Kaufmann & Hascher 2024;

---

<sup>6</sup><https://www.tha.de/Informatik/CHIASM.html>

Kaiser 2024).

Die Aufbereitung der alten B-Plan Bestände birgt einige Herausforderungen. Zum Beispiel führt die fehlende Standardisierung zu schwankender Qualität der Lagegenauigkeit von B-Pläne, wie auch die Limitierung von KI Koordinaten exakt zuzuordnen (Zwick & Wagner 2024). Zusätzlich können sich B-Pläne inhaltlich stark durch z.B. die geltende Gesetzeslage, inhaltlichen Umfang, Anzahl an Änderungen und Gebietskomplexität unterscheiden (Kaufmann & Hascher 2024).

Alte B-Pläne sind allerdings per se nicht schlecht – ganz im Gegenteil, darin steckt viel ungenütztes Potenzial (Maile 2024; Fina 2024). Mit Methoden der KI könnte z.B. versucht werden B-Pläne zu klassifizieren und Indikatoren zu extrahieren, um zu erkennen, welche Festsetzungen veraltet und einer modernen nachhaltigen Stadtentwicklung entgegenstehen (Levell & Novikov 2024; Fina 2024).

Diese Zusammenhänge führen zu folgender Leitfrage: „*Wie können Bestandsdaten klassifiziert und aufbereitet werden, um deren Daten bereits heute nutzen zu können?*“.

## **Verständliche Rechtslage**

Das Baurecht ist nicht einheitlich geregelt (Zwick & Wagner 2024). In Deutschland gelten für jedes Bundesland unterschiedliche Bauordnungsgesetze und auf kommunaler Ebene die von der Gemeinde beschlossenen Festsetzungen der Bauleitplanung (Kaufmann & Hascher 2024). Zusätzlich gilt für jeden B-Plan die Gesetzesgrundlage an dem der B-Plan veröffentlicht wurde (statisches Baurecht), weshalb unterschiedliche Fassungen referenziert werden müssen (Kaufmann & Hascher 2024). Des Weiteren gilt für B-Pläne die sogenannte Festsetzungautomatik, sodass im B-Pläne nicht alle geltenden Festsetzungen explizit aufgeführt werden müssen.

Stand heute sind in der Praxis kommerzielle Rechtsdienstleistungen die Regel, wie z.B. die Platform Juris<sup>7</sup>, da diese aktuelle und ausführliche Zusatzinformationen wie z.B. Kommentare (Gerichtsbeschlüsse) zu einzelnen Paragraphen anbieten, die dabei helfen Baurecht konkret zu interpretieren (Schwindling 2024; Bruch & Bruch 2024). Öffentlich zugängliche Informationen zu Gesetzen und Urteilen sind dagegen nur über mehrere Quellen in unterschiedlicher Qualität zu finden und müssen mit hohem Zeitaufwand zusammengetragen werden (Schwindling 2024). Klare Rahmenbedingungen sind jedoch die Basis für einen effektiven Planungsprozess eines Bauvorhabens, weshalb in einigen Fällen Rechtsberatung hinzugezogen werden muss, da fehlende Informationen die Planung angreifbar machen, was wiederum zu Korrekturschleifen und erhöhten Kosten führt (Maile 2024).

Forschungsprojekte wie das Regulatory Information System (RIS) for Real Estate<sup>8</sup> an der Hochschule Luzern nehmen das Baurecht ins Visier und entwickeln aktuell eine KI Lösung, um Baugesetze maschinell zu parametrisieren und entwicklungshemmende Stellen in der Regulierung zu identifizieren (Levell & Novikov 2024). Dies führt zu folgender Leitfrage: „*Wie können rechtliche Informationen aufbereitet werden, um verständliche Rahmenbedingungen zu schaffen und Interpretationsspielräume zu reduzieren?*“.

---

<sup>7</sup><https://www.juris.de>

<sup>8</sup><https://www.hslu.ch/de-ch/hochschule-luzern/forschung/projekte/detail/?pid=6487>

### 2.2.3 Use Case zur prototypischen Umsetzung

Während der Forschung haben sich für die Entwurfs- und Genehmigungsprozesse vier Stakeholder:innen als besonders relevant herausgestellt: Der **Bauausschuss**, welcher über das gemeindliche Einvernehmen, Ausnahmen und Befreiungen entscheidet, das **Bauverwaltungsamt**, welches die rechtliche Prüfung durchführt und entsprechend Bauvorhaben mit den Festsetzungen aus dem B-Plan vergleicht, die **Architekturbüros**, welche die Bauzeichnungen entwickeln und den Bauantrag einreichen, sowie der **Gemeinde- bzw. Stadtrat** der für die Sicherung der Bauleitplanung die Verantwortung trägt. Aus der Perspektive dieser Stakeholder:innen, entwickelte ich anhand der Leitfragen, aus den Themen-schwerpunkten der Insights, 14 Use Cases, welche durch Anwendung von KI bzw. MLLM die Stakeholder:innen in Entwurfs- und Genehmigungsprozesse unterstützen; siehe Abbildung 13

Für den Bauausschuss, das Bauverwaltungamt und Architekturbüros sind vor allem ein effizienter Zugriff auf Informationen und verständliche Rahmenbedingungen aus dem B-Plan unterstützende Maßnahmen. Dazu zählen Informationen zusammenzufassen und Erklärungen zu Informationen und Entscheidungshilfen, z.B. zu Gerichtsurteilen ähnlicher Bauvorhaben, bereitzustellen. Die Basis für diese Use Cases sind ein Schnellzugriff auf öffentliche relevante Daten und Verknüpfungen zu Querverweise. Entsprechend müssen mehrere Dokumente durchsucht werden können, sodass aus dessen Kontext ein Output generiert werden kann.

Aus Sicht des Gemeinde-/Stadtrats steht vor allem die digitale Transformation und eine nachhaltige Stadtentwicklung im Vordergrund. Hierbei spielen die Weiterentwicklung von B-Pläne eine wesentliche Rolle. Dazu zählen der Input von Bürger:innen effizient auswerten zu können, wie z.B. Stimmungsbilder und thematische Schwerpunkte zu erstellen, Deltas und Ähnlichkeiten zwischen B-Pläne zu erkennen, sowie die Qualität von B-Pläne einschätzen zu können, wie z.B. anhand nachhaltiger Indikatoren. Zusätzlich können die Festsetzungen aus den B-Pläne mit dem aktuellen Baubestand bzw. Baubedarf verglichen werden, um die Planung und Sicherung zu optimieren.

Damit zu diesen Use Cases zukünftige Lösungen konzipiert und prototypisch umgesetzt werden können, sind maschinenlesbare Daten eine wesentliche Voraussetzung. Entsprechend ist das **Extrahieren von Informationen aus unstrukturierten B-Pläne** ein wichtiger Use Case für eine zukünftige digitale Bauleitplanung, welcher im Rahmen der Masterarbeit untersucht und erprobt wurde. Die Forschung beschränkte sich dabei auf die Informationen in einem B-Plan ohne Querverweise, mit dem Ziel, die Forschungsfrage „Inwiefern sind MLLMs in der Lage B-Pläne zu verstehen?“ [RQ1] zu beantworten. Des Weiteren war dieser Use Case im Rahmen der Masterarbeit realistisch umsetzbar, da ausreichend B-Pläne von der Stadt Laichingen und Augsburg vorlagen und die Forschungsergebnisse anschließend mit den jeweiligen Expert:innen evaluiert werden konnten.

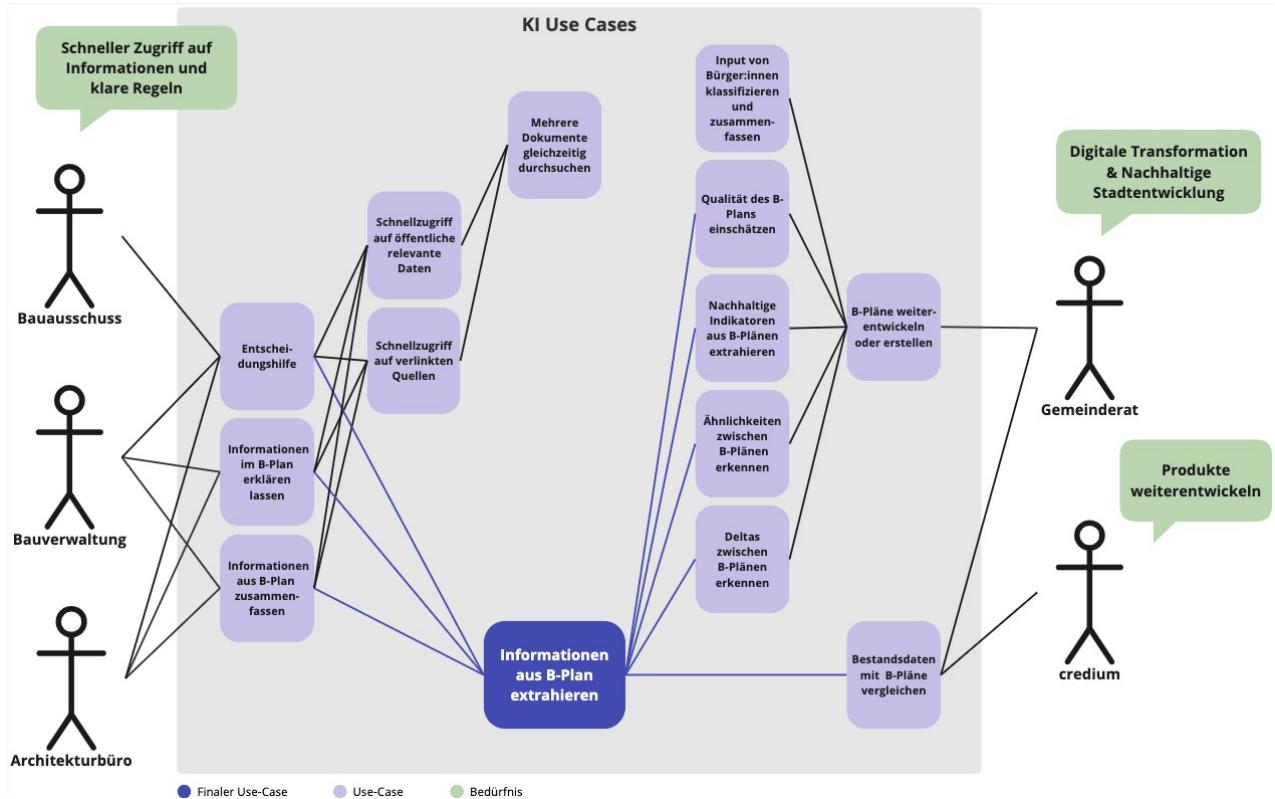


Abbildung 13: Use Case Diagramm für mögliche KI Anwendungsfälle (Eigene Darstellung)

### 3 Multimodal Large Language Models

Multimodale LLMs (=MLLMs) können unter anderem Bild und Text verarbeiten und somit auch die textlichen und zeichnerischen Teile von B-Pläne. Ziel dieses Kapitels ist es, die technischen Grundlagen hinter MLLMs vorzustellen. Die Methodik und Rechercheergebnisse werden auf den folgenden Seiten dargelegt.

#### 3.1 Vorgehen und Methodik

Für die Recherche zu technischen Grundlagen bezüglich MLLMs wurde eine Systematische Literaturrecherche durchgeführt, um relevante wissenschaftliche Arbeiten zu identifizieren, welche wesentliche Konzepte eingeführt haben, auf dessen heutige MLLMs aufbauen. Hierzu wurden zur Auswahl der Literatur thematische Schwerpunkte auf Basis des state-of-the-art Modells *GPT-4o*<sup>9</sup> festgelegt und anschließend die Inhalte aufbereitet, was im Folgenden genauer erläutert wird.

Aktuelle MLLMs wie GPT-4o können Aufgaben in den Bereichen Sprache und Audio, Computer Vision (=CV) und Natural Language Processing (=NLP) lösen (Huggingface 2024). Im Rahmen meiner Masterarbeit wird das Textverständnis von MLLMs bezüglich eines B-Plan analysiert, sodass der Aufgabenbereich auf CV und NLP eingeschränkt wurde. Zum Beispiel beinhaltet der Bereich CV Methoden wie Optical Character Recognition (=OCR) und der Bereich NLP Question Answering und Summarization.

Da die Architektur von GPT-4o nicht öffentlich zur Verfügung steht, analysierte ich repräsentativ die quelloffene state-of-the-art Architektur *LLaVA* von Liu et al. (2023b), welche im Paper „*Visual Instruction Tuning*“ erstmals veröffentlicht wurde. LLaVA eignet sich, da es für Prompt-Anweisungen (Instructions) optimiert und zu Question Answering und Summarization fähig ist, sowie zusätzlich zu klassischen LLMs Bildinformationen erfassen kann. Des Weiteren wurde die LLaVA-Architektur von den Autor:innen einfach gehalten und besteht aus wenigen bekannten Basiskonzepten; siehe Abbildung 25. Dies ermöglichte mir grundlegende Konzepte zur Verarbeitung von Bildern, sowie der Kombination aus Bild und Text via MLLMs herauszuarbeiten und passende Literatur auszuwählen.

Im ersten Schritt fokussierte ich mich auf die Verarbeitung des Textteils, welcher wesentliche Konzepte und Methoden beinhaltet, die in dem Bildverarbeitungsteil ebenfalls genutzt werden. Diesbezüglich habe ich das Paper „*Efficient Estimation of Word Representations in Vector Space*“ von Mikolov et al. (2013) ausgewählt, da in diesem das sogenannte Word2Vec Modell eingeführt wurde, womit Wort-Vektorrepräsentationen (Embeddings) erzeugt werden können, welche die semantische Bedeutung von Wörtern und deren Ähnlichkeit zueinander mathematisch erfassen. Dieses Konzept wird bis heute von state-of-the-art LLMs verwendet und gibt einen Einblick, wie LLMs Sprache erfassen und interpretieren. Anschließend wählte ich das Paper „*Attention Is All You Need*“ von Vaswani et al. (2017) aus, da in diesem die Transformer-Architektur vorgestellt wurde, auf dessen Konzepte heutige state-of-the-art LLMs größtenteils aufgebaut sind. Neben den wesentlichen Bauteilen Input-Embedding-Layer und Positional-Encoding bildet der Attention-Layer das Herzstück einer Transformer-Architektur, dessen Funktionsweise anhand des Decoders GPT-1 bzw. GPT-2 (Radford & Narasimhan 2018; Radford et al. 2019) vertieft und via des Encoders BERT (Devlin et al. 2019) ergänzt wurde.

---

<sup>9</sup><https://platform.openai.com/docs/models/gpt-4o>

Für die Bildverarbeitung nutzten die Autor:innen von LLaVA einen sogenannten Vision Transformer (=ViT) als grundlegenden Baustein, womit das MLLM Bildinformationen verstehen und verarbeiten kann. Diesbezüglich untersuchte ich das Paper „*An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*“ von Alexey et al. (2020), worin die grundlegende Funktionsweise von *Vision Transformer* vorgestellt wurde und unter anderem erklärt, wie Bild-Vektorrepräsentationen (Embeddings) erzeugt werden. Des Weiteren handelt es sich bei dem ViT von LLaVA um ein speziell trainiertes Modell von OpenAI namens CLIP. Das Paper „*Learning Transferable Visual Models From Natural Language Supervision*“ von Radford et al. (2021) geht speziell auf das Training des state-of-the-art MLLM CLIP ein und erläutert, wie ein MLLM eine einheitliche Text- und Bild-Sprache lernen kann. Zusammen ergeben diese Konzepte einen Einblick, wie MLLMs wie GPT-4o Bilder erfassen und interpretieren.

Zusätzlich erweiterte ich die Literaturrecherche um die Konzepte zur Skalierung der maximalen Auflösung eines Input-Bildes, was zur Verarbeitung von B-Pläne relevant ist. Hierzu untersuchte ich die aktuellen Weiterentwicklungen von LLaVA anhand der Paper „*Improved Baselines with Visual Instruction Tuning*“ von Liu et al. (2023a) und „*LLaVA-NeXT: Improved reasoning, OCR, and world knowledge*“ von Liu et al. (2024), sowie das Paper „*Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*“ von Liu et al. (2021).

## 3.2 Ergebnisse

LLMs wurden für die maschinelle Verarbeitung von natürlicher Sprache entwickelt und sind in das Forschungsfeld der natürlichen Sprachverarbeitung (=NLP) einzuordnen (Wikipedia 2024o). Die Schlüsseltechnologie hinter LLMs ist die sogenannte Transformer-Architektur, welche auch für andere Aufgaben, wie z.B. in der Bildverarbeitung, eingesetzt werden kann, was sich heutige MLLMs zunutze machen (Wikipedia 2024o). Im Folgenden werden die technologischen Grundbausteine hinter MLLMs vorgestellt.

### 3.2.1 (Text-)Transformer-Architektur

Die Darstellung von Wörtern als Vektoren, sogenannte Embeddings, ist ein zentraler Ansatz in der maschinellen Sprachverarbeitung. Vor 2018 wurden Embeddings via neuronale Netze, wie z.B. dem Word2Vec Modell von Mikolov et al. (2013), erzeugt bzw. gelernt – sogenanntes Representational Learning. Als Ergebnis wird jedes Wort durch einen eindeutigen Vektor in einem n-dimensionalen Raum repräsentiert (Mikolov et al. 2013). Diese Vektoren werden so trainiert, dass Wörter die in ähnlichen Kontexten vorkommen ähnliche Vektoren haben (Mikolov et al. 2013). Das Resultat ist, dass semantisch ähnliche Wörter im Vektorraum nahe beieinander liegen und Analogien/Beziehungen zwischen Wörtern mathematisch via Vektorrechnungen berechnet werden können (Mikolov et al. 2013); siehe Abbildung 14.

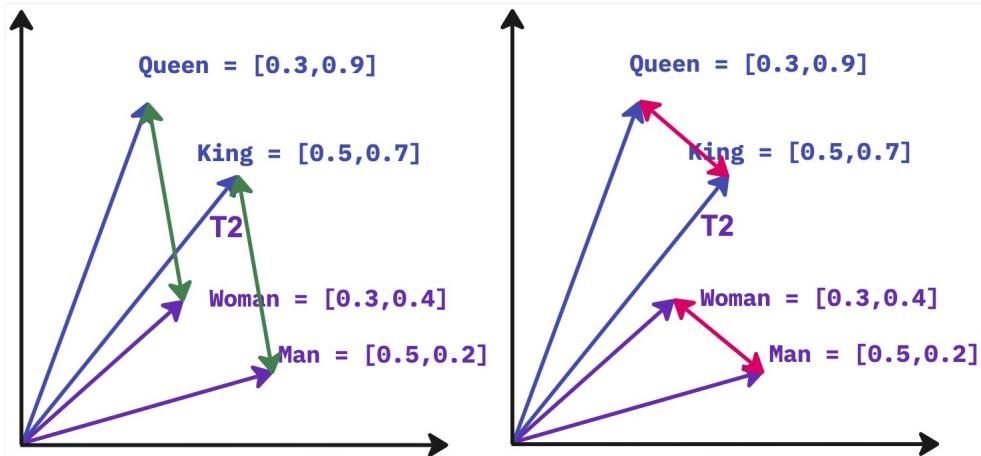


Abbildung 14: Das klassische Beispiel *König + Frau – Mann ≈ Königin* von Mikolov et al. (2013) in 2D (Darstellung von ResearchGate (2024))

Diese Wort-Vektorrepräsentationen werden auch als statische Embeddings bezeichnet, da ein Wort einer spezifischen Bedeutung zugeordnet wird (Peters et al. 2018). Eindeutige Bedeutungsrepräsentationen sind allerdings problematisch, wenn ein Wort mehrere Bedeutungen hat, wie z.B. bei der Homonymie „apple“, was je nach Kontext eine Frucht oder ein Technologie-Institut sein kann (Peters et al. 2018).

Ab 2018 ermöglichten sogenannte Transformer-Modelle dynamische kontextualisierte Embeddings eines Wortes je nach Kontext (Vaswani et al. 2017). Bezuglich des Wortbeispiels „apple“ können jetzt je nach Kontext, wie z.B. „buy an apple and an orange“ und „apple unveiled the new phone“, unterschiedliche Vektorrepräsentationen erzeugt werden. Wie diese sogenannten Context-Embeddings gelernt werden können, veranschaulicht die Grundarchitektur in Abbildung 15. Vaswani et al. (2017) veröffentlichte den ursprünglichen Encoder-Decoder Transformer, wovon sich LLMs wie z.B. *BART* (Lewis et al. 2019) (*Encoder-Decoder* Transformer), *BERT* (Devlin et al. 2019) (*Encoder-only* Transformer) und *GPT-1* (Radford & Narasimhan 2018) (*Decoder-only* Transformer) entwickelten, die wichtige Meilensteine im Bereich NLP repräsentieren (Wikipedia 2024o).

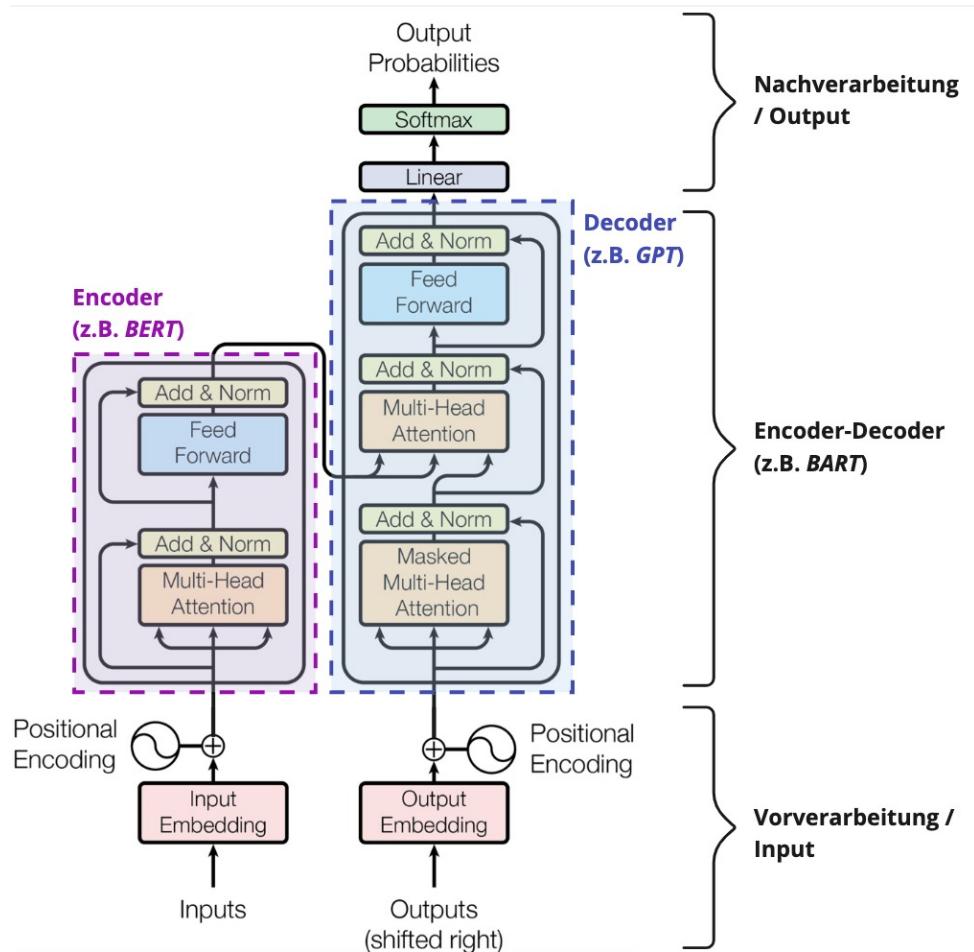


Abbildung 15: Transformer-Architektur (Eigene Darstellung. Angelehnt an Vaswani et al. (2017))

Sowohl für **Encode-Decoder**, **Decoder-only** oder **Encoder-only** Transformer-Architekturen sind die wesentlichen Bestandteile bei der Erzeugung von kontextuellen Embeddings die Layer **Input-Embedding**, **Positional-Encoding** und **Attention** (Vaswani et al. 2017). Im Folgenden werden die einzelnen Schritte durchlaufen, um zu verstehen, wie ein Context-Embedding entsteht:

### 3.2.1.1 Input-Embeddings Layer

Im Folgenden wird der Input-Embedding Layer am Beispiel des GPT-2 Decoder-only Transformer Modells von Radford et al. (2019) erklärt. Um einen Input-Text, wie z.B. „**aaabdaaa-bac**“ in ein Transformer-Modell einzufügen zu können, muss der Text im Layer Input-Embedding zu Embeddings umgewandelt werden; siehe Abbildung 16.

