ASIAN EVALUATION WEEK 2024

Innovations for Influential Evaluation

2–5 September | Conrad Hotel Shanghai, People's Republic of China

#AsianEvaluationWeek #AEW2024

Independent Evaluation ADB

AFDI 亚太财经与发展学院 Asia-Pacific Finance and Development Institute

AI in DEval-Evaluations

Sven Harten

Harnessing the power of Generative AI:
Enhancing evaluations and upholding ethics

Presentation outline

1. DEval genAI implentation process (needs, potential and risk assessment → piloting use cases → guidelines → Institution-wide deployment)
2. Use case: leveraging genAI for the analysis of Big Data
3. Ethical issues /Key points of the guideline
4. Conclusion

INTERNAL. This information is accessible to ADB Management and staff. It may be shared outside ADB with appropriate permission.

Data and types of analyses in DEval-Evaluations

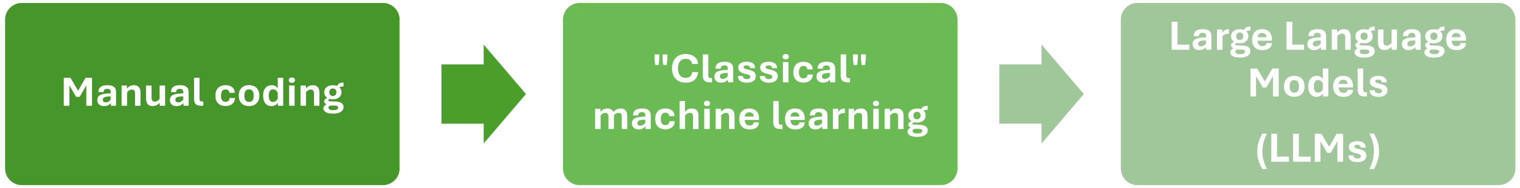
1. Complex evaluations with method-integrated designs
2. Qualitative and quantitative (and causal) analyses
3. Big data
4. Unstructured data
5. Different data types

   (Project databases, financial data, documents, specialist literature, social media, newspapers, interviews, geodata, etc.)
6. Limited time and financial resources

How can artificial intelligence be helpful for analysis?

Text classification with Large Language Models

Development of the possibilities of text classification



Manual coding

"Classical" machine learning

Large Language Models

(LLMs)

Challenge for LLMs:

- **Validity:** LLM doesn't process text like humans, hallucination
- **Reliability:** Output varies according to input and has random elements, black box
- **Objectivity:** LLMs are subject to bias, black box

"Classical" machine-learning

? To what extent have BMZ and state implementing organisations implemented the human rights-based approach?

