Güte unter Zeitdruck

KI-gestützte Inhaltsanalyse als transparente Unterstützung beim Kategorisieren und Kodieren

Christian Atzendorf

f-bb Forschungsinstitut Betriebliche Bildung

Hans Böckler Stiftung

29. Jahrestagung der DeGEval
Evaluation und Zeit

Agenda

1. **Projektkontext, Daten und Kernthesen**
2. **Zentrale Konzepte**
   - Wie LLMs Text generieren
   - Retrieval Augmented Generation (RAG) - in Kürze
3. **Fallstudie: Deduktives Kodieren von Interviews**
4. **Erkenntnisse für KI-gestützte qualitative Kodierung**

Projektkontext und Daten

Projekt

IQ Fachstelle Anerkennung und Qualifizierung (1/2025 – 12/2028, BMAS/ESF Plus)

Fachkräfteeinwanderung – neue Chancen, neue Risiken (10/2025 – 07/2027, Hans-Böckler-Stiftung)

Daten

Strukturierter Kurskatalog extrahiert via API aus KURSNET und mein NOW (80.030 Einträge, 10.708 unterschiedliche Kurstexte)

7 leitfadengestützte Expert*inneninterviews, 42 ausgewählte Passagen (für die Fallstudie)

Aufgabe der KI/Anwendungsfeld

Kurs auf Relevanz für Anerkennung prüfen und wenn möglich Referenzberuf zuweisen

Deduktive Mehrfachkodierung (thematisch/funktional) und Bestimmung der Segmentgrenzen

(Atzendorf et al., 2024; Atzendorf & Judas, 2025)

Fünf Thesen für KI-gestützte Kodierung

Hans Böckler Stiftung

fbb

01

KI kann kodieren. Eine nachvollziehbare Analyse erfordert aber einen kontrollierten und transparenten Arbeitsablauf.

02

RAG unterstützt diesen Prozess, führt jedoch auch weitere Komplexität und mögliche (neue) Fehlerquellen ein.

03

Ein Prompt allein ist kein Kodierprotokoll.

04

Die einzelnen Arbeitsschritte KI-gestützter Kodierung müssen getrennt und anhand spezifischer Gütekriterien beurteilt werden.

05

Der lokale Einsatz von KI ermöglicht die hierfür nötige Kontrolle über Daten und Modelle.

Zentrale Konzepte

Wie LLMs Text generieren

- Text wird in Tokens zerlegt, die aus Wörtern, Wortteilen oder Satzzeichen bestehen können
- Das Modell berechnet eine Wahrscheinlichkeits-verteilung für das nächste Token, wählt eines aus, fügt es dem bisherigen Kontext hinzu und wiederholt den Prozess
- Der erzeugte Text folgt statistischen Mustern des Sprachgebrauchs. Eine plausible Fortführung des Ausgangstextes (so ausgeklügelt sie wirken mag) ist in sich selbst daher kein Beleg für Verständnis, Absicht oder Richtigkeit

(Bender et al., 2021; Vaswani et al., 2017)

[Qual][itative][Inhalts]analyse]

[leist] 20,8%
[dient] 15,6%
[hilft] 9,7%
[liefer] 5,4%
[analys] 4,1%
andere 44,4%

[Qual][itative][Inhalts][analyse][leist][et][eine]
[system][atische][und][interpret][ative][Aus][wertung]
[von][Text][material][,][um][verborgene][Muster][,]
[Themen][und][Bedeut][ungen][zu][erkennen][.]

Prompt: Erkläre kurz, was qualitative Inhaltsanalyse leistet.
Modell: qwen3.5:4b, **Temperatur 0**

Zentrale Konzepte

Retrieval Augmented Generation (RAG)

- Ein Embedding-Modell überführt Texte in einen Vektor. Die zu codierende Passage und die Codekarten werden dabei getrennt verarbeitet, aber im selben erlernten semantischen Raum repräsentiert.
- Die Cosinus-Ähnlichkeit ermöglicht den Vergleich der Richtung dieser Vektoren. Vektoren semantisch ähnlicher Texte weisen tendenziell in eine ähnliche Richtung und erhalten daher einen höheren Ähnlichkeitswert.