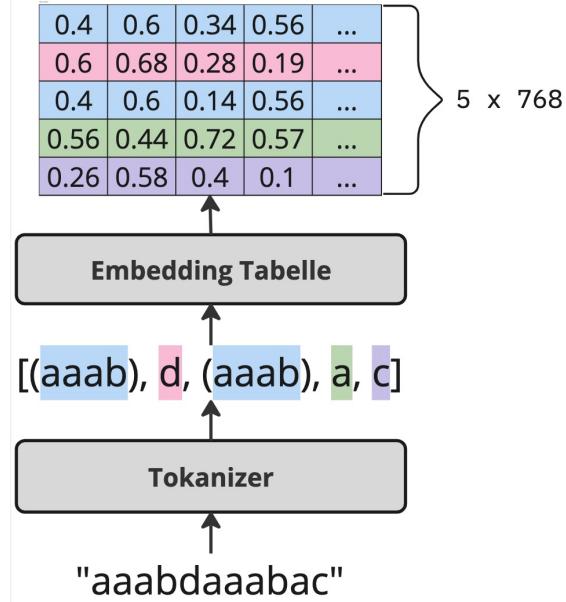


Abbildung 16: Input-Embedding Layer (Eigene Darstellung. Angelehnt an Radford et al. (2019))

Dieser Schritt ist notwendig, da neuronale Netze (=NN) Informationen durch Vektoren, Matrizen und Tensoren verarbeiten. Dazu wird der Text durch ein sogenannten **Tokenizer** in eine Sequenz von sogenannten Tokens zerlegt. Tokens sind einzelne Zeichen oder Kombinationen von Zeichen und dienen als grundlegende Einheit in Transformer Modellen. Die Summe aller Tokens wird als Vokabular bezeichnet. Der Tokenizer von GPT-2 geht dabei systematisch anhand des Algorithmus **Byte Pair Encoding** (=BPE) vor, was im folgenden Beispiel von Wikipedia (2024h) erläutert wird:

1. Text in Zeichensequenz  $T$  zerlegen:  $T = [a, a, a, b, d, a, a, a, b, a, c]$  und Vokabular  $V$  initialisieren ohne Duplikate:  $V = [a, b, d, c]$
2. Häufigstes Zeichenpaar aus  $T$  identifizieren:  $(aa)$
3.  $T$  und  $V$  wie folgt updaten:  $T = [(aa), a, b, d, (aa), a, b, a, c], V = [a, b, d, c, aa]$
4. Den Prozess wiederholen bis die gewünschte Anzahl von Token erreicht ist oder keine häufigen Paare mehr gefunden werden: Nächstes Paar ist  $(ab)$
5.  $T$  und  $V$  wie folgt updaten:  $T = [(aa), (ab), d, (aa), (ab), a, c], V = [a, b, c, aa, ab]$
6. Nächstes Paar ist  $(aab)$
7.  $T$  und  $V$  wie folgt updaten:  $T = [(aab), d, (aab), a, c], V = [a, b, c, aa, ab, aab]$
8. Ende.

Die Tokens  $T = [(aaab), d, (aaab), a, c]$  dienen anschließend als Keys für die interne Embedding-Tabelle, welche zu jedem Token eine Vektorrepräsentation beinhalten. Die Embedding-Tabelle  $E$  ist eine Matrix in der Form:  $E = V \times d$ . Das Vokabular  $V$  von GPT-2 umfasst insgesamt 50.257 Tokens und wurde anhand von Trainingsdaten anhand des BPE Algorithmus erfasst. Die Dimension  $d$  der Embeddings wurde manuell auf  $d = 768$  festgelegt. Dieser Wert wurde durch Experimente und Faktoren, wie unter anderem Modellkapazität und

Rechenaufwand ermittelt. Die Parameter der Embeddings wurden typischerweise mit Zufallswerten initialisiert und anschließend durch das Training des Modells anhand von Trainingsdaten via Backpropagation gelernt. Der finale Output  $O$  des Input-Embedding Layers hängt von der Anzahl  $n$  der Input-Tokens ab und ist eine Matrix der Form  $O = n \times d$ ; im Beispiel wäre das  $O = 5 \times 768$ .

### 3.2.1.2 Positional-Encoding Layer

Die Reihenfolge von Wörtern in einem Satz ist eine wichtige Information für das Verständnis und die Bedeutung eines Satzes (Camacho-Collados & Pilehvar 2018). Zum Beispiel unterscheidet sich der Satz „**Wölfe fressen Schafe**“ von „**Schafe fressen Wölfe**“ ausschließlich von der Wortreihenfolge. Transformer-Modelle nutzen entsprechend einen Positional-Encoding Layer, um die Information der Position der jeweiligen Tokens zu kodieren (Vaswani et al. 2017). Damit Position-Embeddings im Transformer-Modell verarbeitet werden können, werden diese ebenfalls als Vektoren repräsentiert, welche die gleiche Länge haben wie die Input-Embeddings;  $d = 768$  (Vaswani et al. 2017). Dadurch kann ein Position-Embedding zu einem Input-Embedding addiert werden, dessen Ergebnis eine neue Vektorrepräsentation ist, welche die Bedeutung und Position eines Tokens repräsentiert (Vaswani et al. 2017). Der finale Output des Positional-Encoding Layers kann zusammenfassend als Matrix dargestellt werden (Vaswani et al. 2017).

Die Author:innen Vaswani et al. (2017) verwenden ein intelligentes Position-Encoding Schema: für **ungerade Positionen die Sinusfunktion** und für **gerade Positionen die Kosinusfunktion**:

$$\begin{aligned} PE(pos, 2i) &= \sin(pos/10000^{2i/d}) \\ PE(pos, 2i + 1) &= \cos(pos/10000^{2i/d}) \end{aligned}$$

Die Konstante 10.000 ist ein mathematischer „Trick“, sodass der berechenbare Abstand zwischen Position-Embeddings (Vektoren) monoton steigend verläuft. Bei einem Wert von 100 ist dies jedoch nicht der Fall.  $pos$  ist die Position des Tokens im Input,  $i$  ist der jeweilige Parameterindex des Position-Embeddings und  $d$  die Dimension. Mithilfe der trigonometrischen Funktionen bleiben die Werte jedes Parameters des Position-Embeddings zwischen –1 und 1. Ebenfalls ist damit sichergestellt, dass die Position-Embeddings unabhängig von der Anzahl der Input-Embeddings konstant bleiben. Abbildung 17 visualisiert das Schema anhand der Sinusfunktion für das erste Token ( $Index = 0$ ) und dritte Token ( $Index = 2$ ) eines Inputs.

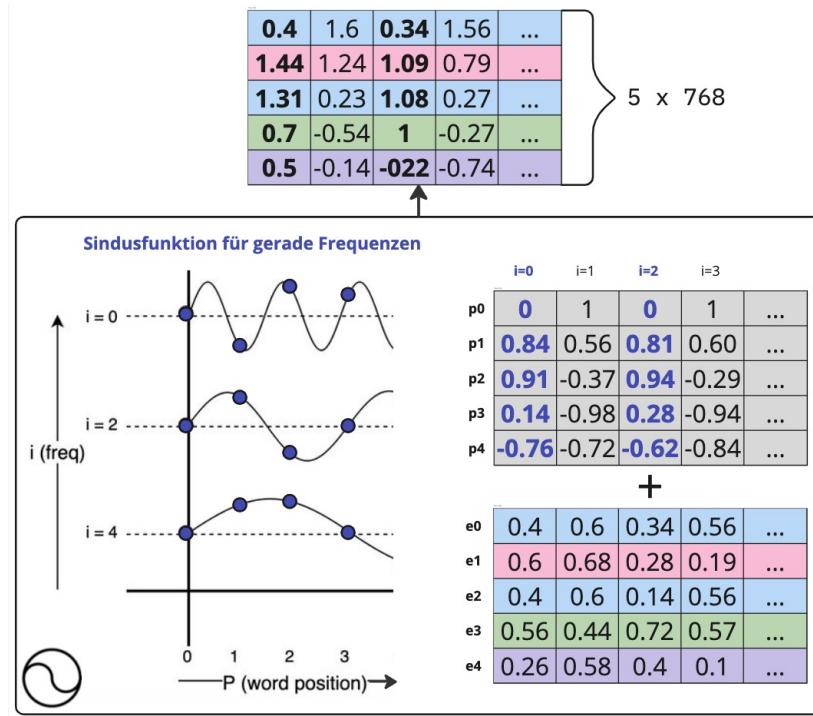


Abbildung 17: Positional-Encoding Layer (Eigene Darstellung. Angelehnt an Anwar & Sayeed (2022))

### 3.2.1.3 Attention Layer

(Multi-Head-)Attention wurde im Paper von Vaswani et al. (2017) vorgestellt und ist der wesentliche Bestandteil eines Transformer-Modells in dem Bezug auf den Kontext eines Tokens. Das folgende Beispiel in Abbildung 18 veranschaulicht wie die Vektorrepräsentation des Tokens „apple“ anhand des Kontexts durch die Attention-Methodik optimiert werden kann. Das Beispiel verwendet Wörter als Tokens, um das Konzept einfacher erklären zu können.

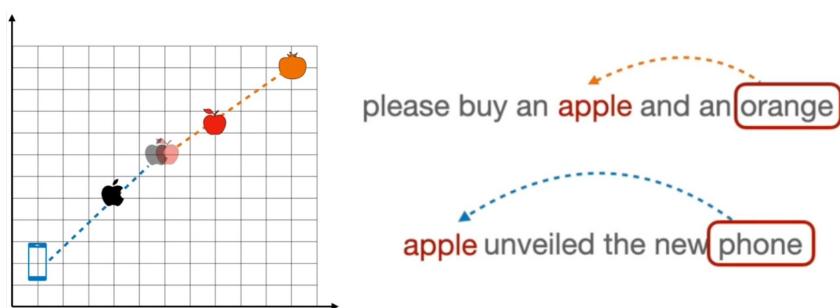


Abbildung 18: Einfluss von Kontext auf den Token „apple“ via Attention (Darstellung von Serrano (2023))

Vor dem Attention-Layer ist das Input-Embedding „apple“ mittig im n-dimensionalen Raum anhand der Trainingsdaten definiert. Via Attention wird anschließend dynamisch, abhängig des Kontexts des Gesamtinputs, das Input-Embedding verschoben, sodass die Bedeutung von „apple“ je nach Kontext einer Frucht oder Technologie-Institut zugeordnet werden kann. Mathematisch vereinfacht formuliert, wird im Attention-Layer ein Attention-Embedding (Vektor) berechnet, welcher anschließen mit dem Input-Embedding addiert wird, was im Ergebnis zu einem Context-Embedding führt. Abbildung 19 veranschaulicht die mathematische Formel  $\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V$  von Vaswani et al. (2017), womit ein Attention-Embedding  $Z$  berechnet wird:

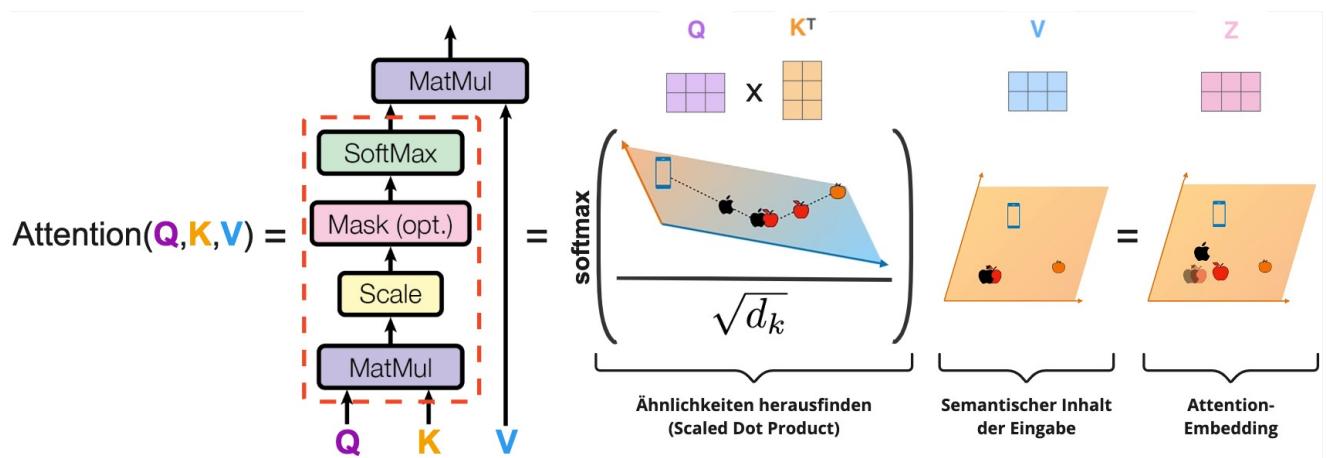


Abbildung 19: Visualisierung von Attention (Eigene Darstellung. Angelehnt an Vaswani et al. (2017) und Serrano (2023))

1. Zu Beginn werden aus dem Input-Embedding  $X$  drei weitere Embeddings erzeugt: **Query-, Key- und Value-Embedding** ( $Q, K, V$ ). Wie in Abbildung 20 zu sehen ist, werden diese Embeddings anhand von Transformations-Matrizen  $W$  berechnet. Die Parameter von  $W$  werden während dem Training des Transformer-Models gelernt.

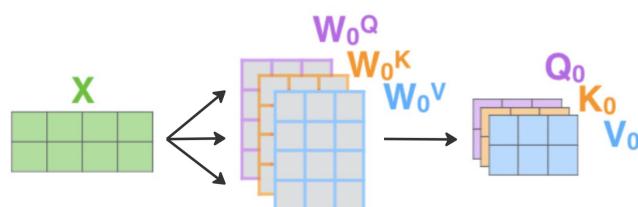


Abbildung 20:  $Q, K, V$  Embeddings (Eigene Darstellung. Angelehnt an Alammar (2020))

2. Ziel der Embeddings  $Q$  und  $K$  ist es das Input-Embedding so zu repräsentieren, so dass Ähnlichkeiten zwischen Tokens bestmöglich erkannt werden können (Serrano 2023). In Abbildung 19 visualisiert das verzerrte Koordinatensystem (Zähler im Bruch), wie die Vektoren modifiziert werden können, um die Trennschärfe der Ähnlichkeiten zwischen den Tokens (*apple, phone, orange*) hervorzuheben (Serrano 2023). Die Ähnlichkeit zwischen Vektoren wird anschließen anhand eines Skalarprodukts berechnet – je größer der Wert, desto ähnlicher sind sich die Vektoren (Embeddings). Die Zahlenwerte werden anschließen mathematisch für die nächsten Rechenschritte angepasst.
3. Ziel des Embeddings  $V$  ist es, den semantischen Inhalt der Eingabe bestmöglich zu repräsentieren (Serrano 2023). Dieser wird wie die Embeddings  $Q$  und  $K$  durch eine Matrix  $W$  berechnet, welche aus gelernten Parametern besteht. Anschließend werden die Ähnlichkeiten ( $\text{softmax}(Q \times K / \sqrt{d_k})$ ) auf die Semantik ( $V$ ) angewendet, woraus ein finales Attention-Embedding  $Z$  entsteht. In Abbildung 19 kann entsprechend beobachtet werden, wie die Vektorrepräsentation von „apple“ ( $Z$ ) entsprechend optimiert werden konnte.

Ein wichtiger Aspekt ist, dass Attention auf gelernte Parameter ( $W$  Matrizen) beruht, welche je nach Trainingsdatensatz und -bedingungen unterschiedlich initialisiert werden. Um bestmögliche Muster und Beziehungen aus der Eingabe extrahieren zu können, wird im Paper von Vaswani et al. (2017) *Multi-Head-Attention* verwendet. Wie Abbildung 20 anhand der Indizes bereits angedeutet hat, können mehrere Instanzen der Matrizen  $W$  (=Heads) genutzt werden, um weitere Varianten von  $Q$ -,  $K$ -,  $V$ -Embeddings zu berechnen, um diese anschließend zu einem bestmöglichen Attention-Embedding  $Z$  zusammenzuführen.

Die Matrizen  $W$  im Attention-Layer sind von entscheidender Bedeutung im Hinblick auf Encoder-only (z.B. BERT) und Decoder-only (z.B. GPT-1) Transformer Modelle. Diese Transformer-Architekturen wurden bereits in Abbildung 15 dargestellt. Bzgl. der Context-Embeddings ist hierbei zu erwähnen, dass die Matrizen  $W$  je nach Modell-Typ unterschiedlich angewendet werden, was zu unterschiedlichen Repräsentationen der Context-Embeddings führt, siehe Abbildung 21.



Abbildung 21: Beispiel von Masked-Attention bzgl. GPT-1 (Eigene Darstellung. Angelehnt an Radford & Narasimhan (2018))

GPT-1 (Generative Pre-Trained Transformer 1) ist ein unidirektionales Modell, das den Kontext nur von links nach rechts erfasst. Diesbezüglich werden die Matrizen  $W$  manipuliert

bzw. maskiert (siehe Nullen in Abbildung 21). Dies führt dazu, dass GPT-1 stärker in generativen Aufgaben wie der Textvorhersage ist (Radford & Narasimhan 2018).

BERT (Bidirectional Encoder Representations from Transformers) verwendet eine bidirektionale Trainingsmethode, die den Kontext eines Wortes sowohl von links als auch von rechts berücksichtigt (komplette Matrix). BERT ist dadurch besser für Aufgaben geeignet, bei denen der vollständige Kontext eines Wortes wichtig ist (Devlin et al. 2019).

### 3.2.2 Vision-Transformer-Architektur

Vision-Transformer(=ViT) wurden erstmals im Paper „*An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*“ von Alexey et al. (2020) vorgestellt und erweitert damit den Anwendungsbereich von Transformer-Modelle um Aufgaben in der Bildverarbeitung. Das Vision-Transformer-Modell von Alexey et al. (2020) unterscheidet sich von einem klassischen Text Encoder-only Transformer-Modell wie BERT prinzipiell nur in dem Input-Embedding Layer. Statt Wort-Embeddings erzeugt ein Vision-Transformer allerdings im Input-Embedding-Layer aus einem Bild sogenannte *Patch-Embeddings*; siehe Abbildung 22.

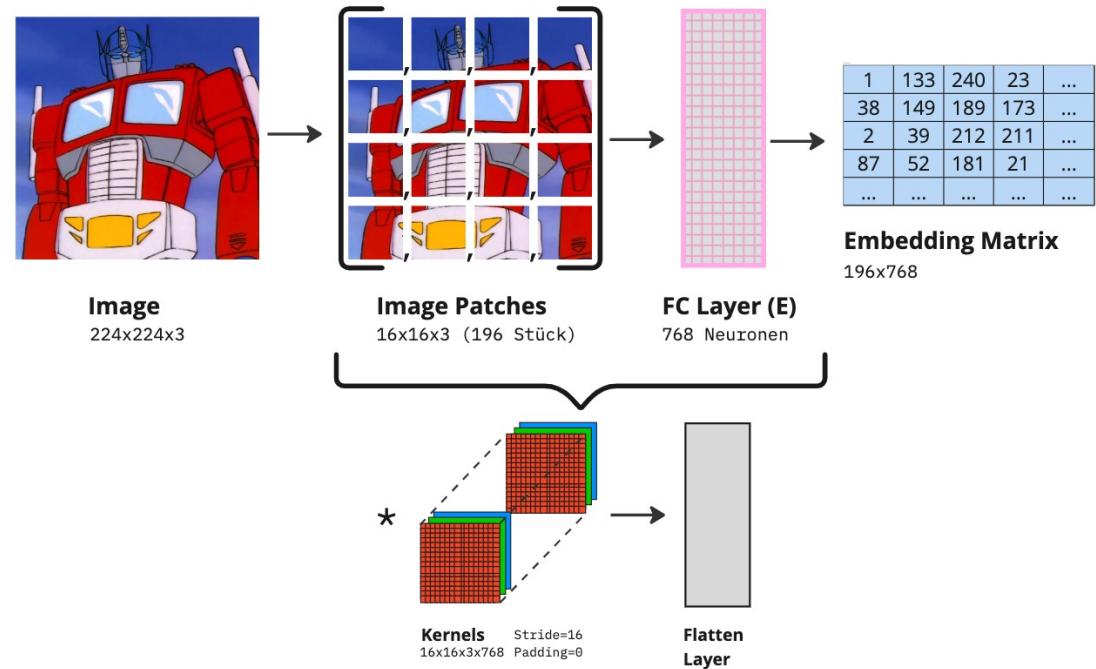


Abbildung 22: Patch Embeddings (Eigene Darstellung. Angelehnt an Alexey et al. (2020))

Damit der ViT ein Bild verarbeiten kann, muss dieses in eine Vektorrepräsentation der Dimension  $d = 768$  transformiert werden. Ein Bild  $X$  besteht jeweils aus den Farbkanälen  $C = 3(R, G, B)$ , Höhe  $H = 244px$  und Breite  $B = 244px$ .  $H \times W$  entspricht der Auflösung des Bildes. Im Paper von Alexey et al. (2020) wird im ersten Schritt ein Bild in eine Sequenz mit  $P \times P$  große Teile, sogenannte Patches, zerlegt bzw. gerastert. Die Patches repräsentieren hierbei Tokens. Die Autoren wählten eine Patchgröße von  $P = 16px$ , wovon sich der Titel des Papers ableiten lässt.

Anders als bei einem Text-Transformer-Modell werden die Patches nicht wie Tokens mit einer Embedding-Tabelle gemappt. Alexey et al. (2020) nutzten dafür eine sogenannte *Linear Projection*, eine trainierbare neuronale fully-connected Schicht  $E$  die die Image-Patches ( $16 \times 16 \times 3$ ) in Patch-Embeddings  $Z$  mit einer Dimension  $d = 768$  transformiert:  $z_0 = [x_p^1 E; x_p^2 E; x_p^3 E; \dots; x_p^N E]$ . Die Operationen des Zerlegens und Projizierens können kombiniert auch als 2D-Konvolution interpretiert werden. Hierbei wäre zusätzlich ein *Flatten Layer* nötig, um die Patch-Embeddings von einer Tensor-Form  $14 \times 14 \times 768$  zu einer linearen Sequenz  $196 \times 768$  umzuschichten. Die Gewichte des Projection-Layers  $E$  werden mit Zufallszahlen initialisiert und während dem Training des ViT-Modells anhand von Trainingsdaten via Backpropagation gelernt. Dadurch lernt der Projection-Layer spezifische Merkmale eines Bilds zu erkennen.  $N = HW / P^2 = 196$  ist die Anzahl der Patch-Embeddings in der Embedding-Matrix  $N \times d$ .

Weiterentwicklungen bzw. Varianten von ViT wie z.B. der *(Faster) Swin Transformer* von Liu et al. (2021), verwenden mehrere Patchgrößen  $P$  in Kombination mit mehreren Attention-Window-Größen und -Positionen, wodurch die resultierenden Image-Embeddings qualitativere Informationen enthalten, um Fähigkeiten wie z.B. Object Detection und Semantic Segmentation zu verbessern; siehe Abbildung 23.

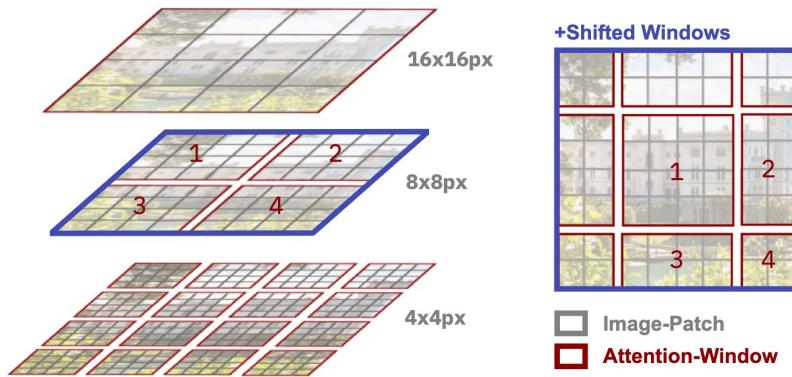


Abbildung 23: Swin-Transformer (Eigene Darstellung. Angelehnt an Liu et al. (2021))

Anstelle der Berechnung aller Patches  $P$ , wie bei ViT ( $16 \times 16$ ), konzentriert sich der Swin-Transformer auf Patches innerhalb eines Windows (Liu et al. 2021). Zusätzlich werden die Positionen der Windows verändert, um den jeweiligen Kontext um naheliegende Pixel zu erweitern (Liu et al. 2021). Des Weiteren werden auch kleinere Patchgrößen ( $4 \times 4$ ,  $(8 \times 8)$ ) verwendet, um eine feingranulare Analysen eines Bildes zu ermöglichen (Liu et al. 2021). Das Grundprinzip erinnert an die sogenannte Sliding Window Methode in Konvolutionsnetzen (=CNN), womit eine kleine Matrix (=Filter) schrittweise über ein Bild geschoben wird. Jeder Schritt wird mathematisch als sogenannte Konvolution bzw. Faltung berechnet, wie in Abbildung 22 bereits angedeutet wurde.

### 3.2.3 Zusammenführung von Text und Vision

Im Folgenden wird repräsentativ das state-of-the-art MLLMs CLIP beleuchtet, welches die Modalitäten Bild und Text in einem Modell verknüpft. Anschließend wird das state-of-the-art MLLM LLaVA präsentiert, welches grundlegende Techniken aus der Sprach- und Bildverarbeitung kombiniert, um komplexe Aufgaben in GPT-4o-Manier zu lösen.

### 3.2.3.1 CLIP (Contrastive Language-Image Pre-Training)

CLIP ist ein multimodales Bild- und Sprach-Modell von (Radford et al. 2021). Es wurde trainiert, um Ähnlichkeiten zwischen Text und Bild zu erkennen, sodass Bilder via Zero-Shot-Prompting klassifiziert werden können. Zero-Shot-Prompting ist eine Technik, bei der ein Modell eine Aufgabe ohne spezifisches Training für diese Aufgabe löst, nur basierend auf der Eingabeaufforderung (=Prompt). Die Fähigkeiten von CLIP wurden durch das sogenannte Contrastiv Pre-Training erreicht, welches Abbildung 24 veranschaulicht. Für das Training wurde von OpenAI der Trainingsdatensatz WIT (=Web-Image-Text) mit 400 Millionen Bild-Text Paaren erstellt.

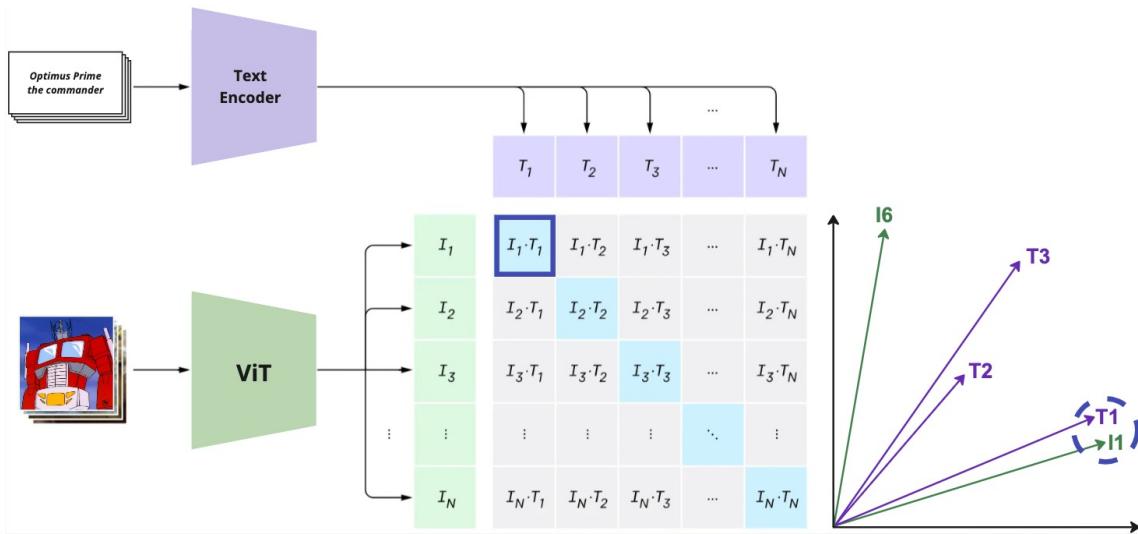


Abbildung 24: Contrastive Pre-Training (Eigene Darstellung. Angelehnt an Radford et al. (2021))

CLIP nutzt eine Encoder-Encoder Transformer-Architektur und kombiniert einen Encoder-only Text Transformer mit einem Encoder-only ViT. Beide Encoder erzeugen Embeddings in gleicher Dimension  $d = 768$ , sodass die Ähnlichkeit zwischen den Embeddings mittels der Berechnung eines Skalarprodukts ermittelt werden kann. Während dem Training werden die Gewichte der Transformer so optimiert, dass zusammengehörende Text-Bild-Vektorpaare (blau hervorgehoben) bestmöglich im n-dimensionalen Raum übereinstimmen, wie z.B. das Paar  $(T_1, I_1)$ . Alle weiteren Kombinationsmöglichkeiten wie z.B.  $(T_2, I_1), (T_3, I_1)$  werden maximal weit voneinander entfernt – daher röhrt der Begriff Contrastiv Pre-Training. Anhand des Trainings lernte das Modell eine einheitliche Vektorsprache für Bild und Text, im Speziellen state-of-the-art Image-Representation, sowie als Nebeneffekt zahlreiche Fähigkeiten wie z.B. OCR.

### 3.2.3.2 LLaVA (Large Language and Vision Assistant)

LLaVA ist ein quelloffenes state-of-the-art MLLM, das für Chat-Anwendungen wie Prompt-Anweisungen (Instructions) optimiert wurde. Es wurde im Paper „*Visual Instruction Tuning*“ von Liu et al. (2023b) veröffentlicht und von Liu et al. (2023a) und Liu et al. (2024)

weiterentwickelt.

Die LLaVA 1.0 Architektur besteht aus drei Hauptkomponenten: ein *Vision Transformer*, ein *Linear Projection Layer* und ein *Large Language Model*, siehe Abbildung 25. Der vorgenannte Vision-Transformer *ViT-L/14*, aus dem *CLIP Model*, erzeugt aus dem Bild-Input  $X_v$  Bild-Embeddings  $Z_v$ . Diese werden anschließend anhand eines trainierbaren *Linear Projection Layers*  $W$  in die Dimension der Text-Embeddings  $H_q$  gemappt, sodass Text-Embeddings  $H_q$  und Bild-Embeddings  $H_v = W \times Z_v$  gemeinsam in einem Sprachmodell verarbeitet werden können. Als Sprachmodell dient das vorgenannte LLM LLaMA von (Touvron et al. 2023), ein autoregressives Decoder-only Transformer Modell, was den Text-Input  $X_q$  und die vorverarbeiteten Bild-Embeddings  $H_v$  entgegennimmt und zu einer Ausgabe  $X_a$  weiterverarbeitet.

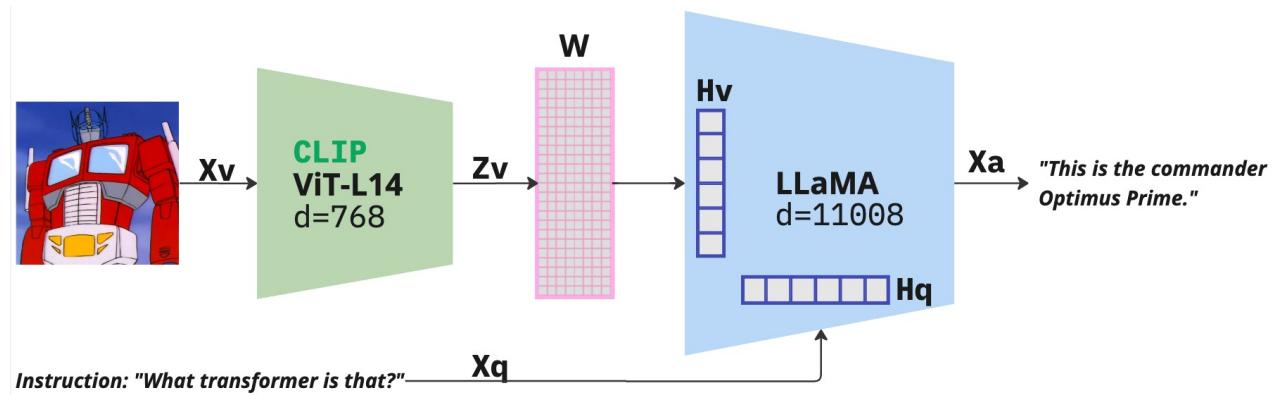


Abbildung 25: LLaVA 1.0 Architektur (Eigene Darstellung. Angelehnt an Liu et al. (2023b))

LLaVA durchlief einen zweistufigen Trainingsprozess, um Fähigkeiten für spezifische Chat-Anwendungen anhand von Prompt-Anweisungen unter der Nutzung multimodaler Trainingsdaten zu lernen, sogenanntes Instruction Tuning, wie in Abbildung 25 angedeutet. Im Pre-Training wurden zunächst nur die Gewichte des Projection-Layers  $W$  gelernt, um Bild-Embeddings der Dimension  $d = 768$  auf die größere Dimension  $d = 11008$  der Text-Embeddings abzustimmen. Anschließend erfolgte im Fine-Tuning die Optimierung der Gewichte des Projection-Layers  $W$  und des Sprachmodells LLaMA, um das Sprachmodell auf multimodale Eingaben anzupassen. Die Gewichte des CLIP Vision Transformers wurden nicht verändert, da die erzeugten Bild-Embeddings bereits für eine Text-Bild Verarbeitung optimiert sind. Um Domänenpezifische Daten optimiert mit LLaVA verarbeiten zu können, ist es möglich das Modell via Full Fine-Tuning oder LoRA Fine-Tuning (Hu et al. 2021) weiter anzupassen. Dazu stehen drei Varianten von LLaVA 1.6 mit 34, 13 und 7 Milliarden Parameter bereit (Liu et al. 2024).

LLaVA 1.6 ist in der Lage eine Bildauflösung von bis zu  $672 \times 672\text{px}$  ( $336 \times 1344\text{px}$ ,  $1344 \times 336\text{px}$ ) einzulesen (Liu et al. 2024). Die Technik hierzu wurde mit LLaVA 1.5 eingeführt, siehe Abbildung 26.

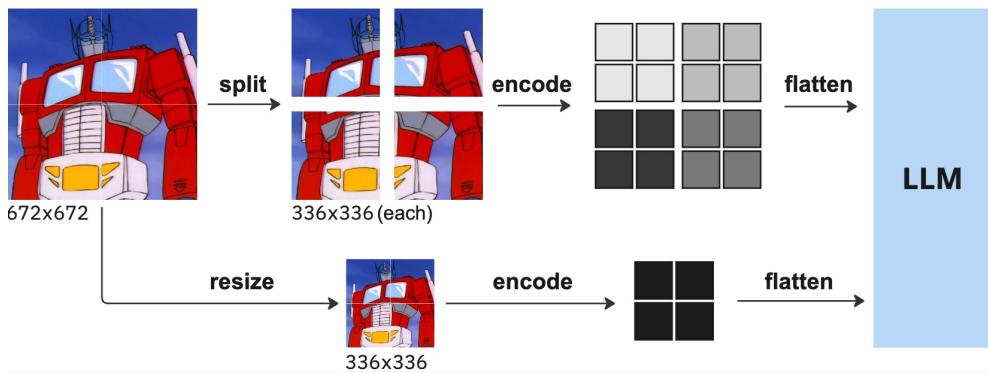


Abbildung 26: Skalierung von LLaVA 1.5 für höhere Bildauflösungen (Eigene Darstellung. Angelehnt an Liu et al. (2023a))

Der CLIP Vision Transformer ViT-L14 kann Bildauflösungen von  $336 \times 336$ px einlesen. Um höhere Bildauflösungen in LLaVA verarbeiten zu können, wird das Bild in vier gleich große Teile geteilt bzw. gerastert, die anschließend unabhängig voneinander via ViT-L14 kodiert werden. Zusätzlich wird das Gesamtbild skaliert und kodiert, um dem LLM einen globalen Kontext zu geben. Durch diesen Prozess konnte die Bildauflösung vervierfacht werden, was unter anderem zu verbesserten Fähigkeiten im Bereich Visual Reasoning und OCR führte.

## 4 Experimente und Evaluation

In diesem Kapitel werden die Planung, Vorbereitung und Durchführung der Experimente, sowie die Evaluation der Ergebnisse vorgestellt. Der entstandene Code steht im Anhang (Kapitel 6) zur Einsicht zur Verfügung.

### 4.1 Vorgehen und Methodik

Ziel der Experimente war es, die Forschungsfrage „*Inwiefern sind MLLMs in der Lage B-Pläne zu verstehen?*“ [RQ1] anhand des Use Case „*Extrahieren von Informationen aus unstrukturierten B-Plänen*“ (Abschnitt 2.2.3) zu beantworten. Vor dem Hintergrund der Prüfungsschritte des Prüfungsprogramms aus Abschnitt 2.2.1.3, sind die **Nutzungsschablone**, **Art der baulichen Nutzung**, **Maß der baulichen Nutzung** und **Bauweise, überbaubare Grundstückflächen** relevante Informationen die im Rahmen der Experimente von einem MLLM extrahiert und verarbeitet wurden. Als Leitlinien für den Aufbau der Experimente dienten die Unterfragen „*Wie gut ist die Qualität der extrahierten Daten?*“ [SQ1.1] und „*Inwiefern kann ein Verständnis zu den extrahierten Informationen geschaffen werden?*“ [SQ1.2]. Abschließend wurden die extrahierten Informationen zur Lösung einzelner Prüfungsschritte eines Bauvorhabens im Bereich des B-Plans L04 DG1 genutzt (siehe Tabelle 2), um die Forschungsfrage „*Können MLLMs die Inhalte von maschinenlesbaren Bebauungsplänen mindestens genauso gut verstehen, wie Menschen dies können?*“ [RQ2] zu beantworten.

Als Datengrundlage dienten 56 B-Pläne, bereitgestellt von der Stadt Laichingen und 3 B-Pläne der Stadt Augsburg. Diese stellen lediglich eine Stichprobe dar und können die geschätzten 750.000 B-Pläne im Raum Deutschland nicht repräsentieren (Rehkop 2024). Die B-Pläne wurden anhand der Kriterien **gestalterische Merkmale**, **inhaltliche Merkmale** und **technische Merkmale** analysiert und gruppiert, sodass der gesamte Datensatz in drei kleinere Datensatzgruppen (=DG) aufgeteilt werden konnte; siehe Abschnitt 4.2.1.2. Innerhalb dieser Datensatzgruppen wurden anschließend für die Experimente aus DG 1 drei B-Pläne und aus DG 2 und 3 jeweils ein B-Plan ausgewählt, welche(r) die jeweilige(n) Datensatzgruppe repräsentieren; siehe Tabelle 2.

Um den abstrakten Begriff Verstehen konkret einordnen zu können, habe ich drei sogenannte Verständnislevel definiert, welche zur Bewertung und Einstufung eines MLLM dienen. Diese Level wurden anhand der nötigen Arbeitsschritte des Genehmigungsprozesses für ein Bauvorhaben entwickelt, welche ich im Abschnitt 2.2.1.3 erläutert habe. Hierbei stellte ich mir die Frage, wie viel Fachwissen, Daten und Interpretationsfähigkeiten zur Lösung der einzelnen Arbeitsschritte nötig sind, woraus folgende drei Verständnislevel entstanden sind: **Grundlagenforschung**, **Entscheidungsbasis** und **Handlungsempfehlung**; siehe Abschnitt 4.2.1.1. Im Rahmen meiner Masterarbeit liegt der Fokus auf dem Verständnislevel Grundlagenforschung, weshalb ich hierfür weitere Sub-Level definiert habe. Diese geben einen konkreten Rahmen der zu lösenden Aufgaben für das MLLM vor, welche an die Prüfschritte des Prüfungsprogramms im Genehmigungsprozess in Abschnitt 2.2.1.3 angelehnt sind. Zusätzlich unterstützen die Sub-Level, als Bewertungskriterien dabei, das KI-System zu bewerten.

Für die Recherche zu möglichen KI-System Architekturen wurden sechs unstrukturierte Expert:innen-Interviews mit Personen aus den Bereichen Forschung und Wirtschaft

durchgeführt. Die teilnehmenden Expert:innen wurden im Hinblick auf ihre berufliche Rolle und Expertise ausgewählt. Durch die Qualitative Forschungsmethode konnte ich viele Perspektiven und Ideen aus der Praxis sammeln, wie ein KI-System in verschiedenen Reifegraden entworfen werden könnte. Basierend auf den Ergebnissen der qualitativen Recherche entschied ich mich für einen pragmatischen, prototypischen *Prompt-Chaining* Ansatz (Anthropic 2024). Durch das Prompt-Chaining können Aufgaben in mehrere Sub-Aufgaben geteilt werden, wodurch die Prompts jeweils eindeutig definiert und unabhängig von anderen Arbeitsschritten evaluiert werden konnten. Abbildung 27 gibt einen Überblick über die finale Prompt-Chain Architektur.

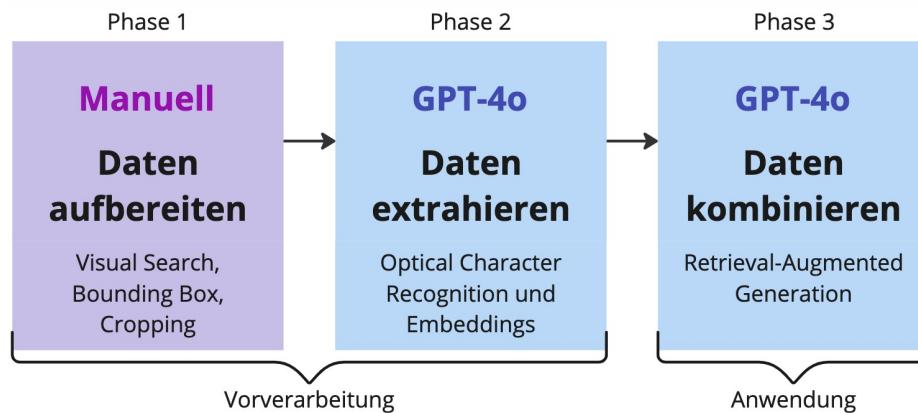


Abbildung 27: Prompt-Chain Architektur des KI-Systems (Eigene Darstellung)

Da in der Masterarbeit der Fokus auf das textliche Verständnis eines B-Plans von MLLMs liegt, wurden die Daten im ersten Schritt manuell vorverarbeitet, sodass sich das Modell auf das Extrahieren und Kombinieren von textlichen Daten konzentrieren konnte. Zusätzlich ist die grafische Vorverarbeitung in Phase 1 laut Visheratin (2024) ein nötiger Arbeitsschritt, da moderne MLLMs hochauflösende Bilder mit hoher Informationsdichte nur sehr begrenzt verarbeiten können, weshalb Methoden wie Visual Search (Wu & Xie 2023) und Visual Cropping angewendet werden müssen, um Details in Bildern erfassen zu können.

Im Rahmen der praktischen Erprobung wurden Experimente zu den einzelnen Hauptteilen eines B-Plans durchgeführt: *Planzeichnung*, *Textteil* und *Zeichenerklärung*. Zu jedem B-Plan-Teil wurden die Vorverarbeitungsphasen 1 und 2 der Prompt-Chain anhand einzelner Aufgaben durchlaufen, sodass in der Anwendungsphase 3 die Ergebnisse via einer vereinfachten Retrieval-Augmented-Generation (=RAG) Anwendung genutzt werden konnten. Die entwickelten Prompts zu den einzelnen Aufgaben repräsentieren eine von vielen möglichen Lösungen. Sie wurden anhand der Prompt Engineering Best Practices (OpenAI 2024a) von OpenAI initialisiert und durch Ausprobieren nach dem „trail and error“-Prinzip optimiert, bis ein zufriedenstellendes Ergebnis erreicht wurde oder sich eine Aufgabe als nicht lösbar herausstellte.

GPT Modelle wie z.B. GPT-4o<sup>10</sup> wurden bisher aufgrund der state-of-the-art Performance als gängige Baseline im Forschungsbereich verwendet. Die Modelle wurden auf großen

<sup>10</sup><https://openai.com/index/hello-gpt-4o>

Datenmengen trainiert und sind durch ihr breites Wissen in der Lage, Aufgaben in einer Zero-Shot-Prompting Manier auszuführen. Hierzu folgte ich der gängigen Praxis und führte die Experimente via Zero-Shot-Prompting basierend auf dem MLLM GPT-4o durch. Zusätzlich steigert der Zero-Shot-Prompting Ansatz die Testbarkeit der generierten Ergebnisse (Choshen et al. 2024). GPT-4o ist wie alle MLLMs nicht deterministisch und kann Informationen erfinden (halluzinieren), wodurch zu einem gleichen Prompt unterschiedliche Antworten generiert werden können. Damit die Antworten möglichst konsistent und getreu der Informationen des B-Plans sind, wurden die Parameter des Modells wie folgt definiert:  $topP = 0$ ,  $seed = 42$ ,  $temperature = 0$ . Diese Einstellung wurde für alle Experimente gleich festgelegt und nicht verändert. Zusätzlich wurde jedes Prompt dreifach ausgeführt, um die Konsistenz der generierten Antworten überprüfen zu können.

Die Ergebnisse der Experimente in Vorverarbeitungsphase 2 wurden mit der Human Evaluation Reference Methode (Cohere 2024b) überprüft. Diese Methode eignete sich als guter Startpunkt, da sich die Evaluation flexibel und individuell nach den Merkmalen des jeweiligen B-Plan richten konnte. Es wäre denkbar mithilfe dieser Methode einen zukünftigen Testdatensatz aufzubauen, zum jetzigen Zeitpunkt dienen die Ergebnisse jedoch nur als Ersteinschätzung, welche für weiterführende Arbeiten verwendet werden können. Während der Evaluation verglich ich jeden generierten Output mit der Grundwahrheit aus den B-Pläne und gab ein Ja/Nein Urteil darüber ab, ob die generierte Antwort eine korrekte und konsistente Antwort darstellte. Die Grundwahrheit für jedes Prompt wurde von mir erarbeitet, dessen Promptqualität sich auf das Bewertungsergebnis auswirkte. Mithilfe der Metriken **Validity (Korrekt)** und **Reliability (Konsistent)** wurde beurteilt, ob die generierten Ergebnisse auf den Informationen in den B-Pläne beruhten und, ob diese verlässlich wiedergegeben wurden. Als Referenz möglicher Metriken diente der Evaluation-Guide von LangSmith (2024).

Der Goldstandard für eine Evaluation ist das Sammeln von Feedback realer Anwendungsnutzer:innen (Cohere 2024a). Zur Evaluation von Anwendungsphase 3 wurde entsprechend ein Expert:innen-Interview mit der Stadt Laichingen durchgeführt, worin die Qualität und die Nützlichkeit der Ergebnisse des KI-Systems anhand des B-Plans L04 (DG1) bewertet wurden. Der B-Plan und die Expert:innen wurden ausgewählt, da mit der Stadt Laichingen bereits in einem Expert:innen-Interview das Prüfungsprogramm zum Baugenehmigungsprozess auf Basis dieses Bauantrags und B-Plans besprochen wurde; siehe Abschnitt 2.2.1.3. Entsprechend konnten die Ergebnisse von GPT-4o optimal mit den Ergebnissen der Expert:innen verglichen und bewertet werden.

## 4.2 Ergebnisse

### 4.2.1 Aufbau der Experimente

#### 4.2.1.1 MLLM Verständnislevel

Auf Basis der Ergebnisse in Kapitel 2.2.1 wurden drei Verständnislevel für MLLM wie folgt definiert:

- **Grundlagenforschung:** Das Level umfasst die Fähigkeit, grundlegende Informationen zu einem Thema in einem B-Plan zu sammeln. Als Datenquelle dient ausschließlich der B-Plan selbst. Zum Beispiel soll das MLLM System in der Lage sein, Werte aus der Nutzungsschablone einer Planzeichnung abzulesen und diese struktu-

riert auszugeben. Hierzu werden zwei Sub-Level definiert: *Informationen extrahieren* (=SL-1) und *Informationen verstehen* (=SL-2). SL-1 umfasst die grundlegende Fähigkeit Informationen aus unstrukturierten B-Pläne extrahieren zu können und ist somit Teil des Use Case „Informationen aus B-Pläne extrahieren“; maßgebend ist eine korrekte Wiedergabe relevanter Informationen aus dem B-Plan:

- **SL-1-A:** Daten aus der Planzeichnung eines B-Plans extrahieren.
- **SL-1-B:** Daten aus der Zeichenerklärung eines B-Plans extrahieren.
- **SL-1-C:** Daten aus dem Textteil eines B-Plans extrahieren.

In SL-2 ist das MLLM System in der Lage, die extrahierten Daten zu verarbeiten, wofür ein grundlegendes Verständnis zu diesen Daten benötigt wird.

- **SL-2-A:** Spezifische Informationen aus den verknüpften Ergebnissen aus SL-1 extrahieren.
- **SL-2-B:** QA Fähigkeiten bzgl. der bekannten Informationen aus SL-1. Zum Beispiel „Ist ein GRZ Wert von 2 laut den Festsetzungen erlaubt?“.

SL-2-B entspricht der höchsten Stufe im Level Grundlagenforschung.

- **Entscheidungsbasis:** Erweitert das Level Grundlagenforschung um die Fähigkeit Querverweise mit einzubeziehen, um eine belastbare Datenbasis bereitzustellen, welche in Entscheidungsprozessen genutzt werden könnte. Entsprechend werden wesentlich mehr Dokumente und Daten als Kontext von dem MLLM mit einbezogen, sodass z.B. Festsetzungen im B-Plan mit zusätzlichen Informationen aus dem geltenden Baurecht vertieft und als Quelle referenziert werden können. Dieses Level wird in der vorliegenden Masterarbeit nicht weiter untersucht.
- **Handlungsempfehlung:** Beschreibt das höchste Verständnislevel und erweitert das Level Entscheidungsbasis um die Fähigkeit, Fachwissen anzuwenden bzw. die Festsetzungen eines B-Plan zu interpretieren. Zum Beispiel kann die Anwendung Entscheidungen lernen und proaktiv Vorschläge zu Fragen bezüglich Ausnahme- oder Befreiungsanträgen generieren, was über die Extraktion und das Verknüpfen von Inhalten hinausgeht. Dieses Level wird in der vorliegenden Masterarbeit nicht weiter untersucht.

#### **4.2.1.2 Datensatz**

Bebauungspläne bestehen in der Regel aus einer Planzeichnung (Teil A), welche Regelungen zur Bodennutzung grafisch festlegt (Wikipedia 2024f). Diese Regelungsinhalte werden durch einen textlichen Teil (Teil B) konkretisiert und ergänzt (§9 BauGB und BauN-VO). Darin werden unter anderem Ziele, Zwecke und wesentliche Auswirkungen des B-Plans begründet und ein Umweltbericht (§9 Abs. 8 i.V.m. §2a BauGB) beigefügt (Wikipedia 2024f). Zusätzlich werden die verwendeten Zeichen und Symbole durch eine Zeichenerklärung (Teil C) auf Basis der PlanZV ergänzt (Wikipedia 2024f).

Da die Geltungsdauer eines B-Plans grundsätzlich nicht begrenzt ist (Fina 2023), sind B-Pläne von 1960 weiterhin rechtskräftig, solange diese geltendem Recht nicht widersprechen (Wikipedia 2024f). Dies führt dazu, dass sich bestehende B-Pläne teilweise sehr deutlich, je nach Erscheinungsjahr, unterscheiden:

- **Gestalterische Merkmale** sind visuelle Elemente, die die Lesbarkeit bestimmen wie z.B. Layout und Schrifttyp.
- **Inhaltliche Merkmale** beziehen sich auf die Auffindbarkeit und Referenzierung relevanter Informationen, wie z.B. Nummerierung, Gliederung und Textumfang.
- **Technische Merkmale** sind formale Aspekte, wie beispielsweise Datentyp und Aufteilung bzw. Anzahl der Dokumente.

Die ausgewählten B-Pläne in Spalte **Dokumente** wurden in den Experimenten verwendet und repräsentieren aufgrund ihrer Merkmale die jeweilige Datensatzgruppe. Die B-Pläne stehen in einer hohen Auflösung bereit und besitzen eine hohe Informationsdichte, so dass erst per Zoom einzelne Informationen leserlich werden. Entsprechend wurden die B-Pläne als Anhang (Kapitel 6) zur Einsicht beigelegt. Die folgende Tabelle 2 zeigt den finalen Datensatz und beschreibt relevante Erkennungsmerkmale hinsichtlich der Experimente pro Datensatzgruppe.

	<b>Gestalterische Merkmale</b>	<b>Inhaltliche Merkmale</b>	<b>Technische Merkmale</b>	<b>Dokumente</b>
<b>Gruppe 1 (=DG1)</b>	<ul style="list-style-type: none"> <li>• Ein Dokument für alle Teile A, B und C</li> <li>• Teilweise sehr große Unterschiede in Layout, Schriftart und Zeichen (wenig Standardisierung)</li> </ul>	<ul style="list-style-type: none"> <li>• Teil B ist auf wenige Festsetzungen reduziert</li> <li>• Inhalt in Teil B ist in nummerierte Themenschwerpunkte gegliedert</li> <li>• Tabellenerklärung der Nutzungsschablone in Teil C variiert</li> <li>• Inhalt in Teil C teilweise thematisch strukturiert</li> <li>• B-Pläne von ca. 1960 bis 2000</li> </ul>	<ul style="list-style-type: none"> <li>• Scan eines analogen B-Plans im PDF Format</li> <li>• Teilweise sind die Dokumente ausgeblendet und dadurch schlecht lesbar</li> <li>• Hohe Auflösung der Dokumente</li> </ul>	<ul style="list-style-type: none"> <li>• L04.pdf</li> <li>• L22.pdf</li> <li>• S03.pdf</li> </ul>
<b>Gruppe 2 (=DG2)</b>	<ul style="list-style-type: none"> <li>• Teil A und C in einem Dokument und Teil B separat</li> <li>• Layout, Zeichen und Schriftart sehr ähnlich (mehr Standardisierung)</li> <li>• Teil C einheitlich zweispaltig: Zeichen links, Definition rechts</li> </ul>	<ul style="list-style-type: none"> <li>• Teil B enthält mehr Informationen und besteht durchschnittlich aus ca. 10 Seiten.</li> <li>• Inhalt in Teil B ist in nummerierte Themenschwerpunkte gegliedert</li> <li>• Tabellenerklärung der Nutzungsschablone in Teil C variiert</li> <li>• Inhalt in Teil C teilweise thematisch strukturiert</li> <li>• B-Pläne ab ca. 2000</li> </ul>	<ul style="list-style-type: none"> <li>• Export/Scan eines DWG/DXF Plans im PDF Format</li> <li>• Scan eines Textdokuments im PDF Format</li> <li>• Daten sind gut lesbar</li> <li>• Hohe Auflösung der Dokumente</li> </ul>	<ul style="list-style-type: none"> <li>• F11pz-ze.pdf</li> <li>• F11-tt.pdf</li> </ul>
<b>Gruppe 3 (=DG3)</b>	<ul style="list-style-type: none"> <li>• Jeder Teil in einem separaten Dokument</li> <li>• Hohes Maß an Standardisierung und einheitlicher Struktur</li> </ul>	<ul style="list-style-type: none"> <li>• Teil B enthält mehr Informationen und besteht aus bis zu 100 Seiten.</li> <li>• Teil B enthält ein nummeriertes Inhaltsverzeichnis</li> <li>• Tabellenerklärung der Nutzungsschablone in Teil C variiert</li> <li>• Inhalt in Teil C thematisch gegliedert</li> <li>• B-Pläne ab ca. 2020</li> </ul>	<ul style="list-style-type: none"> <li>• Teil A, B und C im PDF-Format und zusätzlich in einer XPlanGML Datei</li> <li>• Georeferenzierung</li> <li>• Hohe Auflösung der PDF Dokument</li> <li>• Höchstes Maß an maschineller Lesbarkeit</li> </ul>	<ul style="list-style-type: none"> <li>• BP872A-pz.pdf</li> <li>• BP872A-ze.pdf</li> <li>• BP872A-tt.pdf</li> <li>• BP872A.gml</li> <li>• BP872A.tif/tfw</li> </ul>

Tabelle 2: Datensatz, Gruppierung der Bebauungspläne (Eigene Darstellung)

## 4.2.2 Durchführung der Experimente

Das KI-System ist modular als sogenannte **Prompt-Chain** umgesetzt (siehe Abbildung 27) und via Jupyter Notebooks<sup>11</sup> implementiert und dokumentiert (siehe Anhang (Kapitel 6)).

<sup>11</sup><https://jupyter.org/>

Durch diese Struktur bauen die Experimente aufeinander auf und können reproduziert werden. Die Ordnerstruktur basiert auf dem Cookiecutter Data Science v2<sup>12</sup> Template, eine standardisierte Projektstruktur für die Durchführung und den Austausch von Data Science Projekten.

#### 4.2.2.1 Phase 1: Daten aufbereiten

Die erste Phase umfasst die manuelle grafische Vorverarbeitung der B-Pläne. In einem fortgeschrittenen Reifegrad des KI-Systems könnte diese Phase mit fine-tuned CNN Modellen wie z.B. Faster R-CNN automatisiert werden (Rehkop 2024) oder auch mit Transformer Modellen, sofern genügend Trainingsdaten zur Verfügung stehen. Laut OpenAI hat GPT-4o Schwierigkeiten mit räumlichem Vorstellungsvermögen, was Fähigkeiten mit genauer räumlichen Lokalisierung erfordert (OpenAI 2024b). Entsprechend stellt diese Phase eine wesentliche Unterstützung für GPT-4o dar, die zum Verständnis von B-Pläne maßgeblich beiträgt.

Des Weiteren werden hochauflösende Bilder von GPT-4o automatisch auf die maximal unterstützte Eingabegröße herunterskaliert (OpenAI 2024b). Gerade bei B-Pläne mit hoher Auflösung und Informationsdichte ist das ein Problem, da die Informationen bei zu niedriger Auflösung unleserlich werden. OpenAI bietet hierzu zwei Einstellungen an (OpenAI 2024b):

- **low-res**: Ein Bild mit maximaler Auflösung von  $512 \times 512\text{px}$
- **high-res**: Hierbei wird ein Bild mit maximaler Auflösung von  $768 \times 2000\text{px}$  bzw.  $2000 \times 768\text{px}$  auf  $512 \times 512\text{px}$  herunterskaliert und zusätzlich in mehrere  $512 \times 512\text{px}$  Bildteile zerlegt. Die Technik erinnert an einen *Swin Transformer*, siehe Abbildung 23 im Abschnitt 3.2.2.

Entsprechend wurden die B-Pläne inhaltlich in die Teile **Planzeichnung (=PZ)**, **Zeichenerklärung (=ZE)** und **Textteil (=TT)** zerlegt und auf die maximale unterstützte Bildauflösung skaliert. Dadurch wurden die Themen inhaltlich voneinander getrennt und die Bildauflösung pro Bildausschnitt maximal ausgeschöpft. Testbedingt wurden manche hochauflösenden Bildausschnitte nicht skaliert und in Originalgröße belassen (=raw), um die Effekte der Autoskalierung zu verdeutlichen.

Innerhalb der thematischen Bildausschnitte wurden Informationen zusätzlich mit einer Bounding Box (=bb) markiert, um relevante Informationen hervorzuheben oder zusätzlich via Visual Cropping (=crop) Detailinformationen in einer höheren Auflösung bereitgestellt. Im Projektordner befinden sich unter `./data/raw` die unverarbeiteten Daten und unter `./data/processed` die Ergebnisse des Vorverarbeitungsprozesses. Je nach Vorverarbeitung der Daten wurden diese systematisch anhand der Abkürzungen benannt; z.B. `L04-PZ-bb-crop.png` besteht aus der zugeschnittenen und skalierten Planzeichnung des B-Plan `L04`, worin relevante Inhalte mit einer Bounding Box markiert wurden.

#### 4.2.2.2 Phase 2: Daten extrahieren

In der zweiten Phase wurden Textinformationen aus den B-Pläne extrahiert (OCR) und anschließend persistiert. Hierzu wurde das **GPT-4o** Modell via **OpenAI API**<sup>13</sup> genutzt und mit-

---

<sup>12</sup><https://cookiecutter-data-science.drivendata.org/>

<sup>13</sup><https://python.langchain.com/v0.2/docs/integrations/platforms/openai>

hilfe der Bibliothek *LangChain*<sup>14</sup> implementiert. LangChain ist ein Framework für die Entwicklung von LLM/MLLM Anwendungen, womit die Komplexität der Implementierung des KI-Systems wesentlich reduziert werden konnte. Um die OpenAI API nutzen zu können, ist aufgrund der Datengröße der B-Pläne mindestens ein *Tier 2 Zugang* mit 450000 Tokens/Minute nötig, um nicht in etwaige Rate Limits<sup>15</sup> zu geraten. Zusätzlich wurden die generierten Zwischenergebnisse lokal via der Python Datenbank *storemagic*<sup>16</sup> und als Text-Embeddings in der *Chroma Vektor Datenbank*<sup>17</sup> persistiert. Dadurch konnten die Daten in weiterführenden Prompts wiederverwendet und zusätzlich API Requests eingespart werden, was die Kosten<sup>18</sup> für die API Nutzung senkten und die Ausführungszeit einzelner Prompts beschleunigte. Für die Erzeugung der Text-Embeddings wurde das Modell *text-embedding-3-large*<sup>19</sup> von OpenAI verwendet. Zum derzeitigen Zeitpunkt gab es noch kein Image-Embeddings Modell, sodass nur Textinformationen in der Vektordatenbank gespeichert werden konnten.

Der generelle Aufbau einer Aufgabe (Task) wird im Notebook *./notebooks/hello-world-gpt4o.ipynb* verdeutlicht; siehe Abbildungen 28 und 29.

```
from utils.openai import OpenAI
from utils.runner import Runner

# Initialisierung
instructions = "Du bist ein Assistent zur getreuen Wiedergabe von Informationen aus einem Bebauungsplan. Achte auf Vollständigkeit."
chatGPT = OpenAI(instructions)
runner = Runner()
```

Abbildung 28: Initialisierung des Prompts „Hello World“ (Eigene Darstellung)

```
# Datenquelle
img_path = ".../data/processed/bplaene/1_alles_in_einem_dokument/L04-ZE-TT-crop.png"
pdf_path = ".../data/raw/bplaene/2_zeichnung_textteil_getrennt/F11-PZ-ZE.pdf"

# Aufgabe durchführen
instruction = 'In welchem Jahr ist der vorliegende Bebauungsplan in Kraft getreten? Ausgabe im JSON-Format: {"Gültig seit": <YYYY>}' 
async def run():
    msg1 = await chatGPT.extractTextFromImage(instruction, img_path)
    msg2 = await chatGPT.extractTextFromImage(instruction, pdf_path, "pdf")
    return [msg1, msg2]
results = await runner.async_consistency_check(run)

# Ergebnisse speichern
msg1_l04_year, msg1_f11_year = results[0]
%store msg1_l04_year msg1_f11_year
```

Abbildung 29: Aufbau des Prompts „Hello World“ (Eigene Darstellung)

<sup>14</sup><https://www.langchain.com>

<sup>15</sup><https://platform.openai.com/docs/guides/rate-limits/usage-tiers?context=tier-two>

<sup>16</sup><https://ipython.org/ipython-doc/stable/config/extensions/storemagic.html>

<sup>17</sup><https://python.langchain.com/v0.2/docs/integrations/vectorstores/chroma/>

<sup>18</sup><https://openai.com/api/pricing>

<sup>19</sup><https://platform.openai.com/docs/models/embeddings>

Für die OCR Tasks wurde das Modell mithilfe eines System-Prompts wie folgt initialisiert: „**Du bist ein Assistent zur getreuen Wiedergabe von Informationen aus einem Bebauungsplan. Achte auf Vollständigkeit.**“. Anschließend wurde der Task via User-Prompt formuliert und zusammen mit den decodierten Bildern in den Thread zur Bearbeitung geladen. Abschließend führt der **Runner** den Task dreifach aus und gibt die Ergebnisse aus. Das erste Ergebnis wird jeweils persistiert. Die technischen Implementierungsdetails wurden hinter Utility-Klassen „versteckt“, sodass der Fokus der Prompts auf den Inhalt und die Ergebnisse der jeweiligen Tasks gerichtet ist.

Die Experimente sind in einzelne Jupyter Notebooks aufgeteilt und systematisch im Format **<Aufgabentyp.B-Planteil.Datensatz>** durchnummertiert. Zum Beispiel werden in dem Notebook **<ocr.ze.g1>** Aufgaben vom Typ OCR zu den Zeichenerklärungen im Datensatz 1 oder in dem Notebook **<ocr.pz.g2>** Aufgaben vom Typ OCR zu den Planzeichnungen im Datensatz 2 bearbeitet. Zusätzlich sind die einzelnen Prompts innerhalb der Notebooks thematisch einheitlich indexiert. Zum Beispiel sind unter **Z1.X** Aufgaben definiert, um das Schema einer Nutzungsschablone zu extrahieren, was in der Unteraufgabe **Z1.1** ohne Kontext und Hilfsmittel von GPT-4o gelöst werden soll. Die Tabellen 3, 5 und 4 listen alle **Prompt Ids** mit jeweiliger **Definitionen**, vorkommender **Datentypen** und angewandter **Datensatzgruppe** auf.

Prompt Id	Aufgabe	Datentyp	Datensatzgruppe
Z1.X	Schema der Nutzungsschablone extrahieren: <b>1. Unbearbeitet ohne Kontext</b> <b>2. Bounding Boxen ohne Kontext</b> <b>3. Cropped ohne Kontext</b>	.PNG	1, 2
Z2.X	Erklärungen zu "Art der baulichen Nutzung", "Maß der baulichen Nutzung" und "Bauweise, überbaubare Grundstücksflächen" extrahieren: <b>1. Ohne Kontext</b> <b>2. Mit Kontext (BauNVO)</b>	.PNG, .PDF	1, 2, 3

Tabelle 3: Zeichenerklärung Prompt Ids (Eigene Darstellung)

Prompt Id	Aufgabe	Datentyp	Datensatzgruppe
P1.X	Nutzungsschablonen extrahieren: <b>1. Unbearbeitet ohne Kontext</b> <b>2. Bounding Boxes ohne Kontext</b> <b>3. Cropped ohne Kontext</b>	.PNG	1, 2

Tabelle 4: Planzeichnung Prompt Ids (Eigene Darstellung)

Prompt Id	Aufgabe	Datentyp	Datensatzgruppe
T1.X	Kompletten Textteil extrahieren: <b>1.</b> Ohne Kontext <b>2.</b> Eine Seite pro Thread ohne Kontext	.PNG, .PDF	1, 2, 3
T2.1	Festsetzung zu "Art der baulichen Nutzung" extrahieren – * Mit Kontext (BauNVO) * Similarity Search via Vector Store	.PDF	2, 3
T2.2.X	Festsetzung zu "Art der baulichen Nutzung" extrahieren: <b>1.</b> Seiten extrahieren ohne Kontext <b>2.</b> Seiten extrahieren mit Kontext (BauNVO) <b>3.</b> Vorgefilterte Seiten mit Kontext (BauNVO)	.PDF	3
T2.3	Festsetzung zu "Art der baulichen Nutzung" extrahieren – * Mit Kontext (BauNVO)	.PNG, .PDF	1, 2, 3
T3.1	Festsetzung zu "Maß der baulichen Nutzung" extrahieren – * Mit Kontext (BauNVO) * Similarity Search via Vector Store	.PDF	2, 3
T3.2.X	Festsetzung zu "Maß der baulichen Nutzung" extrahieren: <b>1.</b> Seiten extrahieren ohne Kontext <b>2.</b> Seiten extrahieren mit Kontext (BauNVO) <b>3.</b> Vorgefilterte Seiten mit Kontext (BauNVO)	.PDF	3
T3.3	Festsetzung zu "Maß der baulichen Nutzung" extrahieren – * Mit Kontext (BauNVO)	.PNG, .PDF	1, 2
T4.1	Festsetzung zu "Bauweise, überbaubare Grundstücksflächen" extrahieren – * Mit Kontext (BauNVO) * Similarity Search via Vector Store	.PDF	2, 3
T4.2.X	Festsetzung zu "Bauweise, überbaubare Grundstücksflächen" extrahieren: <b>1.</b> Seiten extrahieren ohne Kontext <b>2.</b> Seiten extrahieren mit Kontext (BauNVO) <b>3.</b> Vorgefilterte Seiten mit Kontext (BauNVO)	.PDF	3
T4.3	Festsetzung zu "Bauweise, überbaubare Grundstücksflächen" extrahieren – * Mit Kontext (BauNVO)	.PNG, .PDF	1, 2
T5.1	Festsetzungen aus B, C und D gesammelt extrahieren – * Mit Kontext (BauNVO) * Similarity Search via Vector Store	.PDF	3
T5.2.X	Festsetzungen aus B, C und D gesammelt extrahieren: <b>1.</b> Seiten extrahieren ohne Kontext <b>2.</b> Seiten extrahieren mit Kontext (BauNVO) <b>3.</b> Vorgefilterte Seiten mit Kontext (BauNVO)	.PDF	3
T5.3	Festsetzungen aus B, C und D gesammelt extrahieren – * Mit Kontext (BauNVO)	.PNG, .PDF	1, 2, 3
T5.4.X	Festsetzungen aus B, C und D gesammelt extrahieren: <b>1.</b> Ohne Kontext <b>2.</b> Eine Seite pro Thread mit Kontext (BauNVO)	.PNG, .PDF	1, 2, 3

Tabelle 5: Textteil Prompt Ids (Eigene Darstellung)

#### 4.2.2.3 Phase 3: Daten kombinieren

In der letzten Phase wurden die extrahierten Daten aus Phase 2 aus der Datenbank gelesen und dem MLLM GPT-4o als Kontext, via Systemnachricht, zur Verfügung gestellt, so dass in einem vereinfachten RAG Workflow der:die User:in in bekannter Question-Answering Interaktion, z.B. in Form eines Chatbots, auf die Daten zugreifen kann. Hierzu wurde das Modell wie folgt initialisiert: „**Du bist ein Chat-Bot der mit Hilfe der vorliegenden Informationen Fragen beantwortet. Referenziere deine Antworten anhand der jeweiligen Kapitelnummer aus der [Zeichenerklärung] und [Textliche Festsetzungen]. Helfen die gegebenen Informationen nicht um eine Antwort zu formulieren, antworte mit [Das ist mir nicht bekannt].**“ Tabelle 6 veranschaulicht den zu betrachteten Ausschnitt der Prüfungsschritte des vorgestellten Prüfungsprogramms aus Abschnitt 2.2.1.3 (Tabelle 1) zum B-Plan L04 in DG1, zu diesem jeweils Prompts entwickelt wurden.

Festsetzungen	Bauantrag	Mensch	entspricht / widerspricht
Art der baulichen Nutzungs	• Wohnhaus	• Nr. 1.1.1 i.V. m. § 1 Abs. 3 BauNVO <b>Allgemeines Wohngebiet (WA)</b> im zeichnerischen Teil des B-Plan • § 4 Abs. 2 Nr. 1 BauNVO Wohngebäude allgemein zulässig	✓
Maß der baulichen Nutzungs	• Lageplan (zeichnerischer und schriftlicher Teil) GRZ 0,17 / GFZ 0,35 • Zahl der Vollgeschosse II	• Nr. 1.1.1 i.V.m. § 1 Abs. 3 BauNVO <b>GRZ 0,4 / GFZ 0,8</b> • Nr. 1.1.3 i.V.m. zeichnerischem Teil <b>Vollgeschosse I</b>	✓ ✗
Bauweise	• Lageplan (zeichnerischer Teil) mit seitlichen Abstandsfächern	• Nr. 1.2 i.V.m. zeichnerischem Teil <b>= offene Bauweise</b>	✓

Tabelle 6: Betrachtete Prüfungsschritte des Prüfungsprogramms im Genehmigungsprozess zum B-Plan L04 in DG1 (Eigene Darstellung. Angelehnt an Kaufmann & Hascher (2024))

Zu jeder Angabe im Bauantrag wurden jeweils zwei Prompts entwickelt:

1. Festsetzungen aus dem B-Plan zu der Angabe bereitstellen.
2. Stellungnahme, ob die Angabe der Festsetzungen im B-Plan widerspricht oder entspricht.

Initial wurden die Prüfungsschritte ohne Lokalisierung des Bauvorhabens umgesetzt. Im Expert:innen-Interview mit der Stadt Laichingen stellten die Expert:innen jedoch fest, dass eine Lokalisierung des Bauvorhabens zu konkreteren Antworten führen könnte (Kaufmann & Hascher 2024), sodass ich im Nachgang ein weiteres Szenario implementierte, welches eine Lokalisierung berücksichtigte.

#### 4.2.3 Evaluation der Ergebnisse

Allgemein konnte ich beobachten, dass die Faktoren – Bildauflösung, Layout, Schrifttyp und Informationsumfang – wesentlich zur Qualität der Ergebnisse beitragen. Im Folgenden werden zu jedem B-Plan-Teil die wesentlichen Erkenntnisse detailliert betrachtet. Die besten Ergebnisse wurden via dicker Schriftstärke in den jeweiligen Tabellen hervorgehoben.

#### 4.2.3.1 Zeichenerklärung

Tabelle 7 zeigt die Ergebnisse zu den Zeichenerklärungen. Das Schema der Nutzungs-schablone (Prompt-Id: Z1.X) konnte für *F11* und *L04* korrekt und konsistent extrahiert wer-den. *BA-872A* wurde zu diesem Fall nicht explizit getestet, da dieser sehr ähnlich zu *F11* ist und entsprechend die Ergebnisse aus *F11* auf *BA-872A* angewendet werden können. Im Rahmen *S03* konnte kein zufriedenstellendes Ergebnis erreicht werden. Je nach Auf-lösung, Informationsdichte und Lesbarkeit konnten bereits mit einer Bounding Box sehr gute Ergebnisse erzielt werden; siehe z.B. Z1.2 / *F11-ZE-bb-crop.png*. Im Fall Z1.2 / L04-*ZE-bb-crop-png* konnte hingegen das Schema noch nicht konsistent korrekt extrahiert werden. Im Rahmen Z1.3 wurde via Visual Cropping die Auflösung des Bildausschnitts des Schemas maximiert, sodass sich auch für *L04* zufriedenstellende Ergebnisse eingestellt hatten. Aus *S03* konnte trotz ausreichender Bildauflösung, aufgrund des komplexen Lay-outs und einen schlechten lesbaren Schrifttyp, keine korrekten Ergebnisse generiert wer-den.

Die Erklärungen zu den drei Haupt-Festsetzungen (Prompt-Id: Z2.X) konnte für die *L04*, *BA-872A* und *F11* korrekt und konsistent extrahiert werden. Zu *S03* konnten wieder auf-grund des komplexen Layouts und Schrifttyps keine zufriedenstellenden Ergebnisse er-ziebt werden. Zusätzlich war in allen Fällen die Erkennung von Symbolen problematisch, da GPT-4o die Symbole ausschließlich mit Schriftzeichen interpretiert bzw. beschreiben kann.

*F11* enthält keine Überschriften, sodass die Gruppierung der Inhalte der Zeichenerklärung tendenziell komplexer war, als zu *L04* und *BA-872A*. Bei Hinzugabe von Hintergrundwissen zu der BauNVO, konnte ich beobachten, dass mehr Informationen als relevant eingestuft wurden, was im Fall Z2.2 / *F11-ZE-crop* zur deutlichen Verschlechterung führte. Des Weite-ren haben sich die Erkenntnisse bezüglich der Bildauflösung und Informationsdichte aus Z1.X bestätigt, was das Ergebnis aus Z2.1 / *F11-ZE* verdeutlicht.

Prompt Id	Dateiname	Datensatzgruppe	Korrekt	Konsistent
Z1.1	L04-ZE-TT-crop.png	1	0/3	1/3
Z1.1	L04-ZE-crop.png	1	0/3	1/3
Z1.1	S03-ZE-two-col-crop.png	1	0/3	2/3
Z1.1	F11-ZE.png	2	0/3	2/3
Z1.1	F11-ZE-crop.png	2	0/3	1/3
Z1.2	L04-ZE-TT-bb-crop.png	1	0/3	3/3
Z1.2	L04-ZE-bb-crop.png	1	1/3	2/3
Z1.2	S03-ZE-two-col-bb-crop.png	1	0/3	3/3
Z1.2	F11-ZE-bb.png	2	1/3	2/3
<b>Z1.2</b>	<b>F11-ZE-bb-crop.png</b>	<b>2</b>	<b>3/3</b>	<b>3/3</b>
<b>Z1.3</b>	<b>L04-ZE-nz-crop.png</b>	<b>1</b>	<b>3/3</b>	<b>3/3</b>
Z1.3	S03-ZE-nz-crop.png	1	0/3	3/3
<b>Z1.3</b>	<b>F11-ZE-nz-crop.png</b>	<b>2</b>	<b>3/3</b>	<b>3/3</b>
<b>Z2.1</b>	<b>L04-ZE.crop.png</b>	<b>1</b>	<b>3/3</b>	<b>3/3</b>
Z2.1	S03-ZE.crop.png	1	0/3	1/3
<b>Z2.1</b>	<b>F11-ZE-crop.png</b>	<b>2</b>	<b>3/3</b>	<b>3/3</b>
Z2.1	F11-ZE.png	2	0/3	2/3
<b>Z2.1</b>	<b>BA-872A-ZE.pdf</b>	<b>3</b>	<b>3/3</b>	<b>3/3</b>
<b>Z2.2</b>	<b>L04-ZE.crop.png</b>	<b>1</b>	<b>3/3</b>	<b>3/3</b>
Z2.2	S03-ZE.crop.png	1	0/3	1/3
Z2.2	F11-ZE-crop.png	2	1/3	2/3
Z2.2	F11-ZE.png	2	0/3	2/3
<b>Z2.2</b>	<b>BA-872A-ZE.pdf</b>	<b>3</b>	<b>3/3</b>	<b>3/3</b>

Tabelle 7: Evaluation der generierten Ergebnisse zu der Zeichenerklärung (Eigene Darstellung)

#### 4.2.3.2 Planzeichnung

Tabelle 8 zeigt die Ergebnisse zu den Planzeichnungen. Die Nutzungsschablonen aus der Planzeichnung konnten korrekt und konsistent aus *L04* und *F11* extrahiert werden, welche repräsentativ für die Datensatzgruppen *DG1* und *DG2* ausgewählt wurden. Da *B-872A* strukturiert im Format XPlanGML vorliegt, können die Informationen über ein Skript vollständig und korrekt extrahiert werden. Die Planzeichnungen sind hochauflösend und bestehen aus einer hohen Informationsdichte, sodass bei dem Herunterskalieren der Bilder selbst ich keine Informationen mehr herauslesen konnte. Dieser Effekt erklärt ebenfalls die schlechten Ergebnisse im Fall *P1.2*. Entsprechend musste die Auflösung der relevanten Informationen maximiert werden, sodass ausschließlich im Fall *P1.3* korrekte und kon-

sistente Ergebnissen erzielt werden konnten. Damit die Werte aus den Nutzungsschablonen ihrer jeweiligen Bezeichnung zugeordnet werden konnten, war die entsprechende Tabellenerklärung aus der Zeichenerklärung als Kontextinformation nötig.

Prompt Id	Dateiname	Datensatzgruppe	Korrekt	Konsistent
P1.1	L04-PZ-crop.png	1	0/3	1/3
P1.1	F11-PZ-crop.png	2	0/3	3/3
P1.2	L04-PZ-bb-crop.png	1	0/3	1/3
P1.2	F11-PZ-bb-crop.png	2	0/3	3/3
<b>P1.3</b>	<b>L04-PZ-nz-crop</b>	<b>1</b>	<b>3/3</b>	<b>3/3</b>
<b>P1.3</b>	<b>F11-PZ-nz-crop</b>	<b>2</b>	<b>3/3</b>	<b>3/3</b>

Tabelle 8: Evaluation der generierten Ergebnisse zu der Planzeichnung (Eigene Darstellung)

#### 4.2.3.3 Textteil

Tabelle 9 zeigt die Ergebnisse zu den Textteilen. Sowohl für *L04* und *L22* im Bildformat als auch *F11* und *BA-872A* im PDF-Format konnten die Inhalte aus dem jeweiligen Textteil korrekt und konsistent extrahiert werden (Prompt-Id T1.X). Speziell zu *F11* und *BA-872A* musste zusätzlich der Umfang der schriftlichen Festsetzungen beachtet werden, da die maximale Tokenanzahl bei GPT-4o bzgl. dem Kontext und Output begrenzt ist. Laut OpenAI Spezifikation<sup>20</sup> gilt aktuell für GPT-4-Turbo eine maximale Output-Länge von 4096 Tokens und eine Kontextgröße von 128000 Tokens. Für GPT-4o ist noch keine konkrete Spezifikation online, sodass die Spezifikation von GPT-4-Turbo als Maßstab genommen wurde. Entsprechend musste der Textteil zu *F11* und *BA-872A* jeweils schrittweise extrahiert werden, um ein vollständiges Ergebnis zu erhalten (Prompt-Id 1.2). Der Output wurde im LaTeX-Format generiert und jeweils mit dem Originallayout verglichen, was nicht zufriedenstellend funktionierte.

Die Experimente zu den Prompt-Ids *T2.X*, *T3.X*, *T4.X* und *T5.X* fokussieren sich inhaltlich auf die drei Haupt-Festsetzungen.

In Datensatzgruppe *DG1(L04, L22)* konnten in den Fällen *T5.3b / L04* und *T2.3, T3.3, T4.4 / L22* teilweise zufriedenstellende Ergebnisse erzielt werden. Im Fall *T.5.3b / L04* musste das Prompt angepasst werden, da die Überschrift zu „Art der baulichen Nutzung“ und „Maß der baulichen Nutzung“ im Textteil zu „Bauliche Nutzung“ zusammengefasst wurde, damit der Output sinnvoll strukturiert generiert werden konnte. Entsprechend war es für GPT-4o nicht möglich die Festsetzung „Art der baulichen Nutzung“ und „Maß der baulichen Nutzung“ einzeln zu extrahieren; siehe *T2.3b, T3.3b, T4.3 / L04*. Des Weiteren enthielt *L22* eine komplexe Tabelle zu der Festsetzung „Maß der baulichen Nutzung“, die in keinen Fällen

<sup>20</sup><https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4>

korrekt extrahiert werden konnte. Allgemein stellte die Festsetzung „Bauweise, überbaubare Grundstücksflächen“ eine inhaltliche Herausforderung dar, sodass hierzu tendenziell zu viele Informationen als relevant extrahiert wurden. Ebenfalls führte die Hinzunahme von Kontextwissen (BauNVO) zu keinen Verbesserungen im Output, sowie hatte ein größerer Umfang der Zeichenerklärungen keinen negativen Einfluss auf das jeweilige Ergebnis.

In Datensatzgruppe **DG2 (F11)** konnten in den Fällen **T2.1, T2.3, T3.1, T3.3** und **T4.3** zufriedenstellende Ergebnisse erzielt werden. **T5.3** und **T5.4.1** führten unabhängig des Kontexts zu unvollständigen Ergebnissen bezüglich der Festsetzung „Art der baulichen Nutzung“ und im Fall **T4.1** wurden zu viele Informationen zur Festsetzung „Bauweise, überbaubare Grundstücksflächen“ extrahiert.

In Datensatzgruppe **DG3 (BA-872A)** wurden aufgrund des Umfanges die meisten Experimente durchgeführt, wovon **T2.2.2, T2.2.3, T2.3.2, T2.3.3, T2.4.2** und **T2.4.3** zufriedenstellende Ergebnisse erzielten. Bei der Extraktion der Seitenzahlen wurden tendenziell zu viele Seiten extrahiert, jedoch waren unter diesen alle relevanten Seiten dabei, ebenfalls waren die Abfragen mit Kontext (BauNVO) deutlich besser im Vergleich zu **T2.2.1, T2.3.1** und **T2.4.1**. Ohne die Vorfilterung der Seiten wurden die Ergebnisse tendenziell unvollständig und enthielten zum Teil irrelevante Informationen; siehe **T3.3** und **T4.3**. Ebenfalls waren die Ergebnisse mit Similarity Search via Vector Store unvollständig, sobald der Inhalt über mehrere Seiten ging, wie z.B. bei der Begründung zu „Maß der baulichen Nutzung“ und „Bauweise, überbaubare Grundstücksflächen“; siehe **T3.1** und **T4.1**. Des Weiteren konnten keine ausreichenden Ergebnisse, unabhängig von der Methode, erzielt werden, sobald alle Themen gesammelt extrahiert werden sollten. Hierzu waren die Ergebnisse zum Teil unvollständig, komprimiert und nicht mehr im gleichen Wortlaut wie im Originaltext; siehe **T5.2.1, T5.2.2, T5.2.3, T5.3, T5.4.1** und **T5.4.2**.

Prompt id	Dateiname	Datensatzgruppe	Korrekt	Konsistent
T1.1	L04-TT-crop.png	1	3/3	3/3
T1.1	L22-TT-crop.[1-3].png	1	3/3	2/3
T1.1	F11-TT.pdf	2	0/3	3/3
T1.1	BA-872A-TT.pdf	3	0/3	3/3
T1.2	F11-TT.pdf	2	3/3	2/3
T1.2	BA-872A-TT.pdf	3	3/3	2/3
T2.1, T3.1, T4.1	F11-TT.pdf	2	3/3, 3/3, 0/3	3/3, 3/3, 3/3
T2.1, T3.1, T4.1	BA-872A-TT.pdf	3	3/3, 0/3, 0/3	3/3, 2/3, 2/3
T5.1	BA-872A-TT.pdf	3	0/3	1/3
T2.2.1, T3.2.1, T4.3.1	BA-872A-TT.pdf	3	0/3, 0/3, 1/3	3/3, 1/3, 1/3
T2.2.2, T3.2.2, T4.3.2	BA-872A-TT.pdf	3	3/3, 3/3, 2/3	2/3, 2/3, 2/3
T2.2.3, T3.2.3, T4.3.3	BA-872A-TT.pdf	3	3/3, 3/3, 3/3	3/3, 2/3, 3/3
T5.2.1	BA-872A-TT.pdf	3	0/3	2/3
T5.2.2	BA-872A-TT.pdf	3	0/3	3/3
T5.2.3	BA-872A-TT.pdf	3	0/3	2/3
T2.3b, T3.3b, T4.3	L04-TT-crop.png	1	0/3, 0/3, 3/3	2/3, 2/3, 2/3
T2.3, T3.3, T4.3	L22-TT-crop.[1-3].png	1	3/3, 0/3, 3/3	2/3, 2/3, 3/3
T2.3, T3.3, T4.3	F11-TT.pdf	2	3/3, 3/3, 3/3	3/3, 2/3, 2/3
T2.3, T3.3, T4.3	BA-872A-TT.pdf	3	3/3, 0/3, 0/3	3/3, 2/3, 2/3
T5.3a	L04-TT-crop.png	1	0/3	2/3
T5.3b	L04-TT-crop.png	1	3/3	3/3
T5.3	L22-TT-crop.[1-3].png	1	0/3	3/3
T5.3	F11-TT.pdf	2	1/3	2/3
T5.3	BA-872A-TT.pdf	3	0/3	-
T5.4.1	L04-TT-crop.png	1	0/3	2/3
T5.4.1	L22-TT-crop.[1-3].png	1	0/3	2/3
T5.4.1	F11-TT.pdf	2	0/3	3/3
T5.4.1	BA-872A-TT.pdf	3	0/3	-
T5.4.2	BA-872A-TT.pdf	3	0/3	-

Tabelle 9: Evaluation der generierten Ergebnisse zu dem Textteil (Eigene Darstellung)

#### 4.2.3.4 Prüfungsschema

Die Ergebnisse aus der Vorverarbeitungsphase 2 zu dem B-Plan *L04* wurden für das RAG Szenario verwendet, welche von Kaufmann & Hascher(2024) analysiert und bewertet wurden. Die extrahierten Werte aus den Nutzungsschablonen aus der Planzeichnung wurden

größtenteils korrekt extrahiert und thematisch den jeweiligen Keys zugeordnet; siehe **P1.3 / L04) in Tabelle 8**(Kaufmann & Hascher 2024). Der Wert „Bauweise“ wurde als Zahl „0“ interpretiert, was in der Originaldatei jedoch ein Dreieckssymbol mit dem Kleinbuchstaben „o“ in der Mitte des Dreiecks ist (Kaufmann & Hascher 2024). Hierzu hätte das Modell mit mehr Hintergrundwissen zumindest den Kleinbuchstaben „o“ ausgeben müssen (Kaufmann & Hascher 2024). Aus der Zeichenerklärung wurden alle wesentlichen Informationen korrekt erkannt und sauber aufgelistet, sowie das Schema der Nutzungsschablone ordentlich als Tabelle wiedergegeben (Kaufmann & Hascher 2024). Wie auch bereits bei der Planzeichnung, machten Symbole teilweise Probleme (Kaufmann & Hascher 2024). Die Festsetzungen aus dem Textteil wurden richtig erkannt (Kaufmann & Hascher 2024). Insgesamt reichte die Qualität der extrahierten Daten als Wissensbasis für die RAG Anwendung aus (Kaufmann & Hascher 2024).

Im zweiten Schritt wurden die Antworten von GPT-4o zu den einzelnen Prüfungsschritten von den Expert:innen bewertet; siehe Tabelle 10. Zur Festsetzung **Art der baulichen Nutzung** konnte GPT-4o korrekte Antworten generieren, welche dem gewünschten Umfang entsprachen (Kaufmann & Hascher 2024). Zur Festsetzung **Maß der baulichen Nutzung** konnte GPT-4o die Inhalte aus der Wissensbasis korrekt kombinieren und referenzieren, sodass die Antworten korrekt generiert wurden (Kaufmann & Hascher 2024). Der Nachteil dieser Antworten war jedoch, dass sich die Inhalte auf alle Gebiete im B-Plan bezogen und die Antworten entsprechend zu allgemein waren (Kaufmann & Hascher 2024). Zur Festsetzung **Bauweise** konnten die relevanten Informationen aus der Wissensbasis von GPT-4o extrahiert und kombiniert werden (Kaufmann & Hascher 2024). Damit die Frage zu der Abstandsfläche jedoch hätte abschließend beantworten können, hätte GPT-4o Informationen zum Bauplan und eine Geo-Referenzierung benötigt, um die Prüfung der Abstandsfläche durchführen zu können (Kaufmann & Hascher 2024). Entsprechend konnte GPT-4o keine Antwort generieren, sondern antwortete mit „Das ist mir nicht bekannt.“, was positiv zu bewerten war, da GPT-4o keine Informationen erfunden hat (Kaufmann & Hascher 2024).

Festsetzungen	Bauantrag	GPT-4o	widerspricht / entspricht	Korrekt	Konsistent
Art der baulichen Nutzungs	<ul style="list-style-type: none"> <li>Wohnhaus</li> </ul>	<ul style="list-style-type: none"> <li>Die erlaubten Gebäudearten sind in den Baugebieten definiert (<b>Kapitel 1.1.1</b>):           <ul style="list-style-type: none"> <li>WR: Reines Wohngebiet (§ 3 BauNVO)</li> <li>WA: Allgemeines Wohngebiet (§ 4 BauNVO)</li> <li>MI: Mischgebiet (§ 6 BauNVO)</li> <li>GE2: Eingeschränktes Gewerbegebiet für nicht wesentlich störende Gewerbebetriebe (§ 8 BauNVO)</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>Ja, ein Wohnhaus ist in den Baugebieten WR (Reines Wohngebiet) und WA (Allgemeines Wohngebiet) erlaubt.</li> </ul> <span style="color: green;">✓</span>	3/3	2/3
Maß der baulichen Nutzungs	<ul style="list-style-type: none"> <li>Lageplan (zeichnerischer und schriftlicher Teil) GRZ 0,17 / GFZ 0,35</li> <li>Zahl der Vollgeschosse II</li> </ul>	<ul style="list-style-type: none"> <li>Die festgesetzten Werte für die Grundflächenzahl (GRZ) und Geschossflächenzahl (GFZ) variieren je nach Baugebiet (<b>Kapitel 1.1.1</b>):           <ul style="list-style-type: none"> <li>GE2: GRZ 0,6, GFZ 1,2</li> <li>WA: GRZ 0,3 oder 0,5, GFZ 0,5 oder 0,8</li> <li>MI: GRZ 0,4, GFZ 0,5</li> <li>WR: GRZ 0,3 oder 0,5, GFZ 0,5 oder keine Angabe</li> </ul> </li> <li>Die Anzahl der erlaubten Vollgeschosse variiert je nach Baugebiet (<b>Kapitel 1.1.3</b>):           <ul style="list-style-type: none"> <li>GE2: II</li> <li>WA: I oder II</li> <li>MI: I oder II</li> <li>WR: I</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>Ja, die Werte GRZ von 0,17 und GFZ von 0,35 liegen unter den maximal erlaubten Werten in allen genannten Baugebieten.</li> <li>Ja, zwei Vollgeschosse sind in den Baugebieten GE2, WA (in einigen Fällen), und MI zulässig.</li> </ul> <span style="color: green;">✓</span> <span style="color: red;">✗</span>	3/3	2/3
Bauweise	<ul style="list-style-type: none"> <li>Lageplan (zeichnerischer Teil) mit seitlichen Abstandsfächern</li> </ul>	<ul style="list-style-type: none"> <li>Die festgesetzten Bauweisen sind (<b>Kapitel 1.2</b>):           <ul style="list-style-type: none"> <li><b>Offene Bauweise:</b> Nur Einzel- und Doppelhäuser zulässig</li> <li><b>Geschlossene Bauweise:</b> Keine spezifische Einschränkung</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>Das ist mir nicht bekannt. Die Informationen über die Einhaltung der seitlichen Abstandsfächern sind nicht in den vorliegenden Daten enthalten.</li> <li>Für Wohngebäude ist in den meisten Fällen die offene Bauweise zulässig. In einigen Gebieten wie WR ist die geschlossene Bauweise zulässig.</li> </ul>	3/3	2/3

Tabelle 10: Evaluation der Prüfungsschritte ohne Lokalisierung (Eigene Darstellung)

Auf Basis des Feedbacks der Expert:innen wurde ein zweites RAG Szenario implementiert, siehe Tabelle 11, in dem GPT-4o die Information zur relevanten Nutzungsschablone für das Bauvorhaben erhielt. Dadurch war GPT-4o nun in der Lage zu der Festsetzung „Maß der baulichen Nutzung“ eindeutige Antworten zu generieren, die im Abgleich mit dem Prüfungsbericht von Kaufmann & Hascher (2024) übereinstimmten; siehe Tabelle 6.

Festsetzungen	Bauantrag	GPT-4o	widerspricht / entspricht	Korrekt	Konsistent
Art der baulichen Nutzungs	• Wohnhaus	• Erlaubt ist ein Gebäude im <b>Allgemeinen Wohngebiet (WA)</b> (§ 4 BauNVO). ( <b>Kapitel 1.1.1</b> )	✓	3/3	3/3
Maß der baulichen Nutzungs	• Lageplan (zeichnerischer und schriftlicher Teil) GRZ 0,17 / GFZ 0,35 • Zahl der Vollgeschosse II	• Die festgesetzte Grundflächenzahl ( <b>GRZ</b> ) ist <b>0,3</b> und die Geschossflächenzahl ( <b>GFZ</b> ) ist <b>0,5</b> . ( <b>Kapitel 1.1.1</b> ) • Es ist <b>ein Vollgeschoss</b> erlaubt. ( <b>Kapitel 1.1.3</b> )	✓ ✗	3/3	3/3
Bauweise	• Lageplan (zeichnerischer Teil) mit seitlichen Abstandsfächern	• Die festgesetzte Bauweise ist die <b>offene Bauweise (O)</b> . ( <b>Kapitel 1.2</b> )	<b>Das ist mir nicht bekannt.</b>	3/3	3/3

Tabelle 11: Evaluation der Prüfungsschritte mit Lokalisierung (Eigene Darstellung)

#### 4.2.4 Interpretation der Ergebnisse

Die Ergebnisse zeigten, dass im Rahmen des Verständnislevel 1 Grundlagenforschung die Aufgaben im Bereich Sub-Level 1, „Informationen extrahieren“, nur teilweise erfüllt werden konnten:

- **SL-1A „Daten aus der Planzeichnung eines B-Plans extrahieren“:**  
Aufgrund der hohen Informationsdichte und Bildauflösung der B-Pläne sind Vorverarbeitungsschritte wie Visual Search und Visual Cropping nötig, sodass alle relevanten Informationen mit ausreichender Auflösung GPT-4o zur Verfügung stehen. Diese Fähigkeiten erfordern ein gezieltes Training des Modells oder könnte laut Rehkop (2024) mithilfe spezialisierter CNN-Modelle effizienter und verlässlicher umgesetzt werden. Ebenfalls gehen beim Extrahieren der Nutzungsschablonen, in Datengruppe DG1 und DG2, die jeweilige Geo-Referenzierung verloren, was laut Schwindling (2024) generell eine Limitierung von KI-Anwendungen mit unstrukturierten B-Pläne darstellt. Des Weiteren sind vorkommende Symbole in den Nutzungsschablonen für GPT-4o problematisch, da das Modell ausschließlich in Textform antworten kann und entsprechend die Symbole „übersetzen“ muss, wozu GPT-4o aktuell das nötige Fachwissen fehlt (Kaufmann & Hascher 2024). Abgesehen von den Einschränkungen konnte GPT-4o die Tabellenstruktur und deren Textinhalte erkennen und die Werte mit der Tabellenerklärung aus der Zeichenerklärung verknüpfen.
- **SL-1B „Daten aus der Zeichenerklärung eines B-Plans extrahieren“:**  
In Datensatz DG1 sind wie in SL-1A bereits beschrieben Vorverarbeitungsschritte

nötig, sodass der Ausschnitt der Zeichenerklärung von GPT-4o analysiert werden konnte. Des Weiteren sind Symbole für GPT-4o, wie in SL-1A erläutert, problematisch. Allgemein fiel es GPT-4o schwer das zweispaltige Layout in den Zeichenerklärungen bestehend aus Symbolen und Texterklärungen zu verstehen, was z.B. beim Extrahieren der Tabellenerklärung der Nutzungsschablone zu Detailfehlern führte. Bei den Erklärungen zu den Haupt-Festsetzungen spielte die Struktur einen entscheidenden Erfolgsfaktor. Waren zum Beispiel Abstände, Überschriften, Nummerierungen und eine seriflose Druckschrift gegeben, konnte GPT-4o sehr gute Ergebnisse erzielen. Bei abnehmender Qualität dieser visuellen Ankerpunkte verschlechterte sich das Ergebnis jedoch signifikant, sodass GPT-4o zum Beispiel Inhalte falsch zuordnete oder wichtige Informationen ignorierte.

- **SL-1C „Daten aus dem Textteil eines B-Plans extrahieren.“:**

In Datensatz DG1 sind wie in SL-1A bereits beschrieben Vorverarbeitungsschritte nötig, sodass der Ausschnitt des Textteils von GPT-4o analysiert werden konnte. Allgemein konnte der Textinhalt aus den jeweiligen Textteilen sehr gut und ohne Halluzinationen von GPT-4o extrahiert werden – unter Beachtung der Limitierungen Output Length und Context Window. Das Layout konnte via GPT-4o jedoch weniger genau reproduziert werden. Im Rahmen der vier Haupt-Festsetzungen spielen wie bei SL-1B visuelle Merkmale eine entscheidende Rolle. Dies verdeutlichen unter anderem die tendenziell schlechteren Ergebnisse via Vector Store, welcher aktuell ausschließlich Text-Embeddings speichern kann, wodurch visuelle Informationen bzw. Beziehungen zwischen den Seiten verloren gehen. Dieser Effekt konnte auch beobachtet werden, wenn GPT-4o die Inhalte via Einzelseiten analysierte. Des Weiteren konnten mit Hinzugabe von minimalem Kontextwissen zu den Haupt-Festsetzungen in den meisten Fällen die Ergebnisse verbessert werden, sodass GPT-4o mehr relevante Inhalte erkannte und die Ergebnisse dadurch vollständiger waren. Ebenfalls stieg bei wachsendem Umfang der Textinhalte der Bedarf eines klaren Fokus bzw. einer eindeutig gerichteten Aufgabe, sodass die Ergebnisse wortgenau und vollständig von GPT-4o wiedergegeben werden konnten.

Auf dieser Basis kann die Unterfrage „*Wie gut ist die Qualität der extrahierten Daten?*“ [SQ1.1] wie folgt beantwortet werden: Die Insights aus SL-1A, SL-1B und SL-1C verdeutlichen, dass GPT-4o in der Lage war Textinformationen in ausreichender Qualität zu extrahieren, was im Rahmen des Expert:innen-Interview mit Kaufmann & Hascher (2024) bestätigt wurde. Die Qualität hängt jedoch sehr stark von der Vorverarbeitung und visuellen Struktur des B-Plan ab. Entsprechend ist anzunehmen, dass GPT-4o Inhalte anhand visueller und textlicher Beziehungen extrahiert, sodass fachliche Zusammenhänge aufgrund mangelndem Domänenwissen übersehen oder falsch interpretiert werden können. Ebenfalls wurden Limitierungen deutlich, wie z.B. die fehlende Geo-Referenzierung der Nutzungsschablonen (Kaufmann & Hascher 2024). Um eine zuverlässige und hohe Qualität zu erreichen, müsste meiner Meinung nach das MLLM für spezifische Aufgaben trainiert werden und Domänenwissen erlernen.

Des Weiteren konnte anhand der Ergebnisse, aus dem Expert:innen-Interview mit Kaufmann & Hascher (2024), im Rahmen des Verständnislevel 1 Grundlagenforschung gezeigt werden, dass Aufgaben im Bereich Sub-Level 2 von (multimodalen) Sprachmodellen wie GPT-4o teilweise sehr gut gelöst werden können:

- **SL-2A „Spezifische Informationen aus den verknüpften Ergebnissen aus SL-1 extrahieren“:**

Die Informationen wurden korrekt von GPT-4o kombiniert, sowie sauber aufbereitet (Kaufmann & Hascher 2024). Ebenfalls wurden die Antworten frei von Halluzinationen im passenden Umfang generiert (Kaufmann & Hascher 2024).

- **SL-2B „QA Fähigkeiten bzgl. der bekannten Informationen aus SL-1“:** GPT-4o war in der Lage, die Angaben aus dem Bauantrag mit den gegebenen Festsetzungen zu vergleichen und inhaltsgetreu zu beantworten (Kaufmann & Hascher 2024). Die jeweiligen Antworten wurden ebenfalls zuverlässig mit Quellen aus dem Kontextwissen referenziert bzw. auf fehlendes Wissen hingewiesen, falls keine Antwort möglich war (Kaufmann & Hascher 2024).

Entsprechend kann die Unterfrage „*Inwiefern kann ein Verständnis zu den extrahierten Daten geschaffen werden?*“ [S01.2] wie folgt beantwortet werden: Aus den Insights von SL-2A und SL-2B lässt sich ableiten, dass große MLLMs wie GPT-4o ein ausreichendes Sprachverständnis besitzen, um domänenspezifische Inhalte korrekt miteinander zu verknüpfen, sowie zielführend mit User:innen-Instruktionen umzugehen. Das Verständnis von MLLMs ist jedoch stark von dem bereitgestellten Fachwissen abhängig, was unter anderem im Prüfungsschritt „Bauweise“, siehe Abschnitt 4.2.2.3, verdeutlicht wurde. Ebenfalls hinterfragte GPT-4o die bereitgestellten Kontextinformationen nicht, was darauf hinweist, dass GPT-4o nur wenig domänenspezifisches Verständnis besitzt. Für ein robustes Verständnis müsste meiner Meinung nach das MLLM grundlegendes Domänenwissen und Retrieval Evaluation-Fähigkeiten (Yan et al. 2024) erlernen.

## 5 Fazit und Ausblick

### 5.1 Beantwortung der Forschungsfragen

Ziel meiner Masterarbeit war es, die Verständnisfähigkeit von MLLMs in der Domäne Bauleitplanung im Hinblick auf B-Pläne zu untersuchen und erproben. Kapitel 4 bietet einen detaillierten Einblick in die Experimente, sowie in die Evaluation und Interpretation der Ergebnisse.

Die Forschungsfrage „*Inwiefern sind MLLMs in der Lage B-Pläne zu verstehen?*“ [RQ1] kann im Rahmen des Use Case „Extrahieren von Informationen aus unstrukturierten B-Pläne“ (siehe Abschnitt 2.2.3), auf Basis der Ergebnisse zu den Unterfragen „Wie gut ist die Qualität der extrahierten Daten?“ [SQ1.1] und „Inwiefern kann ein Verständnis zu den extrahierten Daten geschaffen werden?“ [SQ1.2] (siehe Abschnitt 4.2.4), wie folgt beantwortet werden: Durch die sehr guten Sprachfähigkeiten von GPT-4o, war das Modell in der Lage B-Pläne in Text- und Bildform zu interpretieren, sodass relevante Informationen in ausreichender Qualität extrahiert und korrekt verknüpft werden konnten. Allerdings hing die Qualität der Ergebnisse sehr stark von der Vorverarbeitung und Struktur der B-Pläne ab, damit z.B. fachliche Zusammenhänge und Detailinformationen von GPT-4o erkannt werden konnten. Des Weiteren wurde der Inhalt aus den Vorverarbeitungsschritten von GPT-4o nicht hinterfragt. Dies führt mich zu der Schlussfolgerung, dass GPT-4o nur wenig domänenspezifisches Verständnis besitzt, sodass fachliche Expertise nur einen geringen Einfluss auf die Generierung der Antworten nimmt. Um das Verständnislevel Grundlagenforschung zu meistern, müsste GPT-4o entsprechend für spezifische Aufgaben trainiert werden und grundlegendes Domänenwissen erlernen.

Die Forschungsfrage „*Können MLLMs die Inhalte von maschinenlesbaren Bebauungsplänen mindestens genauso gut verstehen, wie Menschen dies können?*“ [RQ2] kann auf Basis des direkten Vergleichs zwischen dem Vorgehen und den Ergebnissen bzgl. des Prüfschemas im Baugenehmigungsprozess aus dem Interview mit Kaufmann & Hascher (2024) (siehe Abschnitt 2.2.1.3) und den Erkenntnissen aus dem RAG Szenario (siehe Abschnitt 4.2.4) wie folgt beantwortet werden: Zunächst wurden die Prüfschritte vorgegeben, so dass GPT-4o nicht selbst erkennen musste, welche Prüfschritte für die Genehmigung des Bauvorhabens relevant sind. Ebenfalls ist GPT-4o aktuell nicht in der Lage, eine Geo-Referenzierung bzw. einen lagegenauen Vergleich zwischen einem B-Plan und Bauplan durchzuführen, was die Anzahl durchführbarer Prüfschritte innerhalb eines Genehmigungsverfahrens stark limitiert. Die Stärken von GPT-4o sind es, Assoziationen und Beziehungen zwischen Textinhalten zu erkennen. Dadurch können Informationen effizient extrahiert und durchsucht werden. Entsprechend konnte GPT-4o, im Rahmen der minimalen RAG Anwendung, auf Basis des bereitgestellten Kontextwissens sehr gute Ergebnisse erzielen, welche zum Teil der Grundwahrheit von Kaufmann & Hascher (2024) exakt entsprachen. Unter Einbezug der Erkenntnisse aus Forschungsfrage [RQ1] wird deutlich, dass GPT-4o B-Pläne nicht so gut verstehen kann, wie Menschen dies können. Allerdings zeigen die Ergebnisse auch, dass ein Mensch und MLLM als Team B-Pläne womöglich besser verstehen können, als ein Mensch dies ohne Unterstützung von KI könnte.

## 5.2 Anmerkung zu den Forschungsergebnissen

In dieser Masterarbeit wurden für die Durchführung der Interviews, Experimente und Evaluation Qualitative Forschungsmethoden angewendet, weshalb die Forschungsergebnisse einer gewissen Verzerrung unterliegen. Die Insights aus den jeweiligen Interviews in Kapitel 2 und 4 spiegeln einzelne Meinungen und Erfahrungen von Expert:innen wider und bilden entsprechend einen fundierten Einblick, jedoch keine repräsentative Datengrundlage. Des Weiteren wird die Verzerrung durch den konzeptionellen und experimentellen Prozess im angewandten Forschungsteil, durch die Auswahl der untersuchten B-Pläne und implementierten Prompts, weiter verstärkt. Zusätzlich sind die menschlichen Bewertungen der Ergebnisse subjektiv und nur bedingt reproduzierbar.

## 5.3 Weiterführende Forschung

Die Ergebnisse der Forschung machen deutlich, dass viele Anwendungsfälle für KI in der Bauleitplanung existieren, jedoch große MLLMs wie GPT-4o zum jetzigen Zeitpunkt noch nicht für alle Aufgaben geeignet sind und für einen zuverlässigen Einsatz entsprechend trainiert werden müssen. Interessante Ergebnisse konnten mit GPT-4o vor allem erzielt werden, als die Textinformationen klar vorlagen und GPT-4o sich mit Aufgaben zu inhaltlichen Themen und Zusammenhängen beschäftigen konnte. Entsprechend könnte es meiner Meinung nach spannend sein, in weiterführender Arbeit die Einsatzmöglichkeiten von MLLMs zu RAG Szenarien innerhalb der Bauleitplanung zu erforschen.

Für die Aufbereitung der grafischen Daten von B-Pläne möchte ich auf das Interview mit Rehkop (2024) verweisen, wodurch ich einen Einblick gewinnen konnte, wie mit vortrainierten CNN Modellen unstrukturierte Informationen aus B-Pläne zuverlässig und kosten-günstig extrahiert werden können – was ebenfalls in einer weiterführenden Forschungsarbeit genauer untersucht werden könnte.

## 5.4 Schlusswort

Abschließend möchte ich auf die Antwort von Bruch & Bruch (2024) eingehen, die mir auf die Frage „Macht es Sinn den Baugenehmigungsprozess zu automatisieren?“ mit „Ja, so lange sich alle an die Regeln halten“ antworteten. Denn die folgende Interpretation der vermeidlich kurzen Antwort bringt die aktuelle Situation in der Bauleitplanung für mich persönlich auf den Punkt: Digitalisierung allein kann grundlegende Herausforderungen in der Bauleitplanung nicht lösen. Schließlich würden sich dadurch veraltete Regelungen in B-Pläne genauso wenig ändern wie die unterschiedlichen Meinungen und Ziele von Stakeholder:innen. Allerdings ermöglicht die Digitalisierung bzw. der Einsatz von KI, dass Entscheidungen zukünftig datenbasiert und effizient getroffen und B-Pläne entsprechend inhaltlich modernisiert werden können. Dies führt dazu, dass Prozesse beschleunigt werden und nachhaltige Stadtentwicklung, im Sinne der Charta von Leipzig, weiter Fahrt aufnehmen kann.

## 6 Anhang

Die Dokumente können im Github Repository unter folgenden Link eingesehen und heruntergeladen werden: <https://github.com/schwamic/master-thesis-documents>.

### 6.1 Expert:innen-Interviews

- Miro Boards – Expert:innen-Interviews
- Miro Board – Interview Clustering
- Bauplan der Stadt Laichingen
- Praktikumsskript „Öffentliches Baurecht“ der Stadt Laichingen
- Vorlesungsskript „M7 Umfeldplanung (UFP): Klimaneutrale Stadtentwicklung“ von Prof. Dr. rer. nat. Fina

### 6.2 Experimente und Evaluation

- Jupyter Notebooks
- Miro Board – Evaluation
- Datensatz(Verwendete Bebauungspläne)

### 6.3 Präsentationen

- Posterpräsentation am CHIASM Kick-off-Workshop an der THA
- Vorlesung im Studiengang „Digitaler Baumeister“ an der THA
- Tech Talk bei der Firma credium GmbH

### 6.4 Sonstiges

- Exposé zur Masterarbeit
- Ausschreibung der Abschlussarbeit von Prof. Dr. Kratsch

## Glossar

**Affinity Diagram** Ein Affinity Diagram ist eine visuelle Methode um Ideen, Informationen und Beobachtungen aus einer großen Menge an Daten zu organisieren, um Muster zu erkennen und Erkenntnisse zu gewinnen (Foundation 2024c). In der ersten Phase werden die Inhalte anhand von Beziehungen oder Themen gruppiert und markiert. Abschließend werden Erkenntnisse entwickelt (Foundation 2024c). 5

**ALKIS** Das Amtliche Liegenschaftskataster-Informationssystem (=ALKIS) stellt das System zum Nachweis der Geobasisdaten des Liegenschaftskatasters dar (Wikipedia 2024a). Es umfasst die Kartendarstellung sowie die beschreibenden Angaben zu den Flurstücken (u. a. Kennzeichen, Fläche, Nutzungsart) (Wikipedia 2024a). Darüber hinaus werden die Eigentümer in Übereinstimmung mit dem Grundbuch nachrichtlich geführt (Wikipedia 2024a). 17, 18

**Backpropagation** Backpropagation eine Methode zur Gradientenschätzung, die zum Trainieren von Modellen neuronaler Netze verwendet wird. Die Gradientenschätzung wird vom Optimierungsalgorithmus verwendet, um die Aktualisierung der Netzparameter zu berechnen (Wikipedia 2024c). 30, 35

**BART** Bidirectional and Auto-Regressive Transformers ist ein Encoder-Decoder Transformer-basiertes LLM von Meta und kann als Kombination von BERT und GPT interpretiert werden (Lewis et al. 2019). 27

**Bauverwaltungsamt** Teil einer staatlichen, kommunalen, kirchlichen oder einer sonstigen öffentlich-rechtlichen Verwaltung zu Bauangelegenheiten, die mit der Prüfung, Bewilligung/Ablehnung von Bauanträgen betraut ist und Bauabnahmen durchführt (Wikipedia 2023b). 5, 8, 13–15, 17, 23

**BERT** Bidirectional Encoder Representations from Transformers (=BERT), ist ein Encoder-only Transformer-basiertes LLM von Google (Devlin et al. 2019). 27, 33, 34

**BIM** Building Information Modeling (=BIM) ist eine Methode, die die Arbeitsschritte und -prozesse von allen am Bau beteiligten Gewerken automatisieren soll (Werthmann 2022). Via BIM können sämtliche planungsrelevanten Informationen in eine synchronisierte Datenbasis eingespeist, kombiniert und erfasst werden, worauf alle Projektpartner:innen zugreifen können (Werthmann 2022). Grundlage für die Methode ist ein dreidimensionales Gebäudemodell, bestehend aus Daten wie z.B. geometrischen Daten, Kosteninformationen, Terminen und Materialangaben (Werthmann 2022). 17, 18

**Bounding Box** In der Objekterkennung im Bereich Computer Vision werden Bounding Boxen genutzt um den Ort von Objekten anzugeben. Eine Boudning Box ist das kleinste mögliche Rechteck, bei dem das Objekt noch vollständig enthalten ist (Kipp 2024a). 44, 49

**Charta von Leipzig** Die Leipzig-Charta von 2007 und ihr Nachfolgedokument von 2020 betonen zentrale Aspekte einer nachhaltigen, integrierten Stadtentwicklungspolitik, die soziale, ökonomische und ökologische Ziele miteinander verbindet (Wieder et al. 2021). 18, 59

**CLIP** Contrastive Language-Image Pre-training) ist ein Modell von Meta das darauf trainiert wurde, eine einheitliche Vektorsprache zwischen Bild und Text zu lernen (Radford et al. 2021). 26, 36–38

**CNN** Ein Convolutional Neural Network (=CNN) ist ein NN. Es handelt sich um ein von biologischen Prozessen inspiriertes Konzept im Bereich ML. Sie finden Anwendung in zahlreichen Technologien der künstlichen Intelligenz, vornehmlich bei der maschinellen Verarbeitung von Bild- oder Audiodaten (Wikipedia 2024k). 35, 44, 55, 59

**CV** Computer Vision (=CV) besteht aus Methoden zur Erfassung, Verarbeitung, Analyse und zum Verständnis digitaler Bilder sowie zur Extraktion hochdimensionaler Daten aus der realen Welt, um numerische oder symbolische Informationen zu erzeugen.(Wikipedia 2024j). 25

**Design Thinking** Design Thinking ist ein nicht-linearer, iterativer Prozess, um Nutzer:innen zu verstehen, Annahmen in Frage zu stellen, Probleme neu zu definieren und innovative Lösungen zu entwickeln, die anschließend prototypisch erprobt werden. (IxDF 2016). 3

**Double Diamond** Der Double Diamond steht für einen Prozess, bei dem ein Thema zuerst umfassend erforscht wird (divergentes Denken) und anschließend die Informationen strukturiert und zielgerichtete Maßnahmen abgeleitet werden (konvergentes Denken)(Council 2023; Foundation 2024a). Der Prozess besteht aus vier Phasen: Entdecken, Definieren, Entwickeln und Liefern (Council 2023; Foundation 2024a). 3, 5, 67

**Expert:innen-Interview** Das Expert:innen-Interview ist eine qualitative Forschungs methode (Tegan 2024). Hierzu werden Fragen und die Reihenfolge in der sie gestellt werden nicht festgelegt (Tegan 2024). Diese Art von Interviews ist somit sehr flexibel und hilft möglichst viel Wissen zu generieren (Tegan 2024). 3, 4, 17, 39, 41, 48, 56

**Full Fine-Tuning** Die Feinabstimmung ein Ansatz für das Transferlernen, bei dem die Parameter eines vorab trainierten Modells auf neuen Daten optimiert werden (Wikipedia 2024m). 37

**GPT** Generative Pre-Trained Transformers(=GPT)ist ein Decoder-only Transformer-basiertes LLM von OpenAI (Radford & Narasimhan 2018). 27–29, 33, 34, 67

**HMW** How-Might-We (=HMW) ist eine Design Thinking-Methode, in der eine Problemstellung zu einer Frage neu formuliert wird, um anschließend effiziente und zielgerichtete Lösungen entwickeln zu können (Foundation 2024b). HMW ist die Brücke zwischen den Phasen „Definieren“ und „Entwickeln“ im Design Thinking-Prozesses(Foundation 2024b). 5

**HOAI** Die Honorarordnung für Architekten und Ingenieure (=HOAI) ist eine Rechtsverordnung der deutschen Bundesregierung zur Regelung der Honorare für Architekten- und Ingenieurleistungen in Deutschland. Die seit 1. Januar 2021 geltende Fassung regelt die Vergütung der Leistungen von Architekt:innen und Ingenieur:innen, die Planungsleistungen in den Bereichen der Architektur, der Stadtplanung und des Bauwesens erbringen (Wikipedia 2023f). 12, 67

**INSPIRE** INSPIRE ist eine Initiative der europäischen Kommission mit dem Ziel, eine europäische Geodateninfrastruktur für die Zwecke einer gemeinschaftlichen Umweltpolitik zu schaffen (GDI-DE 2024). Die Richtlinie ist am 15. Mai 2007 in Kraft getreten und gilt für alle Mitglieder in Europa (GDI-DE 2024). INSPIRE bezieht sich ausschließlich auf Geodaten die Raum und Umwelt beschreiben. Hierzu gehören z.B. Geografische Namen, Adressen, Verkehrsnetze, Schutzgebiete, Bodenbedeckung oder auch Geologie(GDI-DE 2024). 18

**Lagegenauigkeit** Lagegenauigkeit gibt an wie groß die Abweichung der digital gespeicherten Lagekoordinaten eines Objektes von der Realität ist. Gemessen entweder als (statistische) mittlere Abweichung oder Maximalwert der Abweichung (Auflösung) (Wikipedia 2023d). 22

**LLaMA** Large Language Model Meta AI (=LLaMA) ist ein Decoder-only Transformer-basiertes LLM von Meta (Touvron et al. 2023). 37

**LLaVA** Large Language and Vision Assistant (=LLaVA) ist ein multimodales LLM, das ein allgemeines visuelles und sprachliches Verständnis kombiniert (Liu et al. 2023b). 25, 26, 35–38, 67

**LoRA Fine-Tuning** Low-Rank-Adaption ist eine Technik zur effizienten Feinabstimmung von Modellen. Die Grundidee besteht darin, eine Matrix mit niedriger Dimension zu entwerfen, die dann zur ursprünglichen Matrix hinzugefügt wird, um die Anzahl an Parametern zu reduzieren (Wikipedia 2024m). 37

**ML** Maschinelles Lernen (=ML) entwickelt, untersucht und verwendet statistische Algorithmen, auch Lernalgorithmen genannt. Dazu bildet ein Lernalgorithmus vorgegebene Beispieldaten auf ein mathematisches Modell ab und passt es so an die Beispieldaten an, dass es von ihnen auf neue Fälle verallgemeinern kann (Wikipedia 2024p). 61

**NLP** NLP ist ein Bereich der Linguistik und des maschinellen Lernens, der sich mit dem Verständnis aller Aspekte der menschlichen Sprache befasst. Ziele von NLP sind u.a. einzelne Wörter wie auch dessen Kontext zu verstehen. Zum Beispiel die Identifizierung der grammatischen Bestandteile eines Satzes (z.B. Substantiv, Verb, Adjektiv) oder die Generierung eines neuen Satzes aus einem Eingabetext (z.B. Übersetzen eines Textes in eine andere Sprache, Zusammenfassen eines Textes) (Wikipedia 2023c). 25–27

**NN** Ein neuronales Netz (=NN) ist eine Gruppe miteinander verbundener Einheiten, die als Neuronen bezeichnet werden und Signale aneinander senden (Wikipedia 2024q). 29, 61

**Object Detection** Bei der Objekterkennung handelt es sich um eine Computertechnologie aus dem Bereich der Computer Vision und der Bildverarbeitung, die sich mit der Erkennung von Instanzen semantischer Objekte einer bestimmten Klasse (wie Menschen, Gebäude oder Autos) in digitalen Bildern und Videos beschäftigt (Wikipedia 2023g). 35

**OCR** Optical Character Recognition ist eine maschinelle Methode, um Wörter und Texte in Bildern in maschinell kodierten Text umzuwandeln (Wikipedia 2024s). 25, 36, 38, 44, 46

**OZG** Das Onlinezugangsgesetz (=OZG) wurde am 14. August 2017 erlassen und verpflichtet Bund, Länder und Gemeinden bis spätestens Ende 2022 ihre Verwaltungsleistungen auch elektronisch über Verwaltungsportale anzubieten und diese miteinander zu einem Portalverbund zu verknüpfen (§ 1 OZG) (Wikipedia 2024r). Die OZG führt dadurch zu einem Entwicklungsschub für die Standardisierung im Anwendungsfeld Planen und Bauen (Krause 2022a). 18

**Qualitative Forschungsmethode** Das Ziel der qualitativen Forschung ist es, ein Thema im Detail zu erfassen und zu untersuchen. Diesbezüglich wird keine hohe Anzahl an Teilnehmenden an der Forschung benötigt, sondern qualitativ wertvolle Schlüsse anhand von ausführlichen Ergebnissen gezogen. (Pfeiffer 2023). 3, 40, 59

**Qualitative Inhaltsanalyse** Die qualitative Inhaltsanalyse dient zur systematischen Bearbeitung von Material, z. B. Texten, um die Forschungsfrage einer wissenschaftlichen Arbeit zu beantworten (Pfeiffer 2022). Dabei ist die qualitative Inhaltsanalyse Teil der empirischen Forschung und hilft neue Erkenntnisse zu gewinnen (Pfeiffer 2022). 5

**RAG** RAG ist eine Methode, welche den Kontext eines Prompts um domänen spezifische Daten erweitert, womit das Modell in die Lage versetzt wird fachspezifische konkrete Antworten generieren zu können, statt generische Antworten auf Basis des statischen Trainingswissens (Databricks 2024). 4, 40, 48, 53–55, 58, 59

**Retrieval Evaluation** Retrieval Evaluation misst, wie effektiv der Retriever eines RAG Systems relevante Dokumente identifiziert hat, welche anschließend als Datenquelle zur Generierung einer Antwort genutzt werden (Yan et al. 2024). 57

**Semantic Segmentation** In der digitalen Bildverarbeitung ist die Bildsegmentierung der Prozess der Aufteilung eines digitalen Bildes in mehrere Bildsegmente bzw. Bildobjekte (Wikipedia 2024n). 35

**Similarity Search** Similarity Search ist eine Methode, um nach den ähnlichsten Vektoren basierend auf deren Nähe im Vektorraum zu suchen. Diese Methode kann z.B. in einem Vector Store genutzt werden (Wikipedia 2024t). 52

**Sliding Window** Das Grundprinzip der Sliding Window Technik ist, dass eine kleine Matrix (z.B. 3x3) schrittweise über ein Bild geschoben wird. Erst pixelweise nach rechts, dann am Ende der Zeile zurück auf den Beginn der nächsten Zeile einen Pixel tiefer etc. (Kipp 2024b). 35

**Systematische Literaturrecherche** Mithilfe einer systematischen Literaturrecherche werden wissenschaftliche Veröffentlichungen herausgefiltert, die für eine Forschungsarbeit relevant sind (Solis 2023). 25

**Tensor** Ein Tensor ist ein mathematisches Objekt, das mehrere Werte in einem mehrdimensionalen Raum darstellt und unter anderem zur Beschreibung von Daten im Bereich des maschinellen Lernen verwendet wird. (TensorFlow 2024). 29, 35

**Untere Bauaufsichtsbehörde** Die Untere Bauaufsichtsbehörde kümmert sich um die Einhaltung der öffentlich-rechtlichen Vorschriften auf dem Gebiet des Baurechts. Dementsprechend sind Bauaufsichtsbehörden sowohl für Baugenehmigungsverfahren als auch für Bauordnungsverfahren verantwortlich (Wikipedia 2023a). 13–15

**Use Case** Ein Anwendungsfall (Use Case) bündelt alle möglichen Szenarien, die eintreten können, wenn ein:e Akteur:in versucht, mithilfe des betrachteten Systems ein bestimmtes fachliches Ziel (Business Goal) zu erreichen. Er beschreibt, was inhaltlich bei dem Versuch der Zielerreichung passieren kann und abstrahiert von konkreten technischen Lösungen (Wikipedia 2024b). 3, 5, 23, 39, 42, 58

**Vector Store** Ein Vector Store ist eine Datenbank, die Vektoren zusammen mit anderen Daten speichern kann, welche mit einem Abfragevektor durchsucht werden können, um die meist relevanten Datenbankeinträge zu finden (LangChain 2024). 52, 56

**Visual Cropping** Beim Cropping eines Bildes werden unerwünschte Teile eines Bildes entfernt, um die Komposition zu verbessern, den Fokus auf das Motiv zu legen oder die Größe bzw. Seitenverhältnis zu ändern (Wikipedia 2024l). 40, 44, 49, 55

**Visual Reasoning** Visuelles Denken ist der Prozess der Manipulation des eigenen geistigen Bildes eines Objekts, um zu einer bestimmten Schlussfolgerung zu gelangen (Wikipedia 2021). 38

**Visual Search** Visual Search ist eine Wahrnehmungsaufgabe, die Aufmerksamkeit erfordert und in der Regel ein aktives Absuchen einer visuellen Umgebung nach einem bestimmten Objekt oder Merkmal (dem Ziel) unter anderen Objekten oder Merkmalen (den Ablenkern) beinhaltet (Wikipedia 2023h). 40, 55

**X Bau** XBau ist der Standard für die Kommunikation zwischen den Beteiligten in bauaufsichtlichen Verfahren (Krause 2022b). Er definiert die Strukturen und Inhalte aller Nachrichten, die erforderlich sind, um die Prozesse im jeweiligen Verfahren abzubilden. Durch den Einsatz des XBau-Standards wird die Interoperabilität zwischen den Systemen verschiedener Akteur:innen verbessert und die Produktivität im Genehmigungsprozess gesteigert (Krause 2022b). 17

**XPlanung** XPlanung ist der gesetzlich verbindlich anzuwendende Datenstandard und das Datenaustauschformat für IT-Verfahren, die Planwerke der Raumordnung, Landes- und Regionalplanung, Bauleitplanung und Landschaftsplanung betreffen (Krause 2022a). XPlanung unterstützt den verlustfreien Transfer von Planungsdaten zwischen unterschiedlichen IT-Systemen sowie die internetgestützte Bereitstellung von Planwerken (Krause 2022c). Der Datenstandard wird in der Datenbeschreibungssprache UML mit der Zielsetzung modelliert, die in den relevanten gesetzlichen Grundlagen (BauGB, BauNVO, PlanZV, ROG, BNatSchG und entsprechende Ländergesetze) vorgegebenen Inhalte zu formalisieren und auf vorgegebene Objektklassen abzubilden (Krause 2022c). Aus den UML-Klassendiagrammen wird eine XML-Schema-Datei erzeugt, gegen die XPlanGML-konforme Pläne gültig sein müssen (Krause 2022c). 17, 18

**Zero-Shot-Prompting** Zero-Shot-Prompting bedeutet, dass die Eingabeaufforderung das Modell direkt anweist eine Aufgabe auszuführen, ohne Beispiele oder Demonstrationen (DAIR.AI 2024). 36, 41

## Abkürzungsverzeichnis

**BauGB** Baugesetzbuch. 6, 7, 9, 10, 14–16

**BauNVO** Baunutzungsverordnung. 6, 10, 49, 52

**BPE** Byte Pair Encoding. 29

- B-Plan** Bebauungsplan. 1, 3, 4, 6, 7, 9–17, 19–23, 25, 26, 39–45, 48, 53–56, 58, 59, 67, 68
- CAD** Computer-Aided Design (rechnerunterstütztes Konstruieren). 17
- DG** Datensatzgruppe. 39, 41, 48, 50–52, 55, 56, 68
- FNP** Flächennutzungsplan. 7, 9, 67
- GG** Grundgesetzbuch. 6
- KI** Künstliche Intelligenz. 1, 4, 5, 21–23, 39–41, 43–45, 58, 59, 67
- LBO** Landesbauordnung. 6
- LEP** Landesentwicklungsplan. 7
- LEPROG** Landesentwicklungsprogramm. 7
- LLM** Large Language Model. 25–27, 37, 38, 45
- LpIG** Landesplanungsgesetz. 7
- MLLM** Multimodel Large Language Model. 1, 2, 4, 23, 25, 26, 35, 36, 39–42, 45, 48, 56–59
- PlanZV** Planzeichenverordnung. 42
- ROG** Raumordnungsgesetz. 7
- RP** Regionalplan. 7
- RQ** Research Question. 1, 4, 23, 39, 58
- SQ** Sub Question. 1, 39, 56–58
- THA** Technische Hochschule Augsburg. 18, 21, 60
- ViT** Vision-Transformer. 26, 34–38

# Abbildungsverzeichnis

1	Einarbeitung in die Domäne anhand der Double Diamond Methode (Eigene Darstellung. Angelehnt an Council (2023)) . . . . .	3
2	Überblick der durchgeführten Interviews und eingenommenen Perspektiven (Eigene Darstellung) . . . . .	4
3	Einordnung der Bauleitplanung in das föderale System (Eigene Darstellung) . . . . .	6
4	Einordnung der Bauleitplanung in das öffentliche Baurecht (Eigene Darstellung. Angelehnt an Hascher (2024); Fina (2023)) . . . . .	7
5	Einordnung der Bauleitplanung in die räumliche Gesamtplanung (Eigene Darstellung. Angelehnt an Hascher (2024); Wikipedia (2024e)) . . . . .	8
6	Stakeholder:innen in der Bauleitplanung (Eigene Darstellung. Angelehnt an Kaiser (2024)) . . . . .	8
7	Ausschnitt eines FNP der Stadt München, Fina (2023) . . . . .	9
8	Ausschnitt eines B-Plan der Stadt Augsburg und Erläuterung der Nutzungs schablone (Fina 2023) . . . . .	10
9	Rechtsnatur von Bauleitplänen im beplanten Gebiet (Eigene Darstellung. Angelehnt an Hascher (2024)) . . . . .	11
10	Bauplanungsprozess anhand der HOAI Leistungsphasen (Eigene Darstellung. Angelehnt an Architektenkammer (2021)) . . . . .	12
11	Überblick eines regulären Baugenehmigungsprozesses (Eigene Darstellung. Angelehnt an Hascher (2024)) . . . . .	14
12	Modellbasierter Abgleich der planungsrechtlichen Festsetzungen im CAD Programm (Eigene Darstellung. Angelehnt an Krause (2022b)) . . . . .	17
13	Use Case Diagramm für mögliche KI Anwendungsfälle (Eigene Darstellung) . . . . .	24
14	Das klassische Beispiel <i>König + Frau – Mann ≈ Königin</i> von Mikolov et al. (2013) in 2D (Darstellung von ResearchGate (2024)) . . . . .	27
15	Transformer-Architektur (Eigene Darstellung. Angelehnt an Vaswani et al. (2017)) . . . . .	28
16	Input-Embedding Layer (Eigene Darstellung. Angelehnt an Radford et al. (2019)) . . . . .	29
17	Positional-Encoding Layer (Eigene Darstellung. Angelehnt an Anwar & Sayeed (2022)) . . . . .	31
18	Einfluss von Kontext auf den Token „apple“ via Attention (Darstellung von Serrano (2023)) . . . . .	31
19	Visualisierung von Attention (Eigene Darstellung. Angelehnt an Vaswani et al. (2017) und Serrano (2023)) . . . . .	32
20	<i>Q,K,V</i> Embeddings (Eigene Darstellung. Angelehnt an Alammar (2020)) . . . . .	32
21	Beispiel von Masked-Attention bzgl. GPT-1 (Eigene Darstellung. Angelehnt an Radford & Narasimhan (2018)) . . . . .	33
22	Patch Embeddings (Eigene Darstellung. Angelehnt an Alexey et al. (2020)) . . . . .	34
23	Swin-Transformer (Eigene Darstellung. Angelehnt an Liu et al. (2021)) . . . . .	35
24	Contrastive Pre-Training (Eigene Darstellung. Angelehnt an Radford et al. (2021)) . . . . .	36
25	LLaVA 1.0 Architektur (Eigene Darstellung. Angelehnt an Liu et al. (2023b)) . . . . .	37
26	Skalierung von LLaVA 1.5 für höhere Bildauflösungen (Eigene Darstellung. Angelehnt an Liu et al. (2023a)) . . . . .	38
27	Prompt-Chain Architektur des KI-Systems (Eigene Darstellung) . . . . .	40

28	Initialisierung des Prompts „Hello World“ (Eigene Darstellung) . . . . .	45
29	Aufbau des Prompts „Hello World“ (Eigene Darstellung) . . . . .	45

## Tabellenverzeichnis

1	Prüfschema zur bauplanungsrechtlichen Prüfung der Zulässigkeit nach §30 Abs. 1 und 3 BauGB, Qualifizierter B-Plan (Eigene Darstellung. Angelehnt an Kaufmann & Hascher (2024)) . . . . .	16
2	Datensatz, Gruppierung der Bebauungspläne (Eigene Darstellung) . . . . .	43
3	Zeichenerklärung Prompt Ids (Eigene Darstellung) . . . . .	46
4	Planzeichnung Prompt Ids (Eigene Darstellung) . . . . .	46
5	Textteil Prompt Ids (Eigene Darstellung) . . . . .	47
6	Betrachtete Prüfungsschritte des Prüfungsprogramms im Genehmigungsprozess zum B-Plan L04 in DG1 (Eigene Darstellung. Angelehnt an Kaufmann & Hascher (2024)) . . . . .	48
7	Evaluation der generierten Ergebnisse zu der Zeichenerklärung (Eigene Darstellung) . . . . .	50
8	Evaluation der generierten Ergebnisse zu der Planzeichnung (Eigene Darstellung) . . . . .	51
9	Evaluation der generierten Ergebnisse zu dem Textteil (Eigene Darstellung) . . . . .	53
10	Evaluation der Prüfungsschritte ohne Lokalisierung (Eigene Darstellung) . . . . .	54
11	Evaluation der Prüfungsschritte mit Lokalisierung (Eigene Darstellung) . . . . .	55

## Literatur

- ALAMMAR, J. 2020. The illustrated transformer. <https://jalammar.github.io/illustrated-transformer/> [Abgerufen am 01. Juni 2024].
- ALEXEY, DOSOVITSKIY; BEYER LUCAS; KOLESNIKOV ALEXANDER; WEISSENBORN DIRK; ZHAI XIAOHUA; UNTERTHINER THOMAS; DEHGHANI MOSTAFA; MINDERER MATTHIAS; HEIGOLD GEORG; GELLY SYLVAIN; USZKOREIT JAKOB; und HOULSBY NEIL. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR* abs/2010.11929. URL <https://arxiv.org/abs/2010.11929>.
- ANTHROPIC. 2024. Chain complex prompts for stronger performance. <https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/chain-prompts> [Abgerufen am 25. Juli 2024].
- ANWAR, H., und G. SAYEED. 2022. The storyteller: Computer vision driven context and content generation system - scientific figure on researchgate. [https://www.researchgate.net/figure/Working-of-positional-encoding-in-Transformer-Neural-Networks\\_fig4\\_365934720](https://www.researchgate.net/figure/Working-of-positional-encoding-in-Transformer-Neural-Networks_fig4_365934720) [Abgerufen am 31. Mai 2024].
- ARCHITEKTENKAMMER, BAYERISCHE. 2021. Leistungskatalog. <https://www.byak.de/data/pdfs/Recht/Merkblaetter/M06-HOAI-2021-Leistungskatalog.pdf> [Abgerufen am 28. Mai 2024].

- AUGSBURG, TECHNISCHE HOCHSCHULE. 2024. Forschende der tha unterstützen unternehmen beim ki-wandel. <https://www.tha.de/Informatik/Forschende-der-THA-unterstuetzen-Unternehmen-beim-KI-Wandel.html> [Abgerufen am 10. Mai 2024].
- BAUR, T. 2024. Wie ki planungsverfahren beschleunigen kann. <https://background.tagesspiegel.de/smart-city/wie-ki-planungsverfahren-beschleunigen-kann> [Abgerufen am 17. Mai 2024].
- BLECHER, LUKAS; GUILLEM CUCURULL; THOMAS SCIALOM; und ROBERT STOJNIC. 2023. Nougat: Neural optical understanding for academic documents.
- BRUCH, P., und M. BRUCH. 2024. Interview, siehe Anhang Miro-Board-Interviews.
- CAMACHO-COLLADOS, JOSE, und MOHAMMAD TAHER PILEHVAR. 2018. From word to sense embeddings: A survey on vector representations of meaning.
- CARON, MATHILDE; HUGO TOUVRON; ISHAN MISRA; HERVÉ JÉGOU; JULIEN MAIRAL; PIOTR BOJANOWSKI; und ARMAND JOULIN. 2021. Emerging properties in self-supervised vision transformers.
- CHOSHEN, LESHEM; ARIEL GERA; YOTAM PERLITZ; MICHAL SHMUELI-SCHEUER; und GABRIEL STANOFSKY. 2024. Navigating the modern evaluation landscape: Considerations in benchmarks and frameworks for large language models (LLMs). *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (Irec-coling 2024): Tutorial summaries*, hrsg. von Roman Klinger, Naozaki Okazaki, Nicoletta Calzolari, und Min-Yen Kan. Torino, Italia: ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-tutorials.4>.
- COHERE. 2024a. Evaluating outputs. <https://cohere.com/llmu/evaluating-llm-outputs#introduction> [Abgerufen am 03. August 2024].
- COHERE. 2024b. Human evaluation. <https://cohere.com/llmu/evaluating-llm-outputs#human-evaluation> [Abgerufen am 03. August 2024].
- COUNCIL, DESIGN. 2023. The double diamond. <https://www.designcouncil.org.uk/our-resources/framework-for-innovation/> [Abgerufen am 29. August 2024].
- DAIR.AI. 2024. Zero-shot-prompting reasoning. <https://www.promptingguide.ai/techniques/zeroshot> [Abgerufen am 05. August 2024].
- DATABRICKS. 2024. Retrieval augmented generation. <https://www.databricks.com/de/glossary/retrieval-augmented-generation-rag> [Abgerufen am 03. August 2024].
- DEVLIN, JACOB; MING-WEI CHANG; KENTON LEE; und KRISTINA TOUTANOVA. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.
- FETH, D.; T. JESWEIN; und S. LUDBORZS. 2023. Stand der digitalisierung in der baubranche. <https://www.iese.fraunhofer.de/blog/digitalisierung-baubranche-studie/> [Abgerufen am 29. April 2024].
- FINA, S. 2023. M7 umfeldplanung(ufp): Klimaneutrale stadtentwicklung. Unveröffentlichtes Manuskript im Anhang.

FINA, S. 2024. Interview, siehe Anhang Miro-Board-Interviews.

FOUNDATION, INTERACTION DESIGN. 2024a. 10 insightful design thinking frameworks: A quick overview. [https://www.interaction-design.org/literature/article/design-thinking-a-quick-overview#7.\\_the\\_"double\\_diamond"\\_design\\_process\\_model-design\\_council-7](https://www.interaction-design.org/literature/article/design-thinking-a-quick-overview#7._the_"double_diamond"_design_process_model-design_council-7) [Abgerufen am 29. April 2024].

FOUNDATION, INTERACTION DESIGN. 2024b. How might we (hmw). <https://www.interaction-design.org/literature/topics/how-might-we> [Abgerufen am 29. April 2024].

FOUNDATION, INTERACTION DESIGN. 2024c. What are affinity diagrams? <https://www.interaction-design.org/literature/topics/affinity-diagrams> [Abgerufen am 29. April 2024].

GDI-DE, KOORDINIERUNGSSTELLE. 2024. Inspire-informationsflyer der gdi-de. [https://www.gdi-de.org/download/Flyer\\_INSPIRE\\_de.pdf](https://www.gdi-de.org/download/Flyer_INSPIRE_de.pdf) [Abgerufen am 6. Mai 2024].

GEIRHOS, L. 2024. Interview, siehe Anhang Miro-Board-Interviews.

GISWIKI. 2008. Georeferenzierung. <http://giswiki.org/wiki/Georeferenzierung> [Abgerufen am 12. Mai 2024].

HASCHER, G. 2024. Praktikumsskript: Ausgewählte themen zum Öffentlichen baurecht. Unveröffentlichtes Manuskript im Anhang.

HU, EDWARD J.; YELONG SHEN; PHILLIP WALLIS; ZEYUAN ALLEN-ZHU; YUANZHI LI; SHEAN WANG; LU WANG; und WEIZHU CHEN. 2021. Lora: Low-rank adaptation of large language models.

HUGGINGFACE. 2024. What transformers can do. [https://huggingface.co/docs/transformers/task\\_summary](https://huggingface.co/docs/transformers/task_summary) [Abgerufen am 16. Juni 2024].

IXDF. 2016. What is design thinking (dt)? <https://www.interaction-design.org/literature/topics/design-thinking> [Abgerufen am 05. September 2024].

KAISSER, J. 2024. Interview, siehe Anhang Miro-Board-Interviews.

KARPATHY, A. 2023. Let's build gpt: from scratch, in code, spelled out. <https://www.youtube.com/watch?v=kCc8FmEb1nY&t=3121s> [Abgerufen am 12. Juni 2024].

KAUFMANN, K., und G. HASCHER. 2024. Interview, siehe Anhang Miro-Board-Interviews.

KIM, GEEWOOK; TEAKGYU HONG; MOONBIN YIM; JEONGYEON NAM; JINYOUNG PARK; JINYEONG YIM; WONSEOK HWANG; SANGDOO YUN; DONGYOON HAN; und SEUNGHYUN PARK. 2022. Ocr-free document understanding transformer.

KIPP, M. 2024a. Computer vision. <https://michaelkipp.de/deeplearning/ComputerVision.html#objekterkennung> [Abgerufen am 06. August 2024].

KIPP, M. 2024b. Konvolutionsnetze. <https://michaelkipp.de/deeplearning/Konvolutionsnetze.html#konvolutionsoperator> [Abgerufen am 05. September 2024].

- KRAUSE, K. 2022a. Xleitstelle. <https://xleitstelle.de> [Abgerufen am 06. Mai 2024].
- KRAUSE, K. 2022b. Xplanung. <https://xleitstelle.de/xbau/mehrwert-xbau> [Abgerufen am 06. Mai 2024].
- KRAUSE, K. 2022c. Xplanung. [https://xleitstelle.de/xplanung/ueber\\_xplanung](https://xleitstelle.de/xplanung/ueber_xplanung) [Abgerufen am 06. Mai 2024].
- KÜNSTER, C. 2024. Interview, siehe Anhang Miro-Board-Interviews.
- LANGCHAIN. 2024. Vector store. [https://python.langchain.com/v0.1/docs/modules/data\\_connection/vectorstores/](https://python.langchain.com/v0.1/docs/modules/data_connection/vectorstores/) [Abgerufen am 05. September 2024].
- LANGSMITH. 2024. Rag evaluation summary. <https://docs.smith.langchain.com/concepts/evaluation#rag-evaluation-summary> [Abgerufen am 03. August 2024].
- LEVELL, J., und G. NOVIKOV. 2024. Interview, siehe Anhang Miro-Board-Interviews.
- LEWIS, MIKE; YINHAN LIU; NAMAN GOYAL; MARJAN GHAVVININEJAD; ABDELRAHMAN MOHAMED; OMER LEVY; VES STOYANOV; und LUKE ZETTLEMOYER. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.
- LI, MINGHAO; TENGCHAO LV; JINGYE CHEN; LEI CUI; YIJUAN LU; DINEI FLORENCIO; CHA ZHANG; ZHOIJUN LI; und FURU WEI. 2022. Trocr: Transformer-based optical character recognition with pre-trained models.
- LIU, HAOTIAN; CHUNYUAN LI; YUHENG LI; und YONG JAE LEE. 2023a. Improved baselines with visual instruction tuning.
- LIU, HAOTIAN; CHUNYUAN LI; YUHENG LI; Bo LI; YUANHAN ZHANG; SHENG SHEN; und YONG JAE LEE. 2024. Llava-next: Improved reasoning, ocr, and world knowledge. URL <https://llava-v1.github.io/blog/2024-01-30-llava-next/>.
- LIU, HAOTIAN; CHUNYUAN LI; QINGYANG WU; und YONG JAE LEE. 2023b. Visual instruction tuning.
- LIU, ZE; YUTONG LIN; YUE CAO; HAN Hu; YIXUAN WEI; ZHENG ZHANG; STEPHEN LIN; und BAINING GUO. 2021. Swin transformer: Hierarchical vision transformer using shifted windows.
- MAILE, T. 2024. Interview, siehe Anhang Miro-Board-Interviews.
- MCKINSEY. 2016. Digital europe pushing the frontier, capturing the benefits. <https://www.mckinsey.com/de/~/media/mckinsey/locations/europe%20and%20middle%20east/deutschland/news/presse/2016/2016-06-30/mgi-digital-europe-june-2016.pdf> [Abgerufen am 29. April 2024].
- MIKOLOV, TOMAS; KAI CHEN; GREG CORRADO; und JEFFREY DEAN. 2013. Efficient estimation of word representations in vector space.
- OPENAI. 2024a. Six strategies for getting better results. <https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results> [Abgerufen am 03. August 2024].

- OPENAI. 2024b. Vision. <https://platform.openai.com/docs/guides/vision> [Abgerufen am 28. Juli 2024].
- PETERS, MATTHEW E.; MARK NEUMANN; MOHIT IYER; MATT GARDNER; CHRISTOPHER CLARK; KENTON LEE; und LUKE ZETTLEMOYER. 2018. Deep contextualized word representations.
- PFEIFFER, F. 2022. Qualitative inhaltsanalyse nach mayring in 5 schritten. <https://www.scribbr.de/methodik/qualitative-inhaltsanalyse/> [Abgerufen am 30. April 2024].
- PFEIFFER, F. 2023. Qualitative forschung für die wissenschaftliche arbeit durchführen. <https://www.scribbr.de/methodik/qualitative-forschung> [Abgerufen am 05. September 2024].
- RADFORD, ALEC; JONG WOOK KIM; CHRIS HALLACY; ADITYA RAMESH; GABRIEL GOH; SANDHI-NI AGARWAL; GIRISH SASTRY; AMANDA ASKELL; PAMELA MISHKIN; JACK CLARK; GRETCHEN KRUEGER; und ILYA SUTSKEVER. 2021. Learning transferable visual models from natural language supervision.
- RADFORD, ALEC, und KARTHIK NARASIMHAN. 2018. Improving language understanding by generative pre-training. URL <https://api.semanticscholar.org/CorpusID:49313245>.
- RADFORD, ALEC; JEFF WU; REWON CHILD; DAVID LUAN; DARIO AMODEI; und ILYA SUTSKEVER. 2019. Language models are unsupervised multitask learners. URL <https://api.semanticscholar.org/CorpusID:160025533>.
- REHKOP, M. 2024. Interview, siehe Anhang Miro-Board-Interviews.
- RESEARCHGATE. 2024. Analogien von statischen embeddings. [https://www.researchgate.net/figure/The-classical-king-woman-man-queen-example-of-neural-word-embeddings-in-2D-It\\_fig1\\_332679657](https://www.researchgate.net/figure/The-classical-king-woman-man-queen-example-of-neural-word-embeddings-in-2D-It_fig1_332679657) [Abgerufen am 29. Mai 2024].
- SCHWINDLING, F. 2024. Interview, siehe Anhang Miro-Board-Interviews.
- SERRANO, L.G. 2023. The math behind attention: Keys, queries, and values matrices. [https://www.youtube.com/watch?v=UPtG\\_380q8o&list=PLs8w1Cdi-zvYskDS2icIItfZgxclApVLv&index=3](https://www.youtube.com/watch?v=UPtG_380q8o&list=PLs8w1Cdi-zvYskDS2icIItfZgxclApVLv&index=3) [Abgerufen am 01. Juni 2024].
- SOLIS, T. 2023. Empfehlungen für die systematische literaturrecherche. <https://www.scribbr.de/aufbau-und-gliederung/literaturrecherche/#systematische-literaturrecherche> [Abgerufen am 05. September 2024].
- TEGAN, GEORGE. 2024. Types of interviews in research|guide and examples. <https://www.scribbr.com/methodology/interviews-research/> [Abgerufen am 30. April 2024].
- TENSORFLOW. 2024. Introduction to tensors. <https://www.tensorflow.org/guide/tensor> [Abgerufen am 05. September 2024].
- TOUVRON, HUGO; THIBAUT LAVRIL; GAUTIER IZACARD; XAVIER MARTINET; MARIE-ANNE LACH-AUX; TIMOTHÉE LACROIX; BAPTISTE ROZIÈRE; NAMAN GOYAL; ERIC HAMBRO; FAISAL AZHAR; AURELIEN RODRIGUEZ; ARMAND JOULIN; EDOUARD GRAVE; und GUILLAUME LAMPLE. 2023. Llama: Open and efficient foundation language models.

- VASWANI, ASHISH; NOAM SHAZER; NIKI PARMAR; JAKOB USZKOREIT; LLION JONES; AIDAN N. GOMEZ; LUKASZ KAISER; und ILLIA POLOSUKHIN. 2017. Attention is all you need.
- VISHERATIN, ALEXANDER. 2024. Breaking resolution curse of vision-language models. <https://huggingface.co/blog/visheratin/vlm-resolution-curse> [Abgerufen am 03. August 2024].
- WERTHMANN, C. 2022. Was ist bim? digitales planen einfach erklärt. <https://www.autodesk.com/de/design-make/articles/was-ist-bim> [Abgerufen am 06. Mai 2024].
- WIEDER, S.; H. MAGES; F. EIDNER; BERLIN EINSATEAM; BERLIN STÄDTEBAU UND RAUMORDNUNG E. V.; DEUTSCHER VERBAND FÜR WOHNUNGWESEN; BRANDENBURGISCHE TECHNISCHE UNIVERSITÄT; und COTTBUS-SENFTENBERG. 2021. *Die neue leipzig-charta: Entstehungsprozess und ergebnis*. Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR).
- WIKIPEDIA. 2021. Visual reasoning. [https://en.wikipedia.org/wiki/Visual\\_reasoning](https://en.wikipedia.org/wiki/Visual_reasoning) [Abgerufen am 12. Juni 2024].
- WIKIPEDIA. 2023a. Bauaufsichtsbehörde. <https://de.wikipedia.org/wiki/Bauaufsichtsbehörde> [Abgerufen am 21. Mai 2024].
- WIKIPEDIA. 2023b. Bauverwaltungsamt. [https://de.wikipedia.org/wiki/Bauamt\\_\(Behörde\)](https://de.wikipedia.org/wiki/Bauamt_(Behörde)) [Abgerufen am 21. Mai 2024].
- WIKIPEDIA. 2023c. Computerlinguistik. <https://de.wikipedia.org/wiki/Computerlinguistik> [Abgerufen am 28. Mai 2024].
- WIKIPEDIA. 2023d. Genauigkeit. <https://de.wikipedia.org/wiki/Genauigkeit> [Abgerufen am 17. Mai 2024].
- WIKIPEDIA. 2023e. Geography markup language. [https://de.wikipedia.org/wiki/Geography\\_Markup\\_Language](https://de.wikipedia.org/wiki/Geography_Markup_Language) [Abgerufen am 7. Mai 2024].
- WIKIPEDIA. 2023f. Leistungsphasen nach hoai. [https://de.wikipedia.org/wiki/Leistungsphasen\\_nach\\_HOAI](https://de.wikipedia.org/wiki/Leistungsphasen_nach_HOAI) [Abgerufen am 29. April 2024].
- WIKIPEDIA. 2023g. Optical detection. [https://en.wikipedia.org/wiki/Object\\_detection](https://en.wikipedia.org/wiki/Object_detection) [Abgerufen am 12. Juni 2024].
- WIKIPEDIA. 2023h. Visual search. [https://en.wikipedia.org/wiki/Visual\\_search](https://en.wikipedia.org/wiki/Visual_search) [Abgerufen am 9. September 2024].
- WIKIPEDIA. 2024a. Amtliches liegenschaftskatasterinformationssystem. [https://de.wikipedia.org/wiki/Amtliches\\_Liegenschaftskatasterinformationssystem](https://de.wikipedia.org/wiki/Amtliches_Liegenschaftskatasterinformationssystem) [Abgerufen am 7. Mai 2024].
- WIKIPEDIA. 2024b. Anwendungsfall. <https://de.wikipedia.org/wiki/Anwendungsfall> [Abgerufen am 05. September 2024].
- WIKIPEDIA. 2024c. Backpropagation. <https://en.wikipedia.org/wiki/Backpropagation> [Abgerufen am 14. Juni 2024].

- WIKIPEDIA. 2024d. Bauleitplanung. <https://de.wikipedia.org/wiki/Bauleitplanung> [Abgerufen am 29. April 2024].
- WIKIPEDIA. 2024e. Bauordnungsrecht. <https://de.wikipedia.org/wiki/Bauordnungsrecht> [Abgerufen am 30. August 2024].
- WIKIPEDIA. 2024f. Bebauungsplan (deutschland). [https://de.wikipedia.org/wiki/Bebauungsplan\\_\(Deutschland\)](https://de.wikipedia.org/wiki/Bebauungsplan_(Deutschland)) [Abgerufen am 24. Juli 2024].
- WIKIPEDIA. 2024g. Building information modeling. [https://de.wikipedia.org/wiki/Building\\_Information\\_Modeling](https://de.wikipedia.org/wiki/Building_Information_Modeling) [Abgerufen am 06. Mai 2024].
- WIKIPEDIA. 2024h. Byte pair encoding. [https://en.wikipedia.org/wiki/Byte\\_pair\\_encoding](https://en.wikipedia.org/wiki/Byte_pair_encoding) [Abgerufen am 12. Juni 2024].
- WIKIPEDIA. 2024i. Charta von athen(ciam). [https://de.wikipedia.org/wiki/Charta\\_von\\_Athen\\_\(CIAM\)](https://de.wikipedia.org/wiki/Charta_von_Athen_(CIAM)) [Abgerufen am 7. Mai 2024].
- WIKIPEDIA. 2024j. Computer vision. [https://en.wikipedia.org/wiki/Computer\\_vision](https://en.wikipedia.org/wiki/Computer_vision) [Abgerufen am 12. Juni 2024].
- WIKIPEDIA. 2024k. Convolutional neural network. [https://en.wikipedia.org/wiki/Convolutional\\_neural\\_network](https://en.wikipedia.org/wiki/Convolutional_neural_network) [Abgerufen am 28. Juli 2024].
- WIKIPEDIA. 2024l. Cropping. <https://de.wikipedia.org/wiki/Cropping> [Abgerufen am 05. August 2024].
- WIKIPEDIA. 2024m. Fine-tuning (deep learning). [https://en.wikipedia.org/wiki/Fine-tuning\\_\(deep\\_learning\)](https://en.wikipedia.org/wiki/Fine-tuning_(deep_learning)) [Abgerufen am 12. Juni 2024].
- WIKIPEDIA. 2024n. Image segmentation. [https://en.wikipedia.org/wiki/Image\\_segmentation](https://en.wikipedia.org/wiki/Image_segmentation) [Abgerufen am 12. Juni 2024].
- WIKIPEDIA. 2024o. Large language models. [https://en.wikipedia.org/wiki/Large\\_language\\_model](https://en.wikipedia.org/wiki/Large_language_model) [Abgerufen am 12. Juni 2024].
- WIKIPEDIA. 2024p. Maschinelles lernen. [https://de.wikipedia.org/wiki/Maschinelles\\_Lernen](https://de.wikipedia.org/wiki/Maschinelles_Lernen) [Abgerufen am 8. Mai 2024].
- WIKIPEDIA. 2024q. Neural network. [https://en.wikipedia.org/wiki/Neural\\_network](https://en.wikipedia.org/wiki/Neural_network) [Abgerufen am 14. Juni 2024].
- WIKIPEDIA. 2024r. Onlinezugangsgesetz. <https://de.wikipedia.org/wiki/Onlinezugangsgesetz> [Abgerufen am 6. Mai 2024].
- WIKIPEDIA. 2024s. Optical character recognition. [https://en.wikipedia.org/wiki/Optical\\_character\\_recognition](https://en.wikipedia.org/wiki/Optical_character_recognition) [Abgerufen am 12. Juni 2024].
- WIKIPEDIA. 2024t. Similarity search. [https://en.wikipedia.org/wiki/Similarity\\_search](https://en.wikipedia.org/wiki/Similarity_search) [Abgerufen am 9. September 2024].
- WU, PENGHAO, und SAINING XIE. 2023. V\*: Guided visual search as a core mechanism in multimodal llms. URL <https://arxiv.org/abs/2312.14135>.

YAN, SHI-QI; JIA-CHEN GU; YUN ZHU; und ZHEN-HUA LING. 2024. Corrective retrieval augmented generation. URL <https://arxiv.org/abs/2401.15884>.

YAO, SHUNYU; DIAN YU; JEFFREY ZHAO; IZHAK SHAFRAN; THOMAS L. GRIFFITHS; YUAN CAO; und KARTHIK NARASIMHAN. 2023. Tree of thoughts: Deliberate problem solving with large language models.

ZWICK, S., und M. WAGNER. 2024. Interview, siehe Anhang Miro-Board-Interviews.

