ASIAN EVALUATION WEEK 2024

Innovations for Influential Evaluation

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

#AsianEvaluationWeek #AEW2024

Independent Evaluation ADB

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

Responsibly harnessing the power of Generative AI in UN evaluations

Enhancing evaluations and upholding ethics

Marco Segone
Director, UNFPA Independent Evaluation Office

4 September 2024

Session takeaway

In this session, you will learn:

- How to navigate and uphold ethical considerations in GenAI use in evaluation.
- Practical use cases for harnessing GenAI's capabilities to optimize evaluation practices.
- The importance of the evaluation community in shaping responsible and ethical use of GenAI in evaluation.

The journey

1 **How did we begin?**
AI use cases in evaluation

2 **What is our approach?**
Key features of the Gen-AI powered evaluation strategy

3 **Where are we headed?**
AI pilot in a centralized evaluation and inter-agency meta-synthesis

4 **What did we learn?**
Lessons for evaluation managers and evaluators

5 **Should we embrace or resist GenAI?**
The next steps

1

How did we begin?

Mapping use cases of AI tools

Needs assessment across the evaluation lifecycle

Solution exploration study, based on UNFPA approved Google tools and platforms (Duet AI, Gemini)

2

What is our approach?

Strategy for Gen-AI powered evaluation function at UNFPA

- A pioneering strategy for leveraging the benefits of responsible and ethical GenAI while minimizing risks
- Outlines a roadmap to optimize evaluation processes and products with ethical and responsible use of GenAI
- Focused on achieving greater efficiency, effectiveness, and timeliness in evaluations

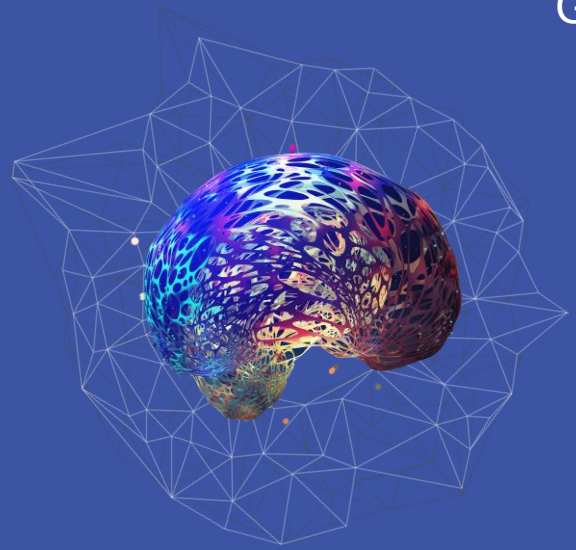
Aligned to UNEG ethical principles for AI use in evaluation (Draft)

- Do no harm
- Transparency & Accountability
- Fairness & Mitigation
- Participation & Inclusiveness
- Privacy & Data Protection
- Accuracy & Reliability
- Human-Centered Approach
- Upholding Human Rights

Strategic principles for leveraging GenAI in evaluation

Demand-driven approach for GenAI-powered evaluation

Diversification and innovation of GenAI tools

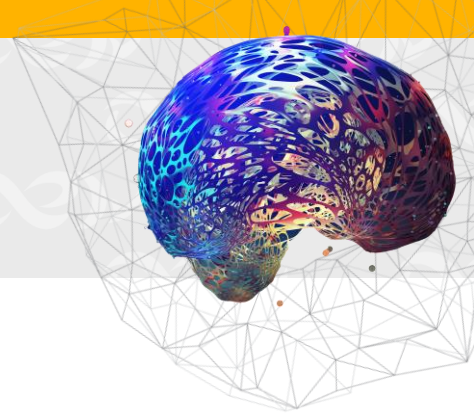


Upholding quality and credibility in evaluation

Adhering to an ethical and human rights-based approach to GenAI use in evaluation

Cultivating GenAI capacity in evaluation, especially in the Global South

Implementation roadmap for GenAI-powered evaluation



- Phased approach to GenAI adoption and deployment
- Developing custom GenAI solutions for evaluation
- Change management and communication
- Iterative and adaptive approach to digital transformation
- Long-term sustainability of GenAI-powered evaluation effort

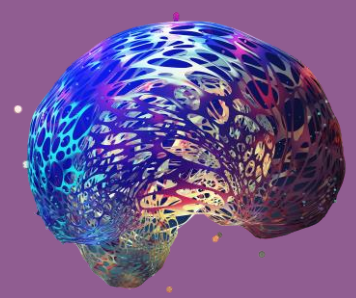
3

Where are we headed?

