

KI für die automatische Erhebung, Annotation und Analyse sprachlicher Daten

(Hands-On Session)

05.12.2025

Workshop der AG Digitale Romanistik

KI statt Crowd?

*Neue Perspektiven auf linguistischen
Sprachdatenerhebung in der Romanistik*

Iris Ferrazzo

everything you say can and
will be used to train a large
language model

LLM

(noun)

1. Like a normal
model,
but cooler.

Überblick

1. Einführung

- a. Large Language Models (LLMs)
- b. LLMs al Crowd

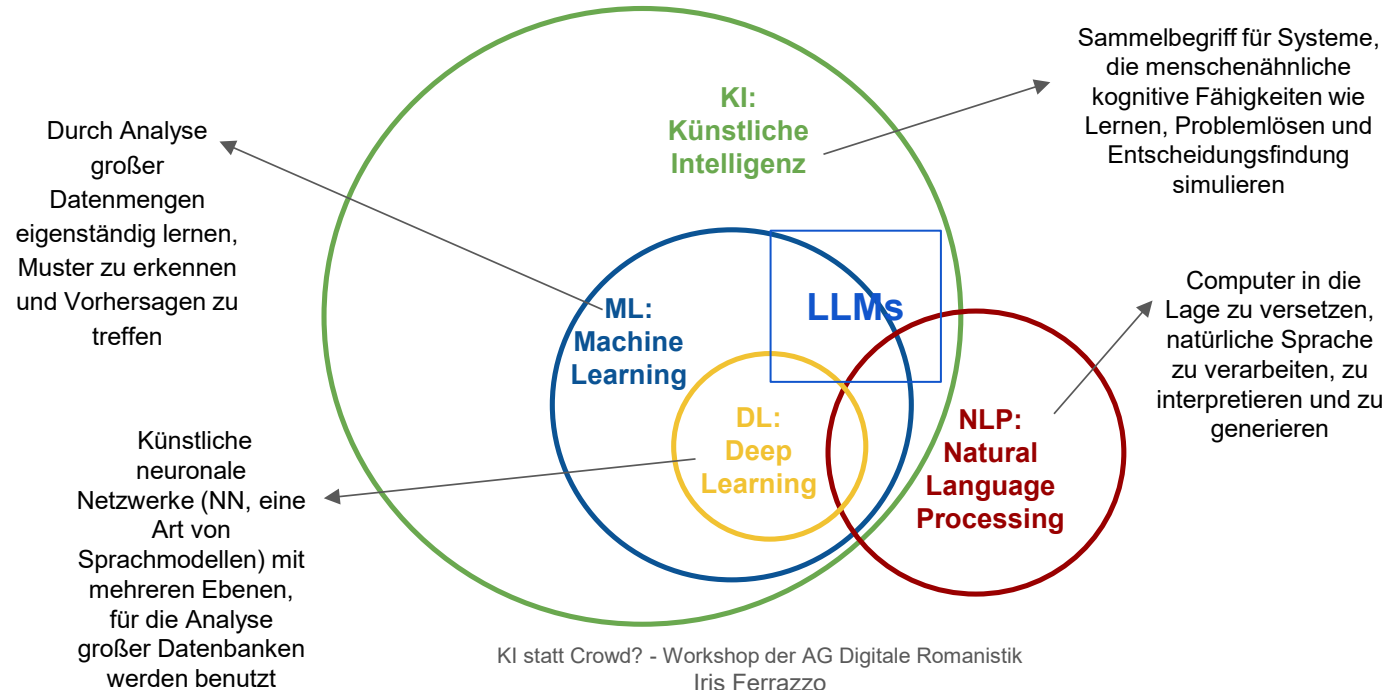
2. Fragestellung: Werden LLMs die nächsten Crowdsourcers?