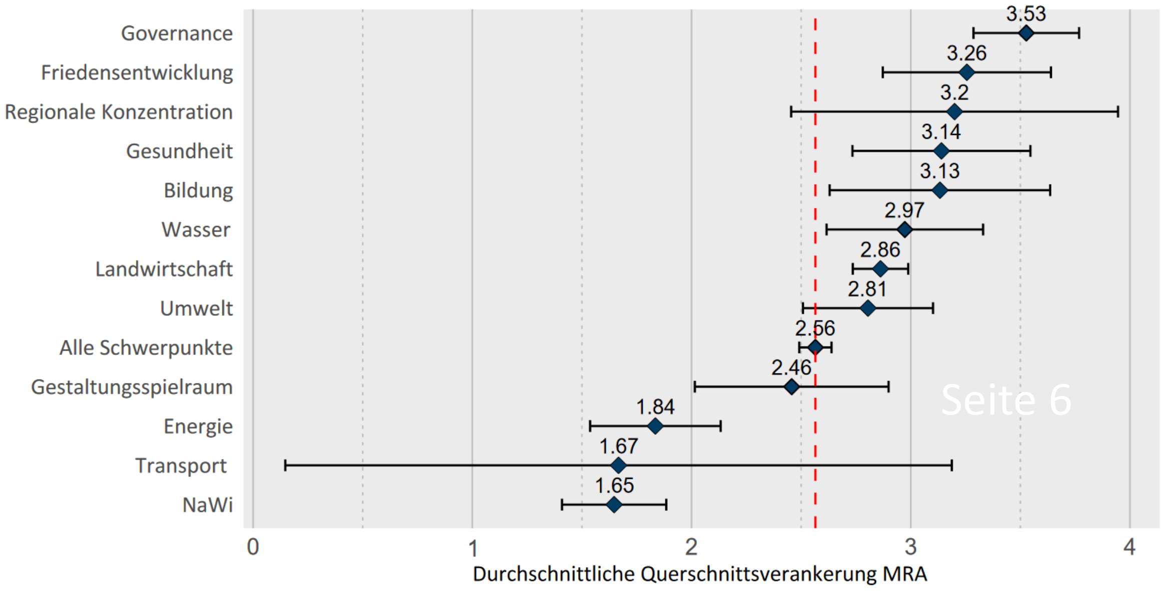
MENSCHENRECHTE
IN DER DEUTSCHEN
ENTWICKLUNGSPOLITIK

Teil 1: Das Menschenrechtskonzept und seine Umsetzung

2021

DEval

Polak et al. 2021



Human Rights Mainstreaming in Projektdokumenten nach Sektoren

"Classical" machine-learning

? What influence does the fragility of the local context have on the success of development projects?

FRAGILE

GERMAN DEVELOPMENT COOPERATION IN FRAGILE CONTEXTS

2019

DEval

Wencker and Verspohl 2019

1 Evaluierungsberichte

2 Entitätenerkennung

The intervention was carried out in Tacloban City in cooperation with the University of the Philippines Visayas.

3 Georeferenzierung

Niekler und Wencker (2019): Text Mining in Evaluation. DEval Policy Brief 1/2019

Possibilities of future use of AI in the evaluation phases

1 Concept, Inception

- Participatory communication tools
- Preliminary analyses on a larger scale

2 Data collection

- Translation
- Transcription
- Interview chatbots
- Big data scraping
- Data preparation
- Anonymization and data security

3 Analysis

- Organizing and summarizing data and documents
- Performing multiple and diverse analyses (qualitative + quantitative, deductive + inductive)
- Prediction

4 Synthesis and reporting

- Automated quality assurance
- Automated reporting, press releases
- bibliography creation
- standardized evaluation

5 Dissemination

- Individualized, interactive presentation of reports(dashboards, chatbots, videos)

Document classification with LLM

? How is BMZ's portfolio structured to support circular economy projects?

Task:

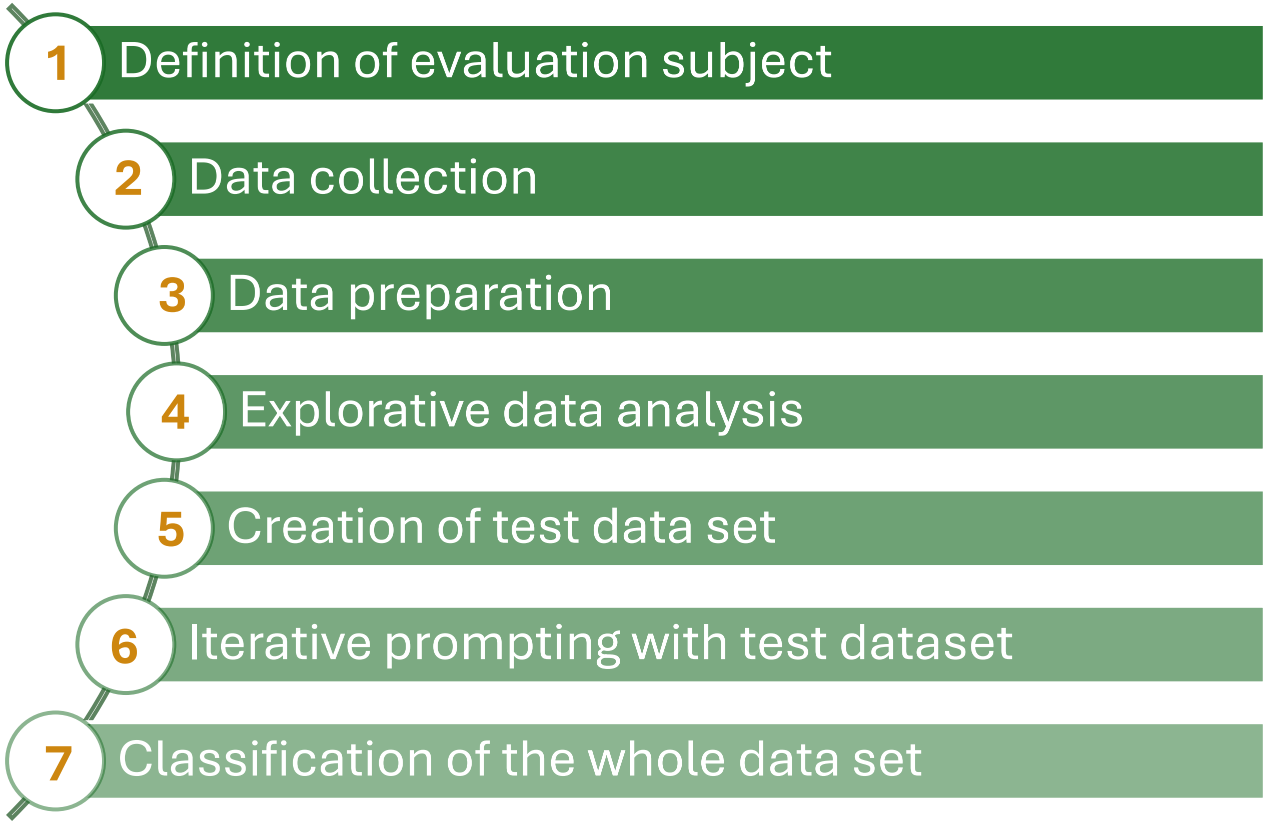
→ **Binary classification of project information from project database**
→ **Circular economy (yes/no)**
→ **Approx. 30,000 projects from 3 years**

Example of a project description for the classification, from OECD-CRS

Circular Economy Solutions to Prevent Marine Litter in Ecosystems Biosphere Protection (2021)

The project "Circular Economy Solutions Preventing Marine Litter in Ecosystems" in India, in collaboration with the MoEFCC (Ministry of Environment, Forest and Climate Change), will carry out technological approaches for tracking and monitoring waste in marine ecosystems. Additionally, the project will work on implementing Extended Producer Responsibility (EPR) to reduce, reuse, and recycle plastics with the involvement of private sector actors such as recyclers and the packaging industry, as well as informal waste workers.

Document Classification Design



1 Definition of evaluation subject

2 Data collection

3 Data preparation

4 Explorative data analysis

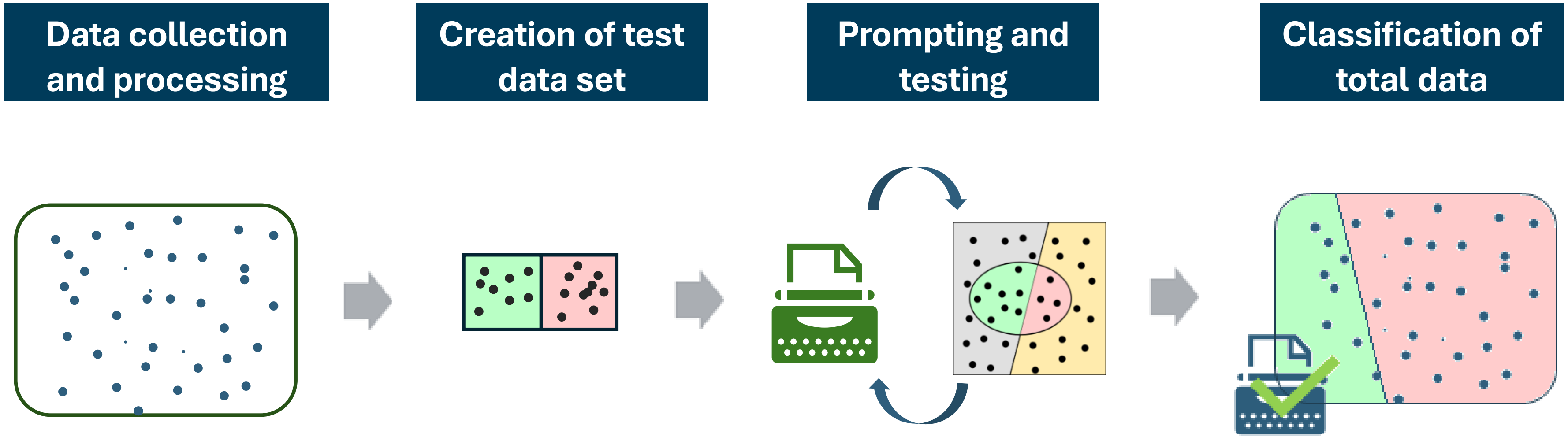
5 Creation of test data set

6 Iterative prompting with test dataset

7 Classification of the whole data set

* Conducted with OpenAI-API and Python

Document Classification Design II



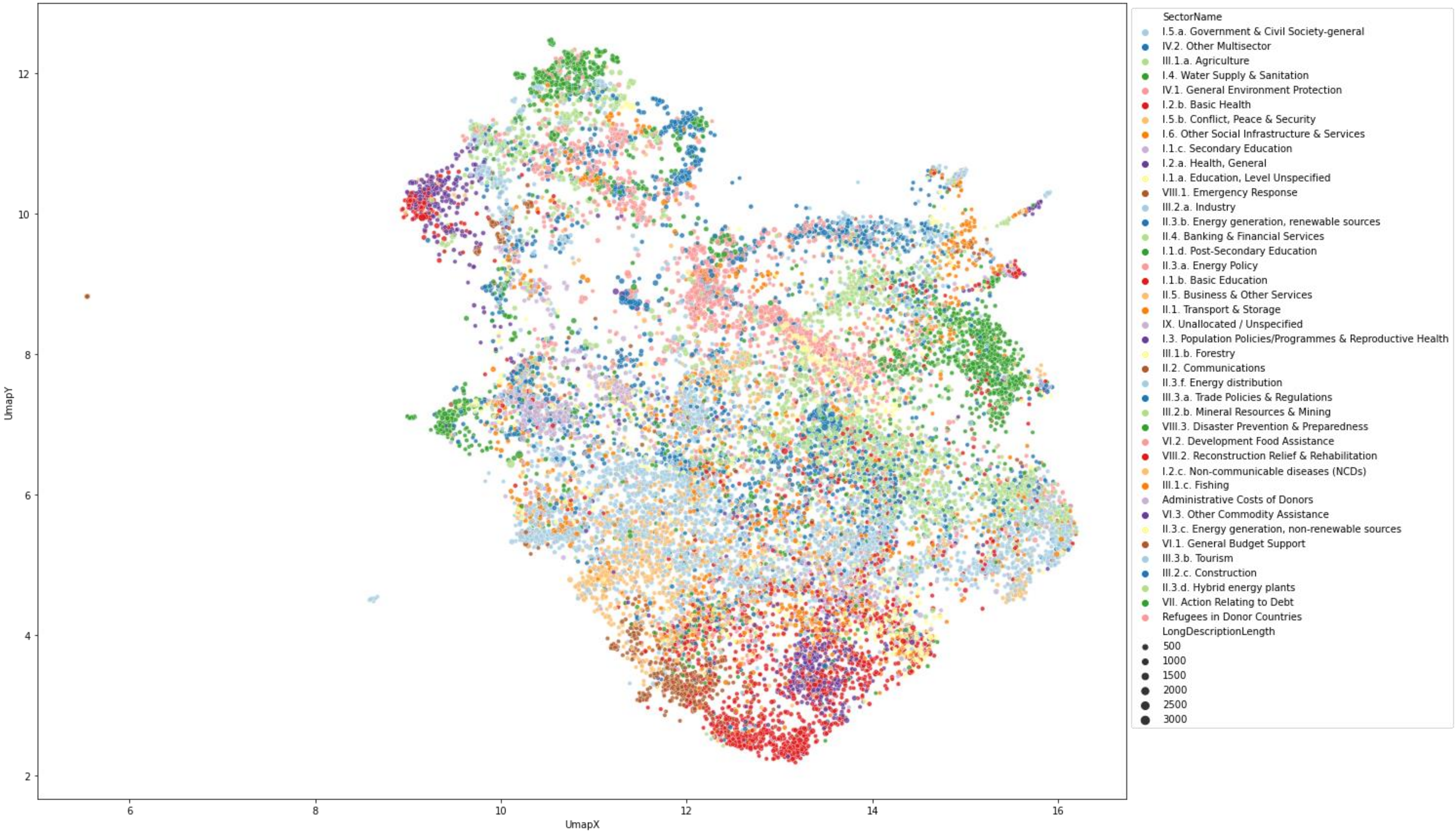
Data collection and processing

Creation of test data set

Prompting and testing

Classification of total data

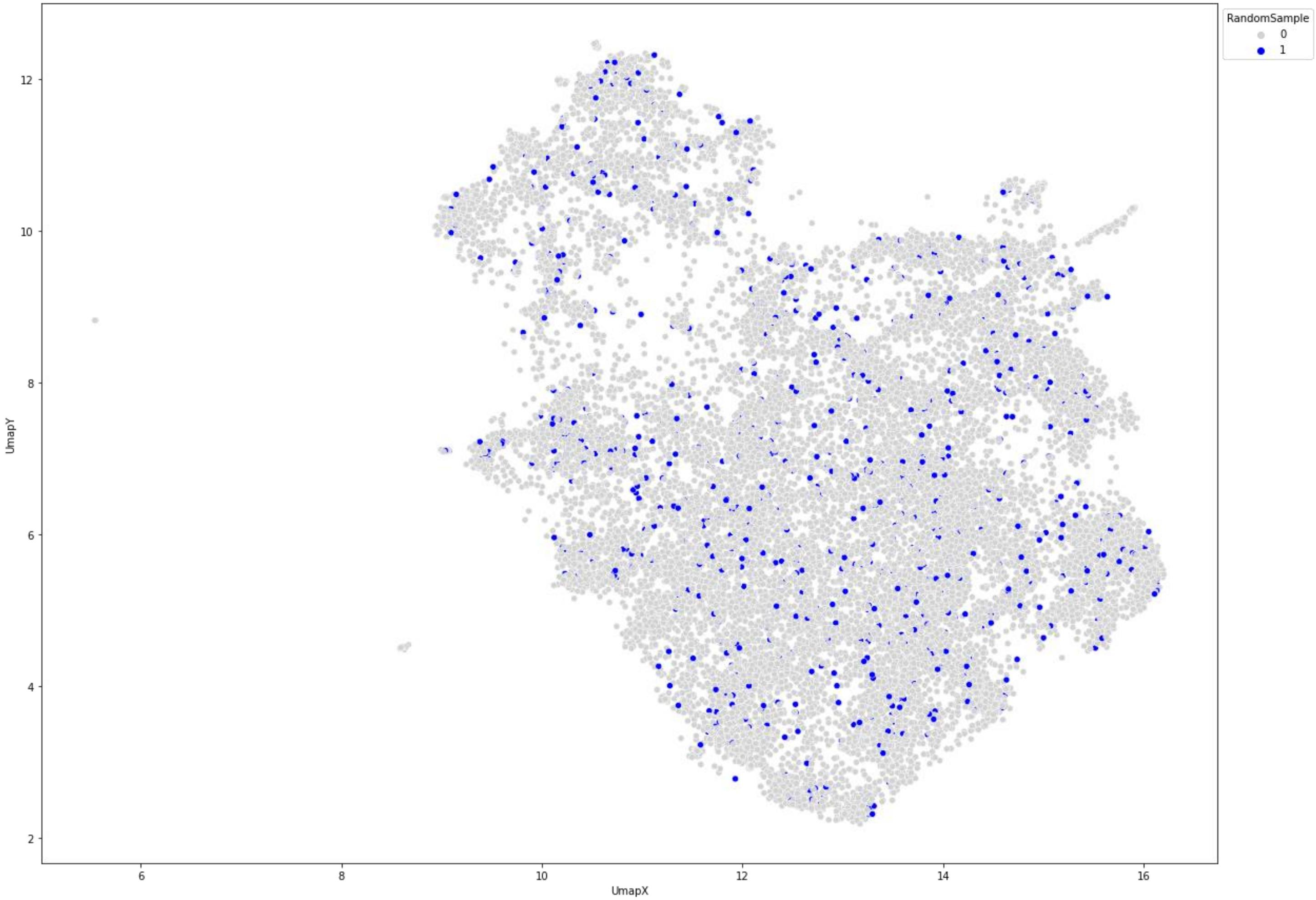
Increase Validity: Explorative Data-analysis and sampling -



“OpenAI’s text embeddings measure the relatedness of text strings.” (Embeddings - OpenAI API)

20.555 Projekte aus der OECD-CRS-Datenbank aus den Jahren 2021 und 2022, markiert nach Sektoren

Increase Validity: Explorative Data-analysis and sampling -



“OpenAI’s text embeddings measure the relatedness of text strings.” (Embeddings - OpenAI API)
20.555 Projekte aus der OECD-CRS-Datenbank aus den Jahren 2021 und 2022

Prompting in Classification Tasks

1. **Context in system message**: e.g. purpose of the task, form
2. **Precise instruction: Clear commands,** omit irrelevant
3. **Step-by-Step Coding Guide:** Clear Steps
4. **Differentiation if necessary:** complex topics in several dimensions
5. **Chain of Thought:** Making the AI analysis process visible
6. **Classification Examples:** Positive and Negative
7. **Creativity parameters:** e.g. adjust temperature
8. **Replicability:** Set Model & Seed
9. **AI self-assessment of the analysis:** certainty of the decision, enough information?
10. **Multiple runs:** with similar prompts, more LLMs

Iterative Prompting: Replicable Chain-of-Thought-Prompting

Projekttitel	Projektbeschreibung	Circular economy		Uncertain or Insufficient Information		
Rehabilitation of decentralised water structures and Waste Management in Syria	Sustainable improvement of drinking water supply, wastewater and waste management to mitigate the impact of the armed conflict on health and quality of life of internally displaced persons and the host community in four governorates.	very likely	because the project involves the improvement of drinking water supply and wastewater and waste management, which are directly related to water and waste management	very unlikely	because the provided information gives a clear indication of the project's focus on water and waste management.	✓
Scholarships for students in Indonesia	Scholarships for students from a poor background at universities of the Federation of Catholic Universities in Indonesia	very unlikely	because the project focuses on providing scholarships to students and does not mention any relation to climate, land, or water management	Some-what unlikely	because the project description is clear and does not suggest any relation to KLW topics.	✗
FAF Agri-Finance Liquidity Facility (ALF)	FAF Agri-Finance Liquidity Facility (ALF)	very unlikely	because the project focuses on providing liquidity for agricultural finance, and there is no direct mention of climate, land, or water management (KLW).	very likely	because the title and description are identical and provide no specific details about the project, making it difficult to confidently categorize.'	?

INTERNAL. This information is accessible to ADB Management and staff. It may be shared outside ADB with appropriate permission.

Lessons learned

1) **Understanding (validity):** Surprisingly high at Gpt-3.5 and 4
2) **Errors in detail:** Usually seem plausible, but errors in detail: especially what is read and how information is processed
3) **Selectivity in complexity:** many possibilities for reasoning a classification decision; Segment long texts
4) **Reliability:** high variation in ambiguous instructions; But: even the same instruction and input can lead to different results
5) **Step-by-step approach:** Separating clear and unclear cases is helpful
6) **Data protection:** Analyses only possible with publicly available data
7) **Coding:** Familiarization with Python a hurdle, but: ChatGPT very helpful
8) **Cost:** 100 reports (10 pages each): GPT 3.5 Turbo: €1.30; GPT-4: €35.91
9) **Preprocessing:** helpful for large amounts of data, including cost and time savings

Example prompt for classification

Please perform the review step by step (from a to f) based on the operationalization below.

a) If the information was collected through various methods of data collection, assign complied with. This could be, for example,

b) when an interview and a survey have been conducted on a fact.

c) if various statistical data or indicators are listed for a fact.

d) Data are compared with data from neighbouring countries.

e) Data from one source can be compared to different years.

f) If data is collected through the involvement of different actors, assign the criterion as met. The involvement of different actors includes, for example, the questioning of different stakeholders and actors. If one of the operationalizations (a, b,c, d, e, or f) is met, categorize the passage as "satisfied." The fulfillment of one aspect is sufficient, even if the others are not fulfilled.

Evaluation standards

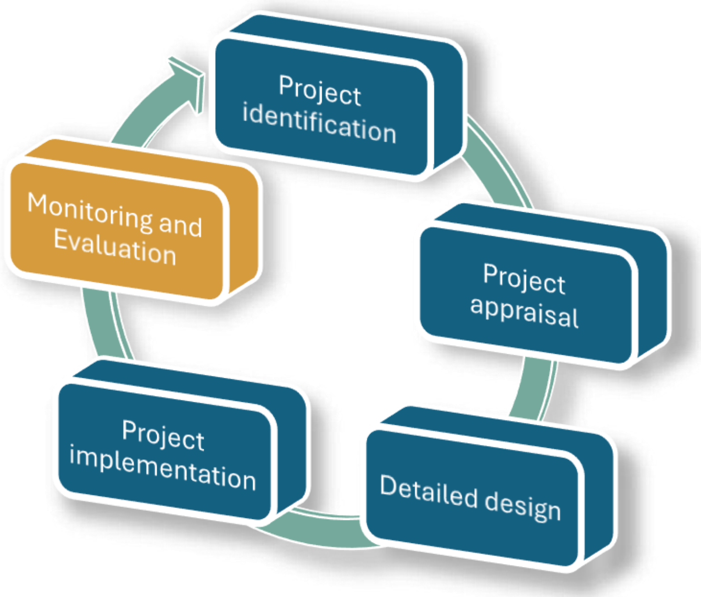
A. **Accuracy, scientific rigour and comprehensibility**

B. **Utility**

C. **Fairness, independence and integrity**

D. Evaluability

E. Comparability



A) accuracy, scientific rigour and comprehensibility

Potentials

- Standardization of evaluation methods and processes through AI application might improve
  - **Objectivity**
  - **intersubjective comprehensibility**
  - **Replicability**
- Combining multiple data collection and anlysis methods can improve the **accuracy** and **validity** (e.g. multiple robustness checks, source verification, error checking, broader analysis)

Risks

- As long as AI remains black boxes, the lack of transparency and traceability prevents **intersubjective comprehensibility**
- Lack of **accuracy** of the output (e.g. hallucination) limits reliability and thus credibility.
- Reproducing bias in the data (e.g. discrimination) minimizes **objectivity**
- **Credibility and competence**: lack of skills, such as intercultural or conflict sensitivity, or mathematical reasoning

a) Credibility and competence of the evaluation team **b)** Theory-based, evaluation design and selection of methods **c)** Context analysis **d)** Use and generation of valid and reliable information and transparent citation of sources **e)** Reasoned conclusions based on scientifically sound methods **f)** Quality assurance and systematic error checking **g)** Data analysis quality h) Completeness of reporting **h)** Replicability and comparability **i)** Consideration of standardized criteria, questions and rating scales

B) Utility and participation

Potentials

- Increased **utility** through individualized presentation of data and results.
- Better **participation**: By using AI, communication channels to stakeholders can be improved more diversely; knowledge and competence asymmetries can be reduced.
- Acceleration of evaluation can lead to more **timely results** for dynamic management processes.
- More **efficient implementation monitoring**

Risks

- Neglecting face-to-face interaction through the use of low-cost AI systems could lead to a **decline in awareness of relevant** issues.
- Focusing on AI technology could amplify digital divide in **participation**
- Evaluators compete in a market flooded with AI products

a) A relevant evaluation subject is defined. **b)** Recipients and stakeholders are identified. **c)** Recipients and stakeholders must be involved in the evaluation process. **d)** Evaluation results must be made usable and disclosed. **e)** The evaluation results are timely. **f)** Implementation planning and monitoring.

C) Fairness, independence and integrity

Potentials

- **Identification of biases** of the evaluation team
- **Independence and impartiality** through standardized algorithms
- **Transparency** through individualized and comprehensible visualization of evaluation processes and results
- Automated **data protection and anonymization**
- **Simulations**: Ethically questionable evaluation designs can be replaced by simulations. (e.g. experiments with agent-based modelling)

Risks

- Bias in the training data could compromise **independence and impartiality**
- **Lack of transparency** of analysis and decision-making processes due to black-box nature of AI.
- **Less independence in the project/policy cycle** due to the replacement of tasks by AI
- Transfer of **ethical actions and decisions** to an automatized AI algorithm
- Risk in **data protection** when using external (commercial) AI systems (processing, security gaps)
- Big Data vs. **data economy**

a) Independence of the evaluation **b)** Predictability and planning security **c)** Ethical approach **d)** Data protection and data economy **e)** Evaluation transparency **f)** Impartial and independent conduct and reporting **g)** Disclosure of values

Elementes of DEval‘s AI-Guideline

1. Data protection and confidentiality
2. Accuracy and error checking
3. Generative AI as a source of facts
4. Transparency
5. Replicability
6. Verifiability
7. Bias
8. Ethics
9. Responsibility

10.Copyright

11.Prompting for interaction with AIs

Conclusion

- **Acknowledge** that a significant technological change is ahead
- **Working together and share knowledge** to adapt to the fast change (e.g. non-commercial AI, use cases and limits, prompt engineering)
- **Risk minimization** at different levels:
  - Individual level (e.g. education)
  - Technical level (e.g. traceability through Explainable AI, Hallucination, Bias, Privacy)
  - Evaluation level (e.g. error checking routines)
  - Organizational and Policy level (e.g. Regulation and standard-setting, resources for R&D)
- **Pragmatism:** Consider AI as an additional team member

Thanks for your attention! Happy prompting ;)

BACKUP

Weitere Folien

Risiken

1) **Big Data** vs. **Datenschutz**: sind gängige Verfahren des Umgangs mit sensiblen Daten noch ausreichend?

2) **Verzerrungen und Vorurteile** aufgrund der Trainingsdaten- und prozesse (Bias)

3) **Intransparenz, fehlende Nachvollziehbarkeit** durch **Blackbox** Künstlicher Intelligenzen

4) Scheinbar menschliche Kompetenzen mit **nichtmenschlichen Fehlern** (u.a. eingeschränkte Logikkompetenzen, Halluzinationen, kein Zugang zur „realen“ Welt)

5) Abgabe von **Verantwortung und Kompetenz** durch Automatisierung

6) **Erosion des Vertrauens** in etablierte Wissensnetzwerke durch KI-Einsatz (u.a. konkurrierende KIs, Deepfakes)

7) **Sozialökologische Nachhaltigkeit** (Probleme mit Urheberschaft des Trainingsmaterials, Entwicklung der KI durch Ausbeutung in Niedriglohnländern, CO^2-Abdruck)

8) Herausforderungen für **Evaluierungsstandards** (z.B. Nachvollziehbarkeit, Transparenz, Unabhängigkeit)

Summary

- AI is **revolutionizing the workplace** with great leaps in innovation in a very short time.

- **Core tasks of the evaluation process** (analysis, information, knowledge production) are likely to be supported/replaced by an AI

- Nevertheless, there are activities that will be **difficult to be replaced by an AI** (e.g. responsible behavior, networking)

- **Technical limitations**, e.g. Hallucination, bias, advanced reasoning

- Potentials and risks:
  - AI can contribute to a more standardized evaluation process
  - Increase in utility through improved participation.
  - Risks related to accuracy, transparency, independence, data protection, and reproduction of biases