(Lewis et al., 2020; Reimers & Gurevych, 2019; Sturua et al., 2024)

RAG-Pipeline für die Mehrfachkodierung
Embedding Modell: jina-embeddings-v3

[…]
Rang 5 (0,559) Unklarheiten und Interpretationsspielräume (K)
Rang 7 (0,546) Kosten der Verfahren (K)
Rang 9 (0,531) Gute Praxis oder Lösungsansatz (M)
Rang 12 (0,526) Ermessens- und Vollzugspraxis (K)

Legende:
(K) von der KI gewählt und im Review akzeptiert
(M) menschlicher Code, von der KI ausgelassen

Fallstudie: Deduktives Kodieren von Interviews

Codekarten-Beispiel: FEG-C128

Codekarte

FEG-C128 Ermessens- und Vollzugspraxis

Hierarchie	**Institutionelle Umsetzung und Verfahren** → Ermessens- und Vollzugspraxis
Definition	Praktische Auslegung und Anwendung rechtlicher Vorgaben durch zuständige Stellen.
Einschließen, wenn	Ermessen oder behördliche Umsetzungspraxis das Ergebnis beeinflussen.
Ausschließen, wenn	Reine Rechtsbeschreibung ohne Anwendungspraxis vorliegt.

Das Modell erhält

- stabile ID
- Hierarchie
- Definition
- Ein- und Ausschlussregeln

Nicht enthalten

- menschliche Kodierung
- Beispiele (aktuell zumindest)
- Ähnlichkeitswerte
- frühere Review-Entscheidungen

Hans Böckler Stiftung

fbb

Fallbeispiel: Deduktives Kodieren in Interviews

Auswahl der Analyseeinheiten

Für sieben Interviews Textpassagen an sechs unterschiedlichen Stellen:

10 %	25 %	40 %	55 %	70 %	85 %

- Je Position Auswahl der nächstgelegenen vollständigen Passage (220 Wörter maximal)
- Organisatorische oder unvollständige Teile wurden ausgeschlossen: keine Zufallsstichprobe

Kodiereinheit
Zielpassage

Kontexteinheit
Ganzer Sprecher*innenblock, nicht das volle Interview

(Fischer, 2025)

Fallbeispiel: Deduktives Kodieren in Interviews

Zyklischer Review-Prozess für KI-unterstütztes Kodieren

Input — **42 Passagen aus 7 Expert*inneninterviews**

01 Menschliche Referenzkodierung

QualCoder
(121 Codes im Codebuch)
109 Passage-Code-Zuordnungen

02 KI-gestützte Kodierung

Embedding Modell:
jina-embeddings-v3
Inferenz Modell:
qwen3.5:4b,
Temperatur 0

03 Evidenz dokumentieren

angeboten / ausgewählt
Auslassungen / Ergänzungen
Code ≠ Segmentgrenze

04 Menschliche Revision

akzeptieren / umcodieren / verwerfen
Segmentgrenzen anpassen
Codebuch überarbeiten