3. Hands On: Replikation von linguistischen Studien aus der Romanistik mithilfe von LLMs

4. Diskussion

1. ~~83~~

Large Language Models (LLMs)



KI statt Crowd? - Workshop der AG Digitale Romanistik
Iris Ferrazzo

1. Large Language Models (LLMs)

Größe des Modells → Large LMs:

Definition: Sprachmodelle, die auf sehr umfangreichen Datensätzen beruhen und komplexe Modellarchitekturen besitzen.

Skalierbarkeit: Werden auf Milliarden von Wörtern trainiert.

Kontextverständnis: Fähigkeit, längere und komplexere Kontexte zu erkennen und zu verarbeiten.

Anwendung: z. B. ChatGPT.

Sprachmodell (Language Model, LM):

Definition: Mathematisches oder statistisches Modell zur Analyse und Erzeugung von Sprache.

Funktion: Lernend aus Textdaten, um Wortfolgen zu verarbeiten und vorherzusagen.

Anwendung: Kontextbasierte Vorhersage des „nächsten Wortes“.

1. LLMs

Wie erzeugen Sprachmodelle Sprache?

→ **Grundprinzip:** Vorhersage des „nächsten Wortes“

1. Automatische Vektorisierung der Eingabe (Word Embedding)

“Le vacanze iniziano tra 20 giorni. Mi sento molto”

wird in eine Zahlenkette umgewandelt. Jedes Wort = einen eigenen Vektor, alle Vektoren werden zu einer gemeinsamen Eingaberepräsentation verkettet.

2. Weitergabe an das Modell

An der numerischen Eingabe werden mathematische Berechnungen durchgeführt.

3. Kontextuelle Aufmerksamkeit

Das Modell erkennt: Das Wort *molto*“ signalisiert, dass ein Adjektiv folgen sollte, das beschreibt, wie sich die sprechende Person fühlt.

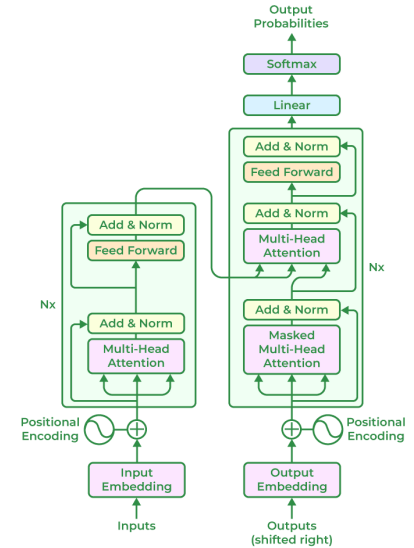
4. Generierung des nächsten Wortes

Das Modell berechnet eine **Wahrscheinlichkeitsverteilung** für mögliche nächste Wörter
z. B.: felice“: 0.35, stanco“: 0.25, ansioso“: 0.15, ...

und wählt das wahrscheinlichste aus → *Le vacanze iniziano tra 20 giorni. Mi sento molto felice”*

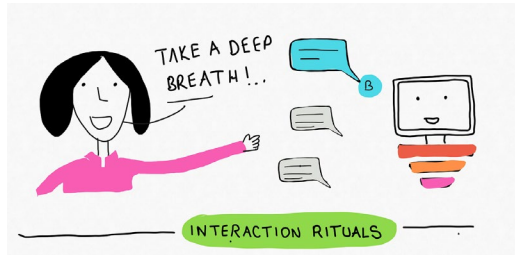
KI statt Crowd? - Workshop der AG Digitale Romanistik