Tool: AILYZE

- Pilot of AI use in evaluation of UNFPA strategic plan
- Pilot of AI use in an inter-agency meta-synthesis exercise

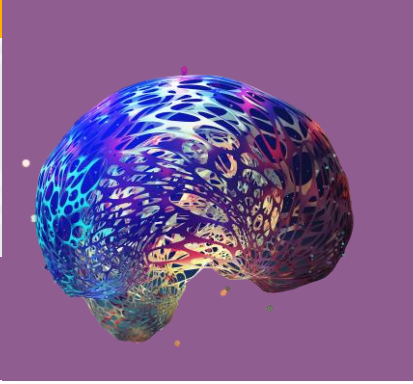
1. Pilot of AI use in evaluation of UNFPA strategic plan



Use case

Qualitative analysis of 150 documents to determine the extent to which Country Programmes are aligned to the corporate Strategic Plan priorities

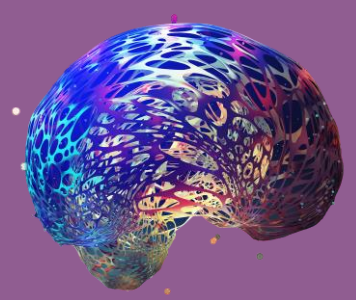
2. Pilot of AI use in UN interagency meta-synthesis to advance the UN Youth Strategy



Use case

- **Report extraction**
  - Keyword search to distil 1348 reports from evaluation databases of 46 UNEG agencies.
  - AI analysis extracted 253 reports, 97% accuracy based on human verification; final dataset of 302 reports from 13 agencies, with manual addition
- **Pilot analysis with 15 reports, 5% of the sample**
  - AI model guided by a conceptual framework and coding structure
  - 3 rounds of pilot testing, with regular AI code checks and validity checks by humans
  - High accuracy in coding and data extraction, with minor false positives/negatives due to concept complexity
- **Coding and synthesis work**
  - Coding structure applied to the full sample and frequency analysis undertaken to identify key patterns/themes
  - AI generated content analyses for selected patterns/themes
  - Humans review and synthesis of key lessons under way

NEW! AI scorecard tool to gauge the pilot results & lessons towards scale up



Metrics

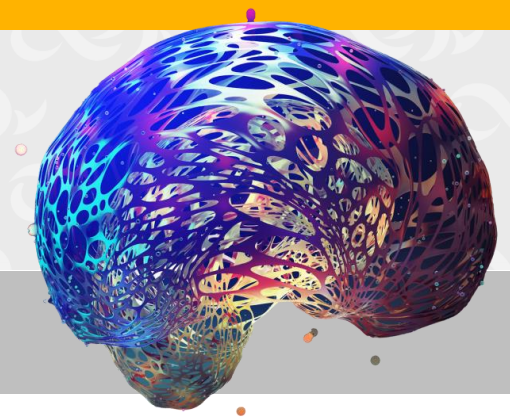
1. **Quality of results**: Accuracy and reliability, deeper and nuanced analysis, actionability
2. **Efficiency**: Time savings, Cost benefit
3. **Ethical considerations**: Fairness/bias mitigation, transparency/explainability, human-AI collaboration, data security & privacy
4. **User experience**

