



# The consequences of AI for evaluation practice in the future

Svenska utvärderingsföreningen 20th anniversary conference - keynote

Steffen Bohni Nielsen

8 December 2023





# Agenda

1. What characterizes the evaluation industry?
2. What is AI in the context of evaluation?
3. How will evaluation practice be affected?
4. What does the future hold?



# 1. What Characterizes the Evaluation Industry?



# Key Features of the Evaluation Industry



## Market dynamic is demand driven

Sensitive to shifting government priorities and sourcing strategