

SEVAL Kongress 2025

Unterarbeitsgruppe

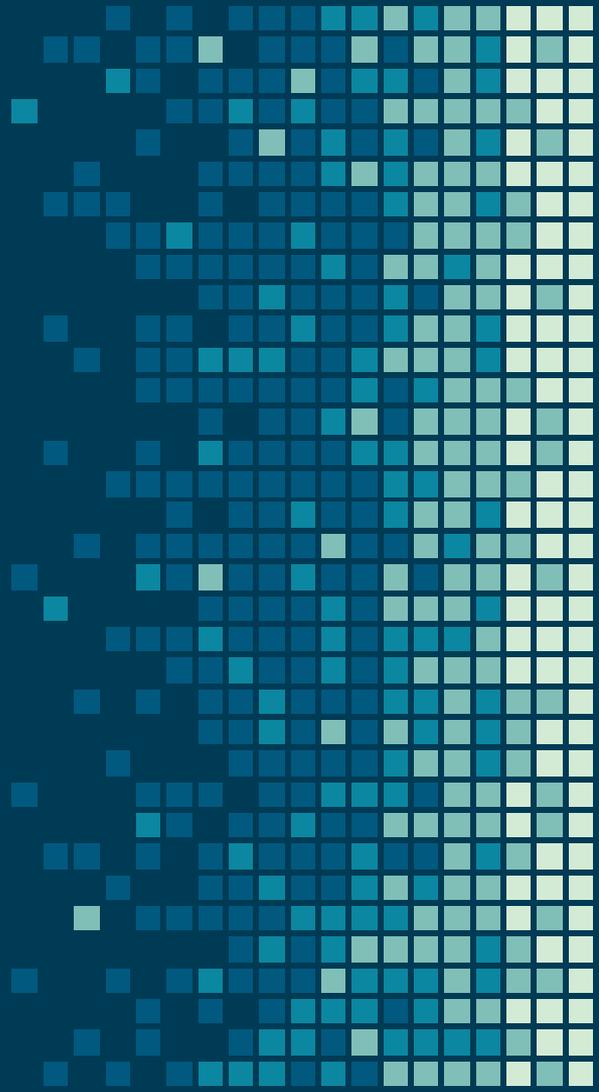
Anwendung von AI in der quantitativen Forschung

Laura Gatto <laura.gatto@cdc.ge.ch>

(Policy Evaluator, Cour des comptes – République et Canton de Genève)

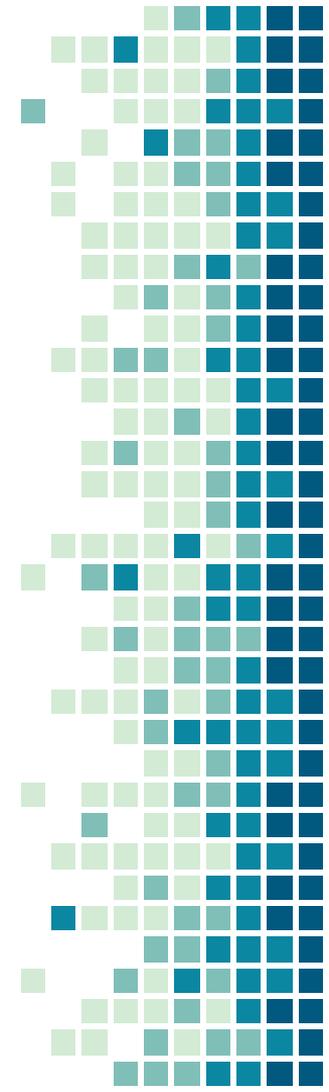
Wolfgang Schneider <wschneider@caritas.ch>

(Advisor & Quality Manager MEAL Caritas Schweiz IZA)



Disclaimer

- Datenschutz..
- Bias..
- AI kann sich irren – Benutzung auf eigene Gefahr
- Wir sind Nutzende von KI, keine Expert:innen



“[AI are] Systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.”

(High-Level Expert Group on Artificial Intelligence, 2019)

AI in dieser Präsentation:

*Nutzung von Large Language Models
(LLMs)*

(z.B. CoPilot, Grok, ChatGPT, Gemini, Claude..)

Überblick Quantitative Daten & LLMs

LLM assistiert

- Datenbereinigung
- Entwurf von Code für die Analyse
- Fehlerbehandlung
- Verfassen von Berichten (?)
- Entwerfen von Umfragen
- Unterrichten von Datenanalyse (?)

Beschleunigt Prozesse

LLM analysiert

- Vollständige Analyse durch LLM-Assistenten
Deskriptive Statistik,
Datenexploration
- Fortgeschrittene Statistik (?)
- Interpretation & vertiefte Erkenntnisse (?)

Automatisiert Analyse

LLM als Instrument der
quantitativen Textanalyse (Big
Data)

Erweiterung der klassischen
NLP-Aufgaben:

- Sentiment-Analyse
- Themenmodellierung,
Clustering
- Extraktion von
Informationen
- Spracherkennung

Die Analyse umfangreicher Texte
leichter zugänglich machen

1. Limitierende Faktoren



«Language Model» & Mathe?

calculate 10000 x (1.097)^6 manually in one go, only share/write the answer, no steps in between, nothing.

18,719,7

Falsch



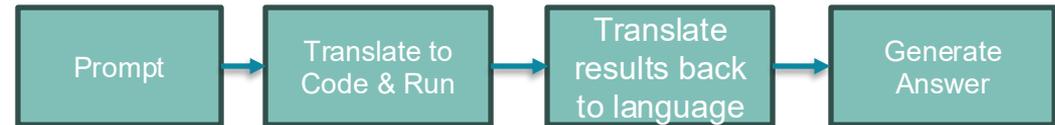
what is 10000 x (1.097)^6? Calculate with python.

Analysiert ^

```
python Details immer anzeigen  | Kopieren  
  
value = 10000 * (1.097)**6  
value  
  
Ergebnis  
17427.687562285468
```

The result is 17,427.69 (approx). [🔗](#)

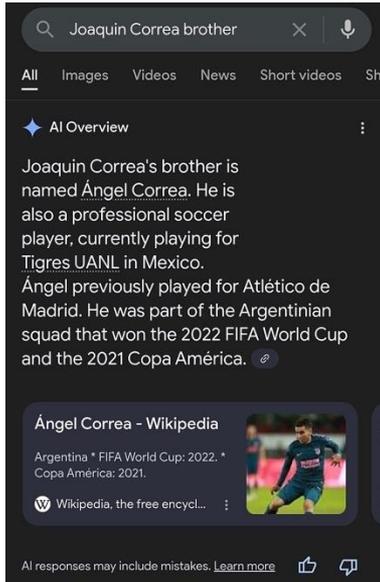
Korrekt



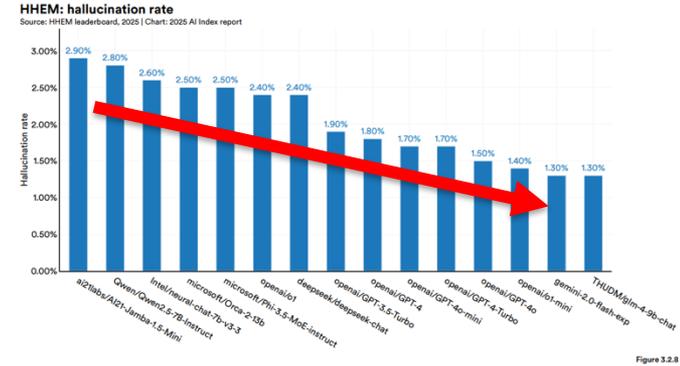
→ Verwenden einer bezahlten Version, um einen direkten Code-Interpreter zu erhalten.

Alternative: Code in LLM erstellen und in einem externen Programm ausführen (STATA, R, Excel...)

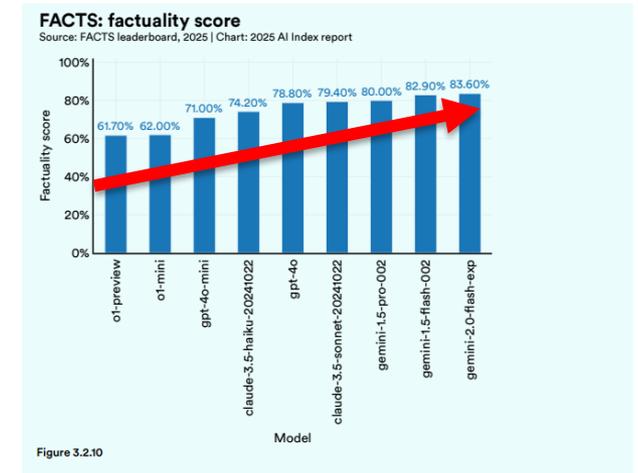
AI Halluzinationen



HHEM: Wie häufig führen LLMs bei der Zusammenfassung von Dokumenten Halluzinationen ein?