Draft AI scorecard for evaluation of the UNFPA strategic plan 2022-2025

	Metric Category	Specific Metric	Description	Measurement	Risk mitigation	Results	Lessons Learned (strategic/process)
1	**Quality of results**	Accuracy and reliability	Degree to which AI-generated results match expert human judgment & analysis, and produce consistent results	Human check conducted on an initial sample of 10 CPDs; further refinement and analysis conducted in two further rounds. Discussion on definitions and discrepancies between AI and human interpretation conducted.	Human review and analysis to "verify" AI results; Identification of limitations with the AI-generated results, including the analysis of highest and lowest performers;	Medium	Nuanced conceptual framework and definitions; time for human verification
		Deeper analysis	Extent to which AI could generate deeper/more nuanced analysis vs human produced analysis	AI-generated analysis of 'highest' and 'lowest' performing Country Offices verified but confidence levels were very low.	Deeper analysis considered flawed and too risky to use. It was discontinued after the pilot phase.	Lower quality analysis	Limitations of data scientist (level of understanding/level of prompting)
		Actionability	Extent to which AI results lead to meaningful contribution to the evaluation/synthesis exercise	Final analysis of CPD operationalization of shifts and accelerators verified and resulted with a medium level of confidence in the results.	Limitations of final analysis included in the methodology annex of the report; Results from the AI-generated analysis used in the evaluation report - both with and without triangulation	Medium	Time for additional rounds of verifications
2	**Efficiency**	Time savings	Comparison of total time spent on report extraction and qualitative analysis by AI vs if the tasks were conducted by humans	Due to the possibility of utilizing AI, the data-set/scope of the analysis was significant, making it challenging to establish a direct time comparison against a purely human-led analysis for the same scope.		Add value of approximate working hours saved or increased, or mention no difference as applicable	Possibility of using the tool expanded the scope; if the tool was not used the analysis design would have been changed
		Cost benefit	Comparison of financial costs of commissioning an AI tool + human time for developing AI frameworks & verifications; vs a full human run exercise	AILYZE cost for the project $1500. Human cost for the full scope $21,700	Extraction: 2 hours* 144 CPDs/7.5 hours/day * $500/day = $19,200 Analysis: 5 days * $500 = $2,500 Total = $21,700	Saving approx 20,200$	Significant cost savings
3	**Ethical considerations**	Fairness/Bias Mitigation	Extent to which AI avoided discriminatory outcomes or perpetuating existing biases in its analysis	Assessment of publicly-available, approved documents presented to the Executive Board resulted in a high degree of confidence in the fairness and absence of bias in the content.	Reduced need for bias mitigation but human review needed given the absence of information on the actual prompts used.	High	Verification needs change based on type of data set, review the prompts
		Transparency/Explainability	Degree to which AI processes were understood by the human team and partners, and explained transparently in the report	Broad processes to be undertaken was explained by contractor but the specific prompts and models used was not divulged.	Lack of transparency led to increased risk mitgation at different stages, e.g. regular check-ins to understand the process/model, drafting a methodological note by AILYZE for explainability; ensuring the evaluation report fully declares the use of AI and AI tools in the methodology	Low	More transparency on models at the contractual stage
		Human-AI Collaboration	Effectiveness of the collaboration between human evaluators and AI	Evaluators interacted with AI through an interlocutor of James; agreement reached to pilot the analysis on a small set of 10 CPDs; he was open and flexible but couldn't predict accurately how many rounds would be needed to complete the analysis; human feedback improved the AILYZE learning - both for the project manager and for the models themselves.	Human-AI collabaoration articulated strongly in contracts, and in accountability of AI company and human contracts. Many meetings to discuss with Allyze	Medium	Collaboration was positive, but there was low transparency, due to which the collaboration was affected; Strong human-AI collab helps to offset over reliance on AI
		Data security & privacy	Degree to which AI model upheld data security & privacy	The chosen dataset (CPDs) already have a low risk on data protection as they are externally available documents	Contractual obligations set on data protection and privacy according to UNFPA policies; trained models deleted	High adherence	Depending on the project, risk mitigation measures should be instituted in the contracts; be aware of UNFPA data protection policy
4	**User experience**	User satisfaction	Overall perception of evaluators regarding AI usability and helpfulness	Even given the lack of transparency and limitations in the results received, the results were used in the evaluation report and were seen as being a helpful and useful addition to the evidence base and overall analysis.		3 (Good)	

4

What did we learn?

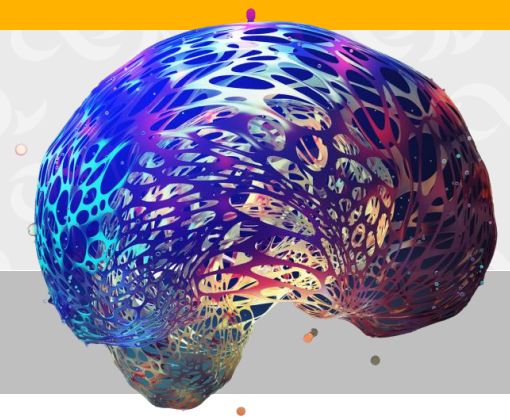


1. Humans are accountable, not machines!

- Incorporate **ethical AI use clauses** into contracts
- Ensure **transparency** regarding AI models at the contractual stage
- Include **AI disclaimer** in report and explain **AI use in the methodology**
- Prioritize **data protection** measures
- Consider time for **human verification and oversight** across multiple rounds of analysis

4

What did we learn?



2. It’s an investment. Be intentional in experimenting and adapt

- **Facilitate collective mindshift** towards AI
- **Upskill** evaluators on AI skills and **build capacities**
- **Clarify scope and definitions** tailored for AI analysis
- Internal data scientist enhance **prompting capabilities**
- Keep track of **efficiency gains** to support scale up of the tool (AI scorecard)
- Be **clear on the initial investment** (financial, human and time)

5

Should we embrace or resist GenAI?

- Human rights, transparency and accountability must light the way, including for GenAI-powered evaluation
- GenAI must benefit everyone, including the one-third of humanity who are still offline
- Evaluation community must take deliberate steps to leverage ethical and responsible GenAI in evaluation

Let’s shape together ethical and responsible GenAI for evaluation

18

UNFPA | IEO Independent Evaluation Office

Access the strategy

GenAI-powered evaluation function at UNFPA

United Nations Population Fund
Independent Evaluation Office

605 Third Avenue
New York, NY 10158 USA

unfpa.org/evaluation
evaluation.office@unfpa.org
@unfpa_eval
@UNFPA_EvaluationOffice