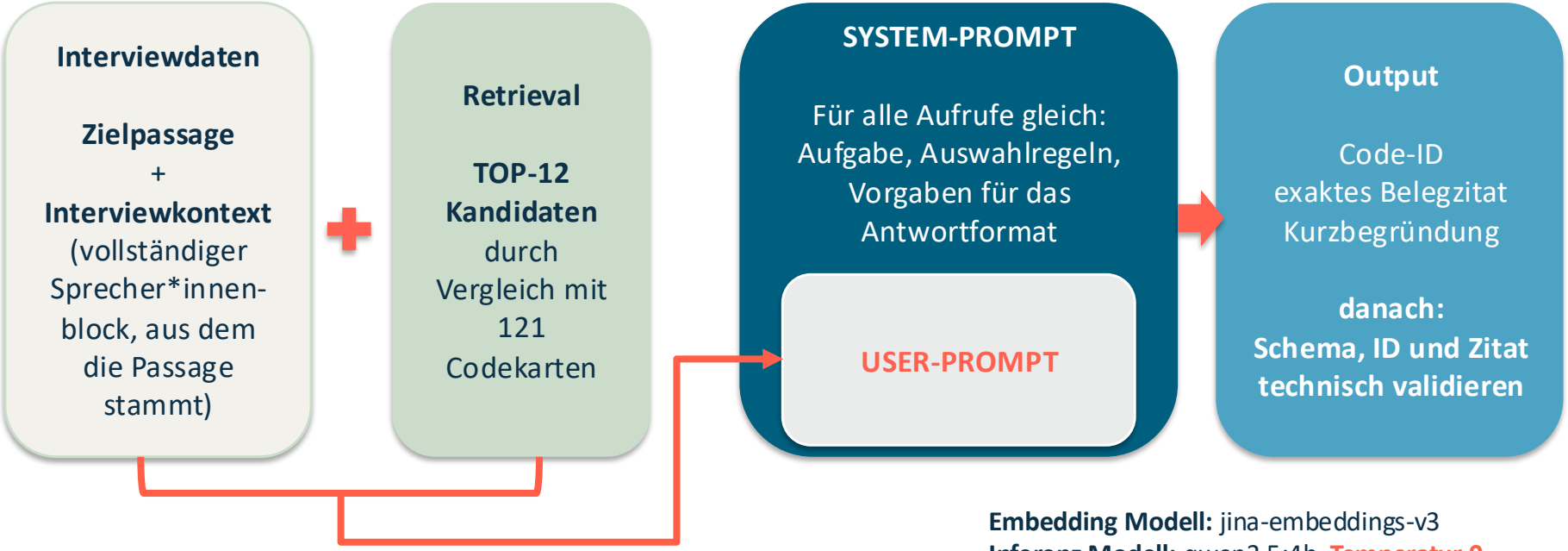
Die Prüfung der Modellausgaben erfolgt durch den dokumentierten Abgleich mit der unabhängigen menschlichen Kodierung und durch nachvollziehbare Revisionsentscheidungen

05 Wiederholen

(Ashwin et al., 2026; Fischer, 2025; Kuckartz & Rädiker, 2024; Than et al., 2025)

Fallbeispiel: Deduktives Kodieren in Interviews

KI-gestützte Kodierung: Wie wurde jede Passage verarbeitet?

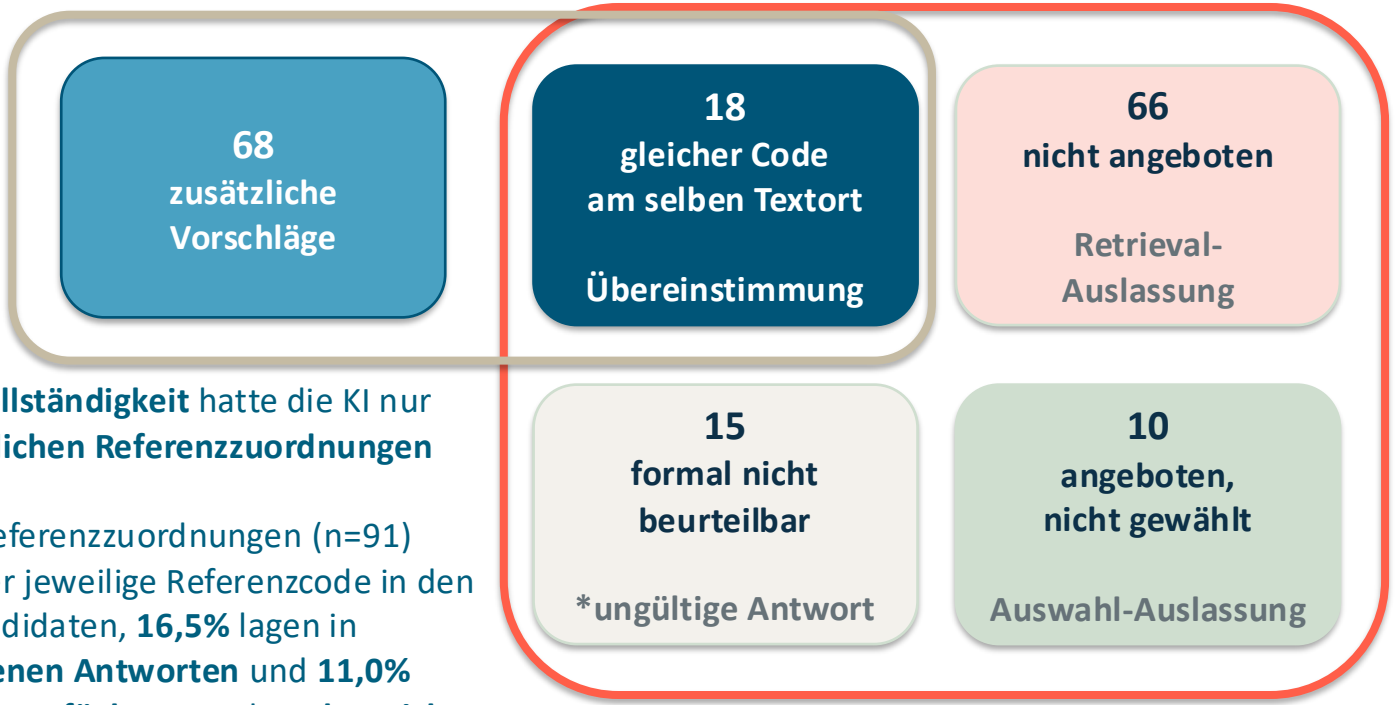


Embedding Modell: jina-embeddings-v3
Inferenz Modell: qwen3.5:4b, **Temperatur 0**

Fallbeispiel: Deduktives Kodieren in Interviews

Abgleich der menschlichen (M) und KI-Kodierung (K)

Ausgangspunkt:
K (n=86), M (n=109)



- In Bezug auf die **Vollständigkeit** hatte die KI nur **16,5% der menschlichen Referenzzuordnungen** reproduziert
- Von den übrigen Referenzzuordnungen (n=91) fehlte bei 72,5% der jeweilige Referenzcode in den zwölf Retrieval-Kandidaten, **16,5%** lagen in **technisch verworfenen Antworten** und **11,0%** waren **als Kandidat verfügbar,** wurden **aber nicht ausgewählt**

***ungültig im Sinne der Vorgaben des System Prompts (7 Übereinstimmung/ 8 Auswahl-Auslassung)**

Kodierbeispiel: BIBB

Die Leute müssen frühzeitig wissen, welche Möglichkeiten sie haben oder haben könnten. Dann müssen sie selbst entscheiden. Aber wie soll man ihnen das verklickern? Menschen wie wir beschäftigen sich seit mehreren Jahren mit diesem Gesetz. **Ich kenne es in- und auswendig und erlebe trotzdem immer wieder, dass die Auslandshandelskammer es erneut erklären muss. Und dann möchte jemand verstehen, was das für den eigenen Einzelfall bedeutet. Das ist und bleibt eine Herausforderung.**

Menschliche Kodierung (Rahmen)

M(1) Gute Praxis oder Lösungsansatz
Begründung: Menschen sollten frühzeitig über ihre Möglichkeiten informiert werden, damit sie selbst eine Entscheidung treffen können.

M(2) Rechts- und Verfahrenskomplexität
Begründung: Die Vermittlung des Gesetzes und seine Anwendung auf individuelle Fälle werden als dauerhafte Herausforderung beschrieben.

KI-Kodierung (fett)

K(1) Rechts- und Verfahrenskomplexität
Begründung: Selbst eine Person, die das Gesetz in- und auswendig kennt, erlebt, dass die Auslandshandelskammer es immer wieder erklären muss.

K(2) Unklarheiten und Interpretationsspielräume
Begründung: Es bleibt schwierig zu verstehen, was die verfügbaren Möglichkeiten für den jeweiligen Einzelfall bedeuten.

Review der zusätzlichen Vorschläge (n=68)

Inhaltliche Entscheidung

48	als zusätzliche Kodierung akzeptiert
20	inhaltlich relevant, aber umcodiert
0	als irrelevant verworfen

Segmentgrenzen der Vorschläge

60	Grenze übernommen
8	Grenze erweitert
0	Grenze eingeengt oder ersetzt

Einordnung der Ergebnisse

- Die inhaltliche Qualität der neuen Kodierungen war durchgängig hoch: alle 68 beurteilten Zusätze waren inhaltlich relevant
- Gegenüber der menschlichen Kodierung gab es zahlreiche Auslassungen (vor allem durch Retrieval-Auslassungen)
- Die KI-Kodierung hat die menschliche nicht ersetzt, sondern qualitativ eher ergänzt

Erkenntnisse für KI-gestützte qualitative Kodierung

***Analyseeinheiten festlegen**

- Verarbeitungseinheit: Zielpassage
- Kodiereinheit: ausgewählte bzw. kodierte Textstelle
- Kontexteinheit: zugehöriger Sprecher*innenblock

***Analyseschritte trennen**

- Vorverarbeitung: Retrieval und Promptaufbau
- Verarbeitung: Kodierung durch das LLM
- Nachverarbeitung: Ausgabe technisch prüfen

Stufenweise evaluieren

- Retrieval: Wurde der Code angeboten?
- Auswahl: Wurde der Code gewählt?
- Review: Passen Code und Segmentgrenze zusammen?

LOKALER KI-EINSATZ
Interviewdaten blieben in eigener Infrastruktur
Modell(version), System- und Userprompt sowie Rohantworten bleiben kontrollier-, prüf- und verfügbar

(CERT-EU, 2023; *Fischer, 2025; Pohle & Thiel, 2020; Than et al., 2025)

Literatur 1/3

Ashwin, J., Chhabra, A., & Rao, V. (2026). Using large language models for qualitative analysis can introduce serious bias. *Sociological Methods & Research, 55*(3), 795–839. https://doi.org/10.1177/00491241251338246

Atzendorf, C., Erfurt, T., Mainka, A.-L., & Roser, L. (2024). Berufliche Anerkennung von Gesundheitsfachkräften mit einer im Ausland erworbenen Berufsqualifikation – Situationsanalyse. https://netzwerk-iq.de/fileadmin/Redaktion/Downloads/FSAQ/FSAQ_SituationsanalyseGFB.pdf

Atzendorf, C., & Judas, K. (2025). Quali gesucht!? Das Angebot von Qualifizierungen im Kontext der Anerkennung ausländischer Berufsabschlüsse [Arbeitshilfe]. IQ Fachstelle Anerkennung und Qualifizierung, Forschungsinstitut Betriebliche Bildung (f-bb) gGmbH. https://netzwerk-iq.de/fileadmin/Redaktion/Downloads/FSAQ/FSAQ_Arbeitshilfe_QualiGesucht_barr.pdf

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

CERT-EU. (2023, May 31). *Potential impact and risks of generative AI in EUIBAs* (Security Guidance 23-002, Version 1.2). https://cert.europa.eu/static/security-guidance/TLP-CLEAR-CERT-EU_Security_Guidance-23-002_v1.2.pdf

Fischer, A. (2025). *Die Zukunft der qualitativen Inhaltsanalyse im Zeitalter künstlicher Intelligenz: Ein Ablaufmodell KI-basierter Inhaltsanalyse* [The future of qualitative content analysis in the age of artificial intelligence: A process model for AI-based content analysis] (f-bb-online 03/25). Forschungsinstitut Betriebliche Bildung (f-bb) gGmbH. https://doi.org/10.25656/01:34405

Kuckartz, U., & Rädiker, S. (2024). *Qualitative Inhaltsanalyse: Methoden, Praxis, Umsetzung mit Software und künstlicher Intelligenz* (6., überarbeitete und erweiterte Auflage). Beltz Juventa.

Literatur 2/3

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/6b493230-Abstract.html

Pohle, J., & Thiel, T. (2020). Digital sovereignty. *Internet Policy Review, 9*(4). https://doi.org/10.14763/2020.4.1532

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), *Official Journal of the European Union L*, 2024/1689 (2024). http://data.europa.eu/eli/reg/2024/1689/oj

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Sturua, S., Mohr, I., Akram, M. K., Günther, M., Wang, B., Krimmel, M., Wang, F., Mastrapas, G., Koukounas, A., Wang, N., & Xiao, H. (2024). *jina-embeddings-v3: Multilingual embeddings with task LoRA* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.10173

Literatur 3/3

Than, N., Fan, L., Law, T., Nelson, L. K., & McCall, L. (2025). Updating “the future of coding”: Qualitative coding with generative large language models. *Sociological Methods & Research, 54*(3), 849–888. https://doi.org/10.1177/00491241251339188

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Kontakt

Christian Atzendorf
Forschungsinstitut Betriebliche Bildung (f-bb)
Stresemannstraße 121
10963 Berlin

Telefon: +4930 4174986-57
Mail: christian.atzendorf@f-bb.de
Web: www.f-bb.de

f-bb Forschungsinstitut Betriebliche Bildung

Hans Böckler Stiftung