Iris Ferrazzo



FastText Vektor von cat : [1.104e-01 -1.541e-01 -1.086e-01 -1.242e-01 -2.360e-02 4.170e-02
-1.600e-01 1.257e-01 6.200e-03 5.200e-03 7.130e-02 -8.560e-02
-3.310e-02 -1.061e-01 6.310e-02 8.580e-02 -6.880e-02 -1.872e-01
-1.921e-01 -4.430e-02 1.238e-01 8.470e-02 1.510e-02 -1.016e-01
1.078e-01 -3.378e-01 -1.830e-02 -2.300e-03 1.385e-01 2.811e-01
3.580e-02 -1.095e-01 -5.200e-02 -4.510e-02 -8.230e-02 -5.870e-02
-5.080e-02 3.360e-02 4.780e-02 -8.560e-02 -3.710e-02 -6.360e-02
-4.530e-02 -5.530e-02 -8.700e-03 1.100e-03 -9.480e-02 7.740e-02
-9.890e-02 -1.195e-01 5.280e-02 2.230e-02 5.860e-02 1.014e-01
-2.968e-01 -1.293e-01 8.500e-03 1.045e-01 9.190e-02 2.762e-01
1.334e-01 6.790e-02 -3.320e-02 8.440e-02 3.510e-02 -1.570e-02
1.330e-01 1.224e-01 -1.300e-03 6.780e-02 1.593e-01 -1.047e-01
2.313e-01 -1.860e-02 -1.269e-01 2.360e-02 2.085e-01 -5.310e-02
-5.670e-02 3.810e-02 1.478e-01 -1.169e-01 -9.000e-03 4.490e-02
6.690e-02 -6.350e-02 -2.034e-01 1.390e-02 -1.159e-01 -5.000e-02
-1.159e-01 -8.850e-02 1.897e-01 -1.472e-01 5.300e-02 -1.090e-02
3.230e-02 -5.820e-02 1.267e-01 6.830e-02 -3.990e-01 -1.301e-01
-2.400e-03 2.270e-02 8.050e-02 -3.720e-02 8.040e-02 1.398e-01
6.810e-02 7.110e-02 2.820e-02 -7.800e-03 -1.700e-02 -2.540e-02
1.667e-01 1.164e-01 1.319e-01 -1.370e-02 9.820e-02 -1.226e-01
4.600e-02 9.140e-02 4.650e-02 -3.400e-03 -1.321e-01 -6.810e-02
-6.400e-02 -3.430e-02 -1.860e-02 -2.298e-01 9.180e-02 4.030e-02
7.400e-03 -6.210e-02 -1.250e-01 5.170e-02 6.920e-02 6.180e-02
-1.824e-01 -7.520e-02 7.470e-02 -9.600e-03 -8.300e-02 5.010e-02
-3.300e-03 1.356e-01 -2.571e-01 -5.500e-03 -4.000e-04 1.770e-02
-9.000e-02 7.860e-02 -1.539e-01 -2.910e-02 -1.590e-01 -4.860e-02
-9.770e-02 3.000e-04 7.350e-02 -1.083e-01 1.013e-01 1.546e-01
1.884e-01 -1.140e-01 1.580e-01 -1.472e-01 1.708e-01 -9.320e-02
-7.860e-02 1.231e-01 1.484e-01 -4.430e-02 -2.649e-01 1.768e-01
7.340e-02 3.600e-02 -5.020e-02 -2.186e-01 -3.850e-02 -7.610e-02
1.390e-01 9.510e-02 -1.950e-02 -4.370e-02 1.781e-01 9.800e-02
1.584e-01 -2.370e-02 -7.110e-02 1.850e-02 -5.130e-02 -1.293e-01
7.880e-02 -5.580e-02 3.080e-02 8.300e-02 4.680e-02 -5.200e-03
-3.490e-02 -7.230e-02 3.224e-01 -1.200e-03 1.292e-01 -3.780e-02
1.746e-01 1.000e-01 1.247e-01 2.330e-02 8.710e-02 -4.490e-02
-1.680e-02 2.800e-02 -9.330e-02 -1.230e-01 -6.300e-02 1.300e-01
1.340e-02 1.140e-02 1.174e-01 5.450e-02 -8.090e-02 9.880e-02
-6.700e-03 -1.170e-01 -1.174e-01 3.190e-02 -1.320e-01 -4.500e-02
-2.650e-02 -1.245e-01 2.410e-02 6.060e-02 2.710e-02 -3.680e-02
2.457e-01 8.300e-02 -6.780e-02 -7.120e-02 -4.530e-02 2.870e-02
4.080e-02 -3.920e-02 4.990e-02 9.090e-02 4.125e-01 2.440e-02
1.600e-01 8.300e-02 -1.780e-02 1.367e-01 -1.509e-01 -5.780e-02
3.390e-02 -1.300e-02 -4.760e-02 9.010e-02 4.670e-02 3.900e-02
-6.860e-02 -2.770e-02 2.000e-04 -4.880e-02 1.268e-01 -1.300e-03
5.320e-02 -1.562e-01 2.600e-02 -1.212e-01 -6.400e-03 -6.100e-03
8.650e-02 -9.980e-02 7.600e-02 -1.770e-02 -1.645e-01 7.320e-02
3.240e-02 -3.000e-03 6.130e-02 -1.077e-01 -9.550e-02 9.340e-02
-6.980e-02 -9.190e-02 -5.650e-02 -1.528e-01 -2.370e-02 -2.281e-01
2.585e-01 -4.490e-02 1.400e-02 9.080e-02 2.393e-01 1.990e-02
-2.330e-02 1.032e-01 1.154e-01 1.243e-01 -6.610e-02 -8.540e-02]

Interaktion mit LLMs



1) Interaktiv: „Chat“-Modus

Dialogszenario über Plattformen, **ohne echte wissenschaftliche Validität**, weil:

- keine Replizierbarkeit der Antworten
- keine Kontrolle über die Interaktion
- unklar bleibt, was zwischen der Benutzeroberfläche und dem eigentlichen Modell geschieht (z. B. zusätzliche Systemprompts, Filtermechanismen, Optimierungen)

1) Coding framework:

Gleiche Klassifikationsprinzipien wie in Machine-Learning-Aufgaben → das Modell wird dazu aufgefordert, bestimmte **Aufgaben an vorgegebenem Sprachmaterial auszuführen** (z. B. linguistische Analysen).

2) Generierung von Text:

Das Modell wird dazu verwendet, **sprachliches Material in Textform zu produzieren**, das anschließend analysiert wird.

2. Interaktion mit LLMs

Prompt engineering:

Gestaltung der richtigen Fragen oder Anweisungen (Prompts), um LLMs zu gewünschten Ergebnissen zu führen

- Bei LLMs → Fokus verschiebt sich von Datenaufbereitung hin zur Formulierung der richtigen Fragen
- ***In-context learning***: Modell lernt direkt über Prompts, ohne zusätzliches *training* oder *fine-tuning*

2. Interaktion mit LLMs

Prompt:

// Was man in ChatGPT eingibt = Anweisungen für das Modell

- Der Prompt sollte alle Infos enthalten, die das Modell braucht, um die gewünschte Ausgabe möglichst gut zu erzeugen (in unserem Kontext: die Aufgabe korrekt/menschenähnlich lösen)
- Der Kontext der Aufgabe sollte detailliert beschrieben werden: Je spezifischer, desto besser
(diese Infos könnten im Trainingsdatensatz nicht enthalten sein!)
- Aber länger ist nicht immer besser!
- → Deshalb müssen Experimente mit verschiedenen Prompts gemacht werden, um den am besten geeigneten zu finden und die beste Modelleleistung zu erreichen

2. Interaktion mit LLMs

Prompt im Code vs. Prompt auf Platforms/Interfaces (z.B. ChatGPT):

→ Kein großer Unterschied!

Was man in die Chat-Oberfläche eingibt, wird im Code zum Prompt.

OpenAI erlaubt die Angabe von drei Rollen (nicht immer alle nötig):

- **user** :
Repräsentiert die Eingabe des Nutzers (z. B. deine Anweisungen in der ChatGPT-Interface)
- **system** :
Beschreibt, was das Modell tun soll und wie es sich im Allgemeinen verhalten bzw. antworten soll
z. B.: "Du bist ein hilfreicher Assistent", "Du bist ein Linguist", "Du bist ein Spanischer Muttersprachler",...
- **assistant** :
Die Antworten des Modells (insbesondere bei Chatbots verwendet)

→ Jede Rolle beinhaltet **content** Feld, wo die Beschreibung hinkommen soll.

KI statt Crowd? - Workshop der AG Digitale Romanistik
Iris Ferrazzo

2. Interaktion mit LLMs

Prompt engineering Strategien:

● Zero-shot Prompting

- Der Prompt weist das Modell direkt an, eine Aufgabe auszuführen
- Es werden keine zusätzlichen Beispiele gegeben
- Beispiel:

„Übersetze den folgenden Satz ins Französische: „Ich liebe Linguistik!““

● Few-shot Prompting

- Ermöglicht **In-Context Learning**, indem Beispiele direkt in den Prompt aufgenommen werden
- Das Modell lernt die Aufgabe anhand dieser Beispiele
- Beispiel:

„Übersetze den folgenden Satz ins Französische:

'Ich liebe Linguistik' → 'J'aime la linguistique.'

'Das Buch ist auf dem Tisch' → 'Le livre est sur la table.'

'Sie lernt Phonetik.' → [Antwort des Modells]

2. Interaktion mit LLMs

● Chain-of-Thought (CoT) Prompting:

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

LLMs als Crowd

- **Stärke:** *Folgen von Anweisungen* (instruction-following abilities), entwickelt während des Pre-Trainings (OpenAI, 2023)

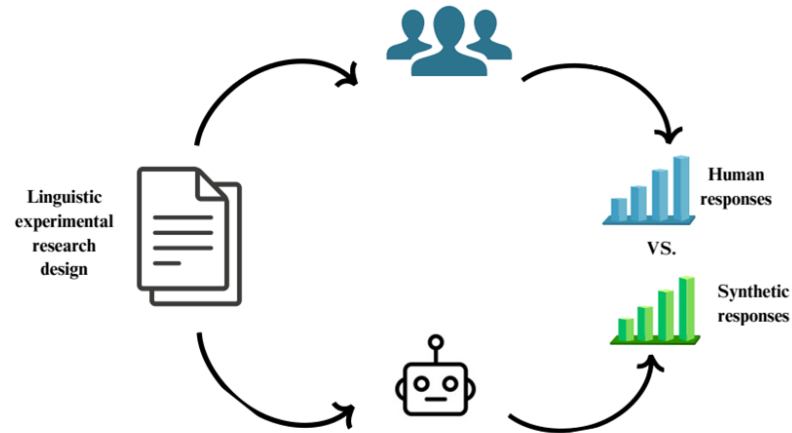
Bereits genutzt anstelle menschlicher Teilnehmer*innen, aber:

- **In NLP:** Annotationsaufgaben (z.B. POS-Tagging, NER, syntaktisches Parsing, usw.)
- **In Linguistik:** sprachlich und kognitiv komplexere Aufgaben, die strukturierten menschlichen Output oder Urteile erwarten

- **Schwäche: Instabilität:**

- Kleine Variationen in der **prompt phrasing** können die Output-Qualität beeinflussen (Webson und Pavlick, 2022)
- **Inkonsistenz** bei wiederholten Anfragen mit identischen Inputs (Gilardi et al., 2023)
- **Strukturierte Formate** wie Multiple-Choice-Fragen sind besonders schwierig (Zheng et al., 2023; Yin et al., 2023; Zhang et al., 2023; Pezeshkpour und Hruschka, 2024)
- **Äußern selten Unsicherheit**, selbst bei mehrdeutigen Inputs (Wang et al., 2024)
- Tendenz, **bejahende** gegenüber negativen oder unsicheren Antworten zu bevorzugen (Jiang et al., 2024; Tjautja et al., 2024)

LLMs als Crowd: Unser heutiger Ansatz



KI statt Crowd? - Workshop der AG Digitale Romanistik
Iris Ferrazzo

Vielen Dank!

iris.ferrazzo@uni-bonn.de



KI statt Crowd? - Workshop der AG Digitale Romanistik
Iris Ferrazzo