4.5 Halluzinationen, moderne KI und die Arbeitswelt

Leitidee

Ein Sprachmodell schreibt einen Satz, der flüssig klingt und grammatisch einwandfrei ist – und trotzdem ist er falsch. Fachleute nennen das **Halluzination**. Das passiert nicht nur kleinen Modellen, sondern auch großen Chatbots. Mit deinem Wissen über n-Gramm-Modelle kannst du jetzt genau erklären, **warum** das passiert. Danach vergleichst du das einfache n-Gramm-Modell mit modernen Sprachmodellen und überlegst, was das für deinen Umgang mit KI bedeutet. Zum Schluss blickst du in die Arbeitswelt: Wie verändert generative KI Berufe, und welche neuen entstehen?

Lernziele

Nach diesem Abschnitt kannst du

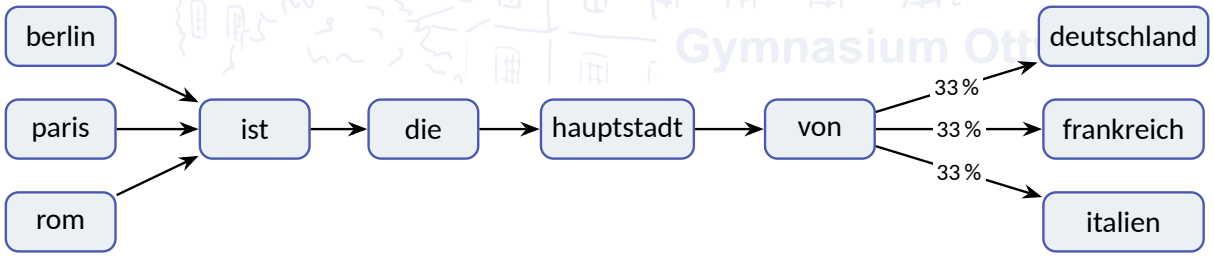
- erklären, was eine Halluzination ist und wie sie bei n-Gramm-Modellen entsteht,
- an einem Beispiel zeigen, warum ein flüssig klingender Satz inhaltlich falsch sein kann,
- Unterschiede zwischen n-Gramm-Modellen und modernen LLMs in Sprachqualität, Korrektheit, Kontextberücksichtigung und Vielseitigkeit beschreiben,
- begründen, warum man Aussagen von KI-Systemen selbst überprüfen muss,
- Auswirkungen generativer KI auf ein Berufsfeld recherchieren und darstellen.

4.5.1 Ein Satz, der nur richtig klingt

Wir trainieren ein Bigramm-Modell mit einem sehr kleinen Korpus aus drei richtigen Sätzen:

```
berlin ist die hauptstadt von deutschland
paris ist die hauptstadt von frankreich
rom ist die hauptstadt von italien
```

Die Bigramm-Tabelle ist schnell aufgestellt. Auf „berlin“, „paris“ und „rom“ folgt jeweils „ist“, darauf „die“, dann „hauptstadt“ und dann „von“. Erst nach „von“ gibt es eine Auswahl:



Bigramm-Modell zum Hauptstadt-Korpus: Erst nach „von“ gibt es drei mögliche Folgewörter. Welche Stadt am Anfang stand, spielt dafür keine Rolle

Jetzt lassen wir das Modell mit dem Startwort „paris“ einen Satz aus sechs Wörtern schreiben, und zwar mit **greedy**. Bis „von“ gibt es keine Wahl: „paris ist die hauptstadt von“. Nach „von“ haben alle drei Länder dieselbe Wahrscheinlichkeit. Greedy nimmt bei Gleichstand das Wort, das im Korpus zuerst vorkommt, also „deutschland“. Das Ergebnis:

> paris ist die hauptstadt von deutschland

Der Satz ist grammatisch perfekt. Jedes Wortpaar darin kommt genau so im Trainingstext vor. Und trotzdem ist die Aussage **falsch**. Mit random wäre es nicht besser: Dann ist der Satz in zwei von drei Fällen falsch, egal mit welcher Stadt er beginnt.

Woran liegt das? Wenn das Modell das Wort nach „von“ auswählt, schaut es **nur auf „von“**. Dass am Anfang „paris“ stand, hat es längst „vergessen“ – es hat es nie beachtet. Das Modell weiß auch nicht, was eine Hauptstadt ist oder was die Wörter bedeuten. Es reiht Wörter nach Wahrscheinlichkeit aneinander. Ob der Satz wahr ist, spielt dabei keine Rolle.

Definition: Halluzination

Eine **Halluzination** ist ein von einem Sprachmodell erzeugter Text, der sprachlich plausibel klingt, aber inhaltlich falsch oder unsinnig ist.

Merke

- **Halluzination** bezeichnet das Phänomen, dass ein Sprachmodell Text erzeugt, der **grammatikalisch bzw. statistisch plausibel**, aber **inhaltlich falsch oder unsinnig** ist.
- Bei n-Gramm-Modellen entsteht Halluzination, weil das Modell **keinerlei „Verständnis“** der Bedeutung hat. Es reiht Wörter nur nach **statistischer Wahrscheinlichkeit** aneinander, unabhängig davon, ob der entstehende Satz sachlich korrekt ist.

Das Wort *Halluzination* kommt eigentlich aus der Medizin. Dort bezeichnet es Wahrnehmungen von Dingen, die gar nicht da sind. Bei Sprachmodellen ist der Begriff ein Bild: Das Modell „erfindet“ etwas, ohne es zu merken. In Wirklichkeit merkt es gar nichts – es rechnet nur.

So kannst du es sagen

So erklärst du eine Halluzination in Fachsprache:

- *„Das Modell hat sich vertan.“ → „Das Modell hat eine Halluzination erzeugt: Der Satz ist grammatisch korrekt, aber inhaltlich falsch.“*
- *„Es hat Paris vergessen.“ → „Das Bigramm-Modell berücksichtigt nur das letzte Wort als Kontext. Deshalb wählt es das Folgewort von ‚von‘ unabhängig vom Satzanfang aus.“*

Arbeitsauftrag 48: Halluzinationen im Hauptstadt-Modell

Verwende das Bigramm-Modell zum Hauptstadt-Korpus.

1. Das Modell erzeugt mit **random** einen Satz aus sechs Wörtern, der mit „rom“ beginnt. Wie groß ist die Wahrscheinlichkeit, dass der Satz inhaltlich richtig ist? Wie groß ist sie, dass er falsch ist?
2. Würde ein **Trigramm-Modell** (Kontext aus zwei Wörtern) das Problem lösen? Begründe.
3. Wie lang müsste der Kontext mindestens sein, damit das Modell immer das richtige Land wählt? Was ist dann der Nachteil?

Meine Antworten:

4.5.2 Halluzinationen im Hund-Katze-Modell

Auch mit dem Hund-Katze-Korpus aus den letzten Abschnitten entstehen leicht seltsame Sätze. Das folgende Programm erzeugt mit random fünf Texte, die mit „die“ beginnen. Es enthält nur die Funktionen, die dafür nötig sind. Die Funktion `generiere_text_random` ist eine kürzere Fassung von `generiere_text` aus dem Abschnitt über Textgenerierung, die immer mit random auswählt.

```python
import random

def folgewoerter(woerter, wort):
    folge = []
    for i in range(len(woerter) - 1):
        if woerter[i] == wort:
            folge.append(woerter[i + 1])
    return folge

def generiere_text_random(woerter, startwort, laenge):
    text = startwort
    aktuelles_wort = startwort
    for schritt in range(laenge - 1):
        folge = folgewoerter(woerter, aktuelles_wort)
        if len(folge) == 0:
            break
        aktuelles_wort = folge[random.randint(0, len(folge) - 1)]
        text = text + " " + aktuelles_wort
    return text

korpus = """
der hund läuft im park der hund spielt mit dem ball
die katze sitzt auf dem baum die katze beobachtet den hund
der hund bellt laut der hund läuft zur katze
die katze springt vom baum die katze rennt weg
der hund läuft ihr hinterher der hund ist müde
"""
woerter = korpus.split()
for nummer in range(5):
    print(generiere_text_random(woerter, "die", 8))
```

AUSGABE

```
die katze sitzt auf dem ball die katze
die katze beobachtet den hund läuft zur katze
die katze springt vom baum die katze rennt
die katze rennt weg der hund ist müde
die katze die katze sitzt auf dem baum
```

Die Ausgabe ist ein **Beispiel**. Bei jedem Start entstehen andere Texte.

- **Zeilen 3–8:** `folgewoerter` sammelt alle Folgewörter eines Wortes, mit Wiederholungen. Diese Liste ist die Lostrommel.
- **Zeilen 10–19:** In jedem Schritt wird ein Los aus der Lostrommel des aktuellen Wortes gezogen und angehängt. Ohne Folgewörter endet der Text.
- **Zeilen 29–30:** Fünf Texte mit je höchstens acht Wörtern werden erzeugt.

Schau dir die erste Zeile an: „die katze sitzt auf dem ball“. Im Korpus sitzt die Katze auf dem **Baum**, und mit dem **Ball** spielt der Hund. Das Modell hat nach „dem“ zwischen „ball“ und „baum“ gelost – beide Wortpaare sind im Korpus richtig, der Satz insgesamt ist eine Erfindung. Auch „den hund läuft zur katze“ ist grammatisch holprig, weil nach „hund“ nicht beachtet wird, dass davor „den“ stand.

Arbeitsauftrag 49: Halluzinationen einordnen

Führe das Programm mehrmals aus. Suche einen erzeugten Satz heraus, der **grammatikalisch plausibel, aber inhaltlich seltsam oder unsinnig** ist. Erkläre in eigenen Worten, **warum** genau dieser Fehler entsteht. Beziehe dich dabei auf die Funktionsweise des n-Gramm-Modells.

Mein Beispielsatz und meine Erklärung:

Versuch: Probiere es aus: Halluzinationen in SoekiaGPT

Wenn deine SoekiaGPT-Version eigene Texte erlaubt, lege eine eigene kleine Textsammlung an, zum Beispiel mit drei oder vier kurzen Steckbriefen von Fantasietieren, die ähnlich aufgebaut sind („der glitzerfuchs lebt im wald und frisst beeren …“). Lass dann Texte erzeugen, die mit dem Namen eines Tieres beginnen. Werden Eigenschaften verschiedener Tiere vermischt? Schau dir in der n-Gramm-Anzeige auch längere n-Gramme an: Welche Wortfolgen stammen aus nur einem einzigen Steckbrief?

4.5.3 n-Gramm-Modelle und moderne LLMs

Moderne Chatbots beruhen auf **Large Language Models (LLMs)**. Sie schätzen ebenfalls Wahrscheinlichkeiten für die Fortsetzung eines Textes, sind aber in vieler Hinsicht viel leistungsfähiger als unser n-Gramm-Modell. Ein LLM wurde mit unvorstellbar viel mehr Text trainiert und betrachtet nicht nur ein oder zwei Wörter, sondern sehr lange Textabschnitte als Kontext. Dadurch „vergisst“ es den Satzanfang nicht so schnell wie unser Modell.

Trotzdem bleibt das Grundprinzip gleich: Ein LLM erzeugt Text, **der wahrscheinlich klingt**. Es schlägt nicht in einer geprüften Faktensammlung nach. Wenn eine Information in den Trainingsdaten selten, widersprüchlich oder gar nicht vorkommt, erzeugt es trotzdem eine flüssige Antwort. Bekannte Beispiele sind erfundene Quellenangaben mit Büchern oder Artikeln, die es nicht gibt, falsche Jahreszahlen oder erfundene Zitate.

Merke

Moderne **LLMs** unterscheiden sich von einfachen n-Gramm-Modellen u. a. in:

- **Sprachqualität:** LLMs erzeugen deutlich flüssigeren, grammatikalisch korrekteren Text.
- **Kontextberücksichtigung:** LLMs können sehr viel längere Kontexte einbeziehen (nicht nur 1–2 vorherige Wörter).

- **Korrektheit:** LLMs erzeugen häufig zuverlässigere Antworten, **können aber weiterhin halluzinieren**. Das n-Gramm-Modell dient hierfür als **vereinfachte Analogie**: Auch LLMs erzeugen Text Wort für Wort nach Wahrscheinlichkeiten und prüfen ihn nicht an Fakten. Die Ursachen bei modernen LLMs sind aber wesentlich komplexer und nicht mit dem einfachen Modell gleichzusetzen.
- **Vielseitigkeit:** LLMs können unterschiedlichste Aufgaben bewältigen (Übersetzen, Zusammenfassen, Programmieren …), während ein einfaches n-Gramm-Modell nur Text im Stil seines Trainingskorpus fortsetzt.

Arbeitsauftrag 50: Vergleichstabelle

Stelle n-Gramm-Modelle und moderne LLMs in einer Tabelle gegenüber. Beschreibe für jedes Kriterium in eigenen Worten, was du beobachtet hast, und nenne – wo möglich – ein Beispiel aus diesem Kapitel.

Kriterium	n-Gramm-Modell	modernes LLM
Sprachqualität		
Korrektheit		
Kontextberücksichtigung		
Vielseitigkeit		

4.5.4 Fakten selbst prüfen

Weil auch große Sprachmodelle halluzinieren können, gilt: **Was ein KI-System schreibt, ist ein Vorschlag, keine geprüfte Tatsache.** Gerade weil die Texte so überzeugend klingen, fallen Fehler leicht nicht auf. Das ist auch für unsere Gesellschaft wichtig: Falschinformationen (**Desinformation**) verbreiten sich schnell, und wer Aussagen nicht prüft, kann leichter getäuscht oder manipuliert werden. In einer Demokratie ist es wichtig, dass sich Menschen ihre Meinung auf Grundlage zutreffender Informationen bilden.

So gehst du vor, wenn du eine Aussage eines KI-Systems verwenden willst:

- **Unabhängig prüfen:** Suche die Aussage in mindestens einer seriösen, unabhängigen Quelle (z. B. Lehrbuch, Fachlexikon, Website einer Behörde oder Hochschule, Qualitätszeitung).
- **Quellen nachschlagen:** Nennt das System Quellen, prüfe, ob es sie wirklich gibt und ob dort tatsächlich das steht, was behauptet wird.
- **Besonders vorsichtig sein** bei Zahlen, Jahreszahlen, Namen, Zitaten und Literaturangaben. Hier entstehen Halluzinationen besonders leicht.
- **Aktualität beachten:** Ein Sprachmodell selbst kennt nur, was in seinen Trainingsdaten stand. Über neuere Ereignisse weiß ein Chatbot nur etwas, wenn er zusätzlich im Internet sucht – und auch dann können Fehler entstehen.

Merke

KI-Systeme erzeugen Texte nach Wahrscheinlichkeit, nicht nach Wahrheit. Wichtige Aussagen muss man deshalb **selbst mit seriösen Quellen überprüfen** – das schützt vor Irrtümern, Desinformation und Manipulation.

Arbeitsauftrag 51: Stellung nehmen

Deine Mitschülerin sagt: „Was der Chatbot schreibt, stimmt bestimmt. Er klingt doch total sicher und hat sogar Quellen angegeben.“

Nimm Stellung zu dieser Aussage. Verwende dabei, was du über Halluzinationen weißt.

Meine Stellungnahme:

4.5.5 Berufliche Orientierung: Generative KI in der Arbeitswelt

Sprachmodelle und andere **generative KI-Systeme** (Systeme, die Texte, Bilder, Musik oder Programmcode erzeugen) werden inzwischen in vielen Berufen eingesetzt. Manche Tätigkeiten werden dadurch schneller oder einfacher, manche fallen weg, und es entstehen neue Aufgaben – zum Beispiel das geschickte Formulieren von Anweisungen an eine KI, das **Prompt Engineering**, oder das Prüfen und Überarbeiten von KI-Ergebnissen.

Wie stark ein Beruf betroffen ist, lässt sich nicht pauschal sagen. Deshalb recherchierst du selbst, am besten zu zweit oder zu dritt.

Arbeitsauftrag 52: Generative KI und Berufe

Recherchiert die Auswirkungen generativer KI-Systeme auf **ein** Berufsfeld eurer Wahl, zum Beispiel:

- Journalismus und Erstellen von Inhalten für Websites und soziale Medien,
- Softwareentwicklung,
- Grafikdesign und Illustration,
- Kundenservice und Support,
- Übersetzen und Dolmetschen.

Leitfragen für eure Recherche:

1. Welche Tätigkeiten dieses Berufsfeldes werden bereits heute durch generative KI unterstützt oder verändert?

2. Entstehen durch generative KI **neue** Aufgaben oder sogar neue Berufsbilder in diesem Bereich (z. B. Prompt Engineering)?
3. Welche Tätigkeiten lassen sich aus heutiger Sicht **nicht** oder nur schwer durch KI ersetzen – und warum?
4. Wie bewertet ihr diese Entwicklung? Welche Chancen und welche Risiken seht ihr?
5. Findet (falls möglich) eine **reale Person** mit Bezug zu diesem Berufsfeld und recherchiert ihre Berufsbiografie (Ausbildung, erste Stelle, heutige Tätigkeit), z. B. über ein Interview, eine Unternehmenswebsite oder ein Karriereportal.

Haltet eure Ergebnisse in Stichpunkten in der Ergebnisvorlage unten fest und bereitet einen kurzen mündlichen Beitrag von etwa einer Minute vor.

Tipp

Seriöse Quellen für die Recherche sind zum Beispiel: das Berufsinformationsportal **BERUFENET** und die Seite **planet-beruf.de** der Bundesagentur für Arbeit, das Berufsinformationszentrum (BIZ) der Agentur für Arbeit, Berufsverbände und Kammern (z. B. IHK, Handwerkskammer), Hochschulen sowie Berichte und Interviews in Qualitätsmedien.

Achte bei jeder Quelle darauf, **wer** sie veröffentlicht hat und **wann**. Die Entwicklung ist schnell, ältere Texte können überholt sein. Übernimm Zahlen (z. B. „Anteil der Jobs, die …“) nur, wenn du die Quelle nennen kannst, und notiere, worauf sich die Zahl bezieht. Wenn du einen Chatbot zur Recherche nutzt, prüfe seine Aussagen wie im Abschnitt *Fakten selbst prüfen* beschrieben.

Ergebnisvorlage: Leitfrage	**unsere Ergebnisse (Stichpunkte)**
Berufsfeld	
1. bereits veränderte oder unterstützte Tätigkeiten	
2. neue Aufgaben und Berufsbilder	
3. (noch) nicht ersetzbare Tätigkeiten und warum	
4. Chancen	
4. Risiken	
5. Berufsbiografie einer realen Person	
Quellen (Herausgeber, Titel, Datum)	

Notizen für den mündlichen Beitrag:

4.5.6 Weitere Übungen

Hier kannst du das Gelernte weiter üben. Die Punkte im Titel zeigen, wie schwer eine Aufgabe ist: ●○○ leicht, ●●○ mittel, ●●● knifflig. Fang mit den leichten an.

Mehrere Aufträge verwenden ein Bigramm-Modell, das mit diesem Korpus aus zwei Steckbriefen trainiert wurde:

> lina wohnt in ottweiler und spielt handball tom wohnt in neunkirchen und spielt schach

Arbeitsauftrag 53: Halluzination oder nicht? ●○○

Das Modell hat die folgenden Sätze erzeugt. Entscheide jeweils, ob der Satz zu den Steckbriefen im Korpus passt oder ob er eine **Halluzination** ist.

Nr.	erzeugter Satz	passt / Halluzination
1	lina wohnt in ottweiler und spielt handball	
2	tom wohnt in ottweiler und spielt schach	
3	tom wohnt in neunkirchen und spielt schach	
4	lina wohnt in neunkirchen und spielt handball	
5	tom wohnt in neunkirchen und spielt handball	

Arbeitsauftrag 54: Richtig oder falsch? ●○○

Entscheide, ob die Aussage richtig oder falsch ist. Trage „r“ oder „f“ ein. Verbessere die falschen Aussagen unten.

Nr.	Aussage	r oder f?
1	Eine Halluzination ist ein Text, der plausibel klingt, aber inhaltlich falsch oder unsinnig ist.	
2	Ein n-Gramm-Modell halluziniert, weil es absichtlich falsche Dinge schreibt.	
3	Moderne LLMs können nicht mehr halluzinieren.	
4	LLMs können viel längere Kontexte einbeziehen als n-Gramm-Modelle.	
5	Ein einfaches n-Gramm-Modell setzt nur Text im Stil seines Trainingskorpus fort.	
6	Mit dem n-Gramm-Modell sind die Ursachen von Halluzinationen bei LLMs vollständig erklärt.	

Verbesserungen:

Arbeitsauftrag 55: Kriterien zuordnen ●○○

Ordne jede Beobachtung einem der vier Kriterien zu: **Sprachqualität**, **Kontextberücksichtigung**, **Korrektheit** oder **Vielseitigkeit**.

Nr.	Beobachtung	Kriterium
1	Ein LLM übersetzt einen Text ins Englische und fasst ihn danach zusammen.	
2	Das Bigramm-Modell achtet bei der Wahl des nächsten Wortes nur auf das letzte Wort.	
3	Ein LLM schreibt flüssige, grammatisch richtige Sätze.	
4	Ein LLM nennt eine falsche Jahreszahl, obwohl die Antwort sehr sicher klingt.	
5	Das n-Gramm-Modell erzeugt „den hund läuft zur katze“.	
6	Ein LLM bezieht sich in seiner Antwort auf eine Frage, die viel weiter oben im Gespräch stand.	

Arbeitsauftrag 56: Halluzinationen erklären ●●○

Verwende das Bigramm-Modell zu den Steckbriefen.

1. Bestimme die Folgewörter von „in“ und von „spielt“ mit ihren Wahrscheinlichkeiten.
2. Das Modell erzeugt mit **random** einen Text aus **7 Wörtern**, der mit „tom“ beginnt. Nenne alle Texte, die dabei entstehen können.
3. Wie viele dieser Texte passen zum Korpus, wie viele sind Halluzinationen?
4. Erkläre mit der Funktionsweise des Bigramm-Modells, warum die Halluzinationen entstehen.

Meine Antworten:

Arbeitsauftrag 57: Kundenservice und KI ●●○

Im Kundenservice eines Online-Shops fallen diese Tätigkeiten an:

a) häufige Standardfragen im Chat beantworten (z. B. „Wo ist mein Paket?“),
b) Antworten auf E-Mails vorformulieren,
c) eine sehr verärgerte Kundin am Telefon beruhigen und eine faire Lösung mit ihr finden,
d) eingehende Anfragen nach Themen sortieren,
e) entscheiden, ob ein Kunde in einem Sonderfall sein Geld zurückbekommt.

1. Ordne jede Tätigkeit zu: „wird schon heute durch generative KI unterstützt“ oder „aus heutiger Sicht schwer zu ersetzen“. Begründe die zweite Gruppe.
2. Nenne eine **neue** Aufgabe, die in diesem Berufsfeld durch generative KI entsteht.

Meine Zuordnung und Begründung:

Arbeitsauftrag 58: Hilft ein längerer Kontext? ●●●

Verwende wieder den Korpus mit den Steckbriefen.

1. Ein **Trigramm-Modell** (Kontext aus zwei Wörtern) erzeugt einen Text, der mit „tom wohnt“ beginnt. Kann dabei noch „tom wohnt in ottweiler“ entstehen? Begründe.
2. Bestimme, wie viele Wörter der Kontext **mindestens** umfassen muss, damit das Modell ab „tom wohnt in“ nur noch den richtigen Steckbrief erzeugen kann. Begründe für den Wohnort **und** für das Hobby.
3. Beschreibe, welchen Nachteil ein so langer Kontext hat.

Gymnasium Ottweiler

Meine Antworten:

4.5.7 Das Wichtigste in Kürze

- Eine **Halluzination** ist ein Text, der sprachlich plausibel klingt, aber inhaltlich falsch oder unsinnig ist.
- Beim n-Gramm-Modell entsteht sie, weil nur die letzten ein bis zwei Wörter als Kontext zählen und die Bedeutung keine Rolle spielt: Jedes Wortpaar stammt aus dem Korpus, die Kombination kann trotzdem falsch sein. Längere n-Gramme helfen nur, bis das Modell bloß noch abschreibt.
- **LLMs** sind besser in Sprachqualität, Kontextberücksichtigung und Vielseitigkeit und meist zuverlässiger – sie halluzinieren aber aus demselben Grund.
- Aussagen von KI-Systemen muss man **selbst mit seriösen Quellen prüfen**. Das schützt vor Desinformation und Manipulation.
- Generative KI verändert Berufe: Tätigkeiten werden unterstützt oder fallen weg, neue Aufgaben entstehen.

Selbstcheck

- ☐ Ich kann erklären, was eine Halluzination ist.
- ☐ Ich kann an einem n-Gramm-Beispiel zeigen, warum ein flüssiger Satz falsch sein kann.
- ☐ Ich kann begründen, warum ein längerer Kontext hilft, aber auch Nachteile hat.
- ☐ Ich kann n-Gramm-Modelle und LLMs in vier Kriterien vergleichen.
- ☐ Ich kann beschreiben, wie ich eine Aussage eines KI-Systems überprüfe.

- ☐ Ich kann zu einem Berufsfeld recherchieren, wie generative KI es verändert, und meine Quellen angeben.