FACTS: Wie gut können LLMs Antworten erstellen, die sowohl sachlich korrekt als auch detailliert genug sind, um zufriedenstellende Antworten zu geben? (Benchmark erfordert Zusammenfassung, Frage- und Antwortgenerierung, Tatsachenfeststellung und Erklärung)



By AI Overviews, prompted by the query "Joaquin Correa brother" on 10 August 2025 - AI Overviews, prompted by the query "Joaquin Correa brother" on 10 August 2025, Public Domain, <https://commons.wikimedia.org/w/index.php?curid=172963779>

AI Halluzinationen

- LLMs klingen selbstbewusst, können aber falsche/irrelevante Antworten geben
 - Grund-Training fördert «**erraten**» des nächsten Wortes
 - LLM-Entwickler probieren mit Finetuning/Tools Halluzinationen wieder einzudämmen - **aber evtl. nicht 100%ig möglich**
- Zusätzlich:
 - Fehler in Trainingsdaten, Kodierung, Dekodierung von Daten
 - Fehlende Informationen in den Trainingsdaten
 - Fehler beim „Verstehen“ der Aufforderung, beim „Interpretieren“ der Daten, beim Vorhandensein aller Zusammenhänge

Immer Kontrollieren!

- Nur verwenden, was man versteht / verifizieren kann
 - Man muss in der Lage sein, Output zu hinterfragen & zu bewerten
- Erzeugen von Code → ermöglicht validierung des Analyseprozesses

2. Demo: LLM-basierte Datenanalyse



Schritte der Datenanalyse

Datenerhebung

Vorverarbeitung

Exploration,
Deskriptive Statistik

Vertiefte Analyse
(statistische
Modellierung)

Auswertung und
Berichterstattung

DEMO



System & Dataset

- ChatGPT mit GPT-5-Modell und „Plus“-Konto
- Dataset (1532 Antworten)
 - Kleiner Datensatz einer Umfrage aus vier Projekten in drei Ländern
 - Wichtige Elemente des Datensatzes:
 - Aktueller Beschäftigungsstatus: Daten für alle Befragten
 - Beschäftigungsstatus vor 12 Monaten: 955 Antworten
 - Wahrgenommene Veränderung der finanziellen Situation in den letzten 12 Monaten: 1.506 Antworten, kategorisiert als negativ, unverändert oder positiv.
 - Aktuelles Einkommen: Simulierte Einkommensdaten für alle
 - Gründe für finanzielle Veränderungen: 813 Antworten.

Datenvorverarbeitung (ChatGPT, GPT-5)

lets do data analysis! Start with cleaning this data set. Data cleaning
Checking quality, missing data, coding, merging.

Analyzed 

Here's a quick overview of the dataset before cleaning:

- **Rows:** 1532
- **Columns:** 15
- **Issues identified:**
 1. **Long/complex column names** – hard to work with.
 2. **Missing data** – many columns (especially from column 4 onward) have values only for 813 rows or fewer.
 3. **Inconsistent formatting** – leading spaces in some column names.
 4. **Multilingual labels** – some French columns in an otherwise English dataset.
 5. **Possible misaligned columns** – one column has no closing parenthesis.

Next steps:

1. **Rename and simplify columns.**
2. **Trim whitespaces.**
3. **Check & quantify missing data.**
4. **Identify if merging is needed (only one sheet here, so likely not).**
5. **Clean datatype issues.**

Proceeding with step-by-step cleaning now. 

Proceeding with step-by-step cleaning now. 

Analyzed 

```
python Always show details   Copy code

# Step 1: Strip whitespace and simplify column names
df.columns = df.columns.str.strip() # Remove leading/trailing spaces

# Create simplified, programmatic column names
simplified_columns = {
    df.columns[0]: "Currently_Employed",
    df.columns[1]: "Employed_12mo_Ago",
    df.columns[2]: "Financial_Change",
    df.columns[3]: "Current_Income_Chg",
    df.columns[4]: "Income_Source_Change",
    df.columns[5]: "Market_Price_or_Access_Change",
    df.columns[6]: "Agri_Business_Change",
    df.columns[7]: "Financial_Resource_Access_Change",
    df.columns[8]: "Tech_Infrastructure_Change",
    df.columns[9]: "Skills_Knowledge_Change",
    df.columns[10]: "Policy_External_Actors_Change",
    df.columns[11]: "Service_Access_Change",
    df.columns[12]: "Community_Network_Change",
    df.columns[13]: "Natural_Resources_Change",
    df.columns[14]: "Household_Dynamics_Change"
}

# Rename columns
df.rename(columns=simplified_columns, inplace=True)

# Step 2: Check missing data
missing_summary = df.isnull().sum().to_frame(name="Missing_Values")
missing_summary["%_Missing"] = (missing_summary["Missing_Values"] / len(df) * 100).round(2)

# Step 3: Check data types
```

Code

953	Yes	Yes	-1
954	Yes	Yes	-1
955	Yes	Yes	-1
956	Yes	NA	0
957	No	NA	-1
958	No	NA	0
959	No	NA	0
960	No	NA	0



953	1	1	-1
954	1	1	-1
955	1	1	-1
956	1		0
957	0		-1
958	0		0
959	0		0
960	0		0

Your cleaned dataset is ready:

 [Download Cleaned-Data-Set.xlsx](#)

Output: cleaned excel

Let me know if you want visualizations, stats, or modeling next. 

Transformation

make buckets: 0-25%, 25-50%, 50-75%, 75-100% <- add a new column for this

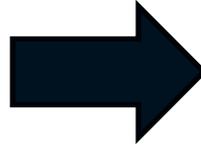
Analyzed ▾

✔ Income buckets added under new column: Income Bucket

Bucket	Income Range (CHF)	Count
0-25%	0 - 250	281
25-50%	250 - 500	446
50-75%	500 - 750	424
75-100%	750 - 1000	381

Would you like:

- A download with this new column?
- Cross-analysis (e.g., income bucket × job status or financial change)? [↗](#)



```
python Always show details Copy code  
  
# Load the reuploaded file  
file_path = "/mnt/data/test-data-set.xlsx"  
df = pd.read_excel(file_path)  
  
# Clean column names  
df.columns = df.columns.str.strip().str.replace('\n', '', regex=True).str.replace(' ', '', regex=True)  
  
# Recalculate income buckets  
income_col = 'Current Income (HH, CHF)'  
max_income = df[income_col].max()  
bins = [0, 0.25 * max_income, 0.50 * max_income, 0.75 * max_income, max_income + 1]  
labels = ['0-25%', '25-50%', '50-75%', '75-100%']  
  
# Add Income bucket column  
df['Income Bucket'] = pd.cut(df[income_col], bins=bins, labels=labels, right=False)  
  
# Save to Excel  
output_bucketed = "/mnt/data/income_buckets_final.xlsx"  
df.to_excel(output_bucketed, index=False)  
  
output_bucketed
```

Code

Entry-ID	Employment not or other	12 months	Current Income (HH, CHF)	Income Bucket
1	1	1	-1	813 75-100%
2	0	0	0	794 75-100%
3	1	0		271 25-50%
4	1	0	0	515 50-75%
5	1	0	1	749 50-75%

Buckets

let's do a comparison of these two (where both entries are). How many people gained jobs, lost jobs, stayed with job, without job?

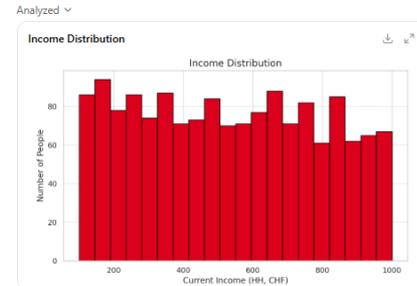
Grafik & Tabellen

What would be the best plot to visualize the income distribution to experts? Generate it.



Column: Currently Employed

- Type: Binary (1 = Yes, 0 = No)
- Distribution:
 - ✔ Yes (1): 850 individuals (55.5%)
 - ✘ No (0): 682 individuals (44.5%)



Here's a simple histogram of income — better suited for non-experts:

- Easy to read: shows how many people fall into each income range.
- ✔ Great for presentations or stakeholder summaries.

Analyzed ▾

Job Status Comparison (Current vs 12 Months Ago)

(1,174 respondents with both entries available)

Transition	Count
✔ Stayed with job	558
✘ Stayed without job	211
☑ Gained job	147
☒ Lost job	39

Clear insight:

- Most people kept their job.
- ~3× more gained jobs than lost them.

Continue with the financial situation change column next? ↗



in%?

Analyzed ▾

Job Status Comparison (%)

Transition	Count	Percent
✔ Stayed with job	558	58.4%
✘ Stayed without job	211	22.1%
☑ Gained job	147	15.4%
☒ Lost job	39	4.1%

Lässt Einträge mit fehlenden Daten korrekt aus.

Statistik

Nur verwenden, was man versteht!

- Richtige Methodik?
- Versteckte bias? (sampling, non-responses, ...)
 - Versteckte variablen?
- Interpretation? (e.g. Kausalität)

you said Staying unemployed strongly correlates with a worsened financial situation. is this really the case?

Good question — let's test that claim properly rather than relying on eyeballing.

Hypothesis

H_0 : Financial situation is independent of job transition

H_1 : Financial situation is associated with job transition

We can test this with a Chi-squared test of independence on the cross-tab.

Let me run the test to see if the relationship is statistically significant.

Analyzed 

Chi-Squared Test Result

Metric	Value
Chi ² Statistic	104.17
Degrees of Freedom	6
p-value	< 0.0001
Significant?	<input checked="" type="checkbox"/> Yes (at $\alpha = 0.05$)

Code Generierung

generate a downloadable r-script

Analysefehler ▾

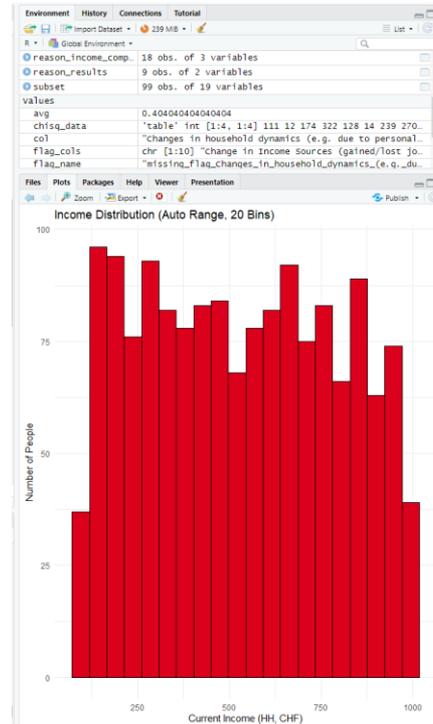
Analysiert ▾

Your R script is ready. You can download it here:

[Download clean_data_script.R](#)



```
52 print(cross_tab)
53
54 # chi-squared test
55 chisq_data <- table(
56   data$job_transition,
57   data$S11.3. Has your financial situation changed compared to 12 months ago? (1 improved, 0 stayed the same, -1 gotten w
58 )
59 chi_result <- chisq.test(chisq_data)
60 print("chi-squared Test:")
61 print(chi_result)
62
63 # Create income buckets
64 max_income <- max(data$`Current Income (HH, CHF)` , na.rm = TRUE)
65 data$income_bucket <- cut(
66   data$`Current Income (HH, CHF)` ,
67   breaks = c(0, 0.25*max_income, 0.5*max_income, 0.75*max_income, max_income + 1),
68   labels = c("0-25%", "25-50%", "50-75%", "75-100%"),
69   right = FALSE
70 )
71
72 # Income histogram (non-expert)
73 ggplot(data, aes(x = `Current Income (HH, CHF)`) +
74   geom_histogram(fill = "#000000", color = "black", bins = 20)) +
75   labs(title = "Income Distribution", x = "Income (CHF)", y = "Count") +
76   theme_minimal()
77
78 # violin + box plot (expert)
79 ggplot(data, aes(x = "", y = `Current Income (HH, CHF)`) +
80   geom_violin(fill = "#000010", trim = FALSE) +
81   geom_boxplot(width = 0.1, fill = "white")) +
82   labs(title = "Income Distribution (Violin + Box Plot)", y = "Income (CHF)", x = "") +
83   theme_minimal()
84
85 # Reason analysis: average income change if cited as positive
86 reason_cols <- c(
87   "market_price_access_change",
88   "changes_in_agricultural_production / business performance (decline/growth)",
89   "changes_in_access_to_financial_resources (gained/lost access to credit, remittances, loans)",
90   "skills_knowledge_change",
91   "changes_in_governmental_policies / external actors activities (benefited from / impaired by government or other external
92   "changes_in_access_to_services (coaching, health, other support)",
93   "changes_in_access_to_public_spaces and green spaces")
```



Läuft perfekt beim ersten Versuch

Nicht die gleiche Implementierung wie in Python!

Fragen/Diskussionen

- Don't forget
 - Datenschutz..
 - Bias..
 - AI kann sich irren → Halluzinationen, Bias

→ Nur verwenden, was man versteht / verifizieren kann!

Wir suchen Mitentdecker:innen!

Laura Gatto <laura.gatto@cdc.ge.ch>

(Policy Evaluator, Cour des comptes - République et Canton de Genève)

Wolfgang Schneider <wschneider@caritas.ch>

(Advisor & Quality Manager MEAL Caritas Schweiz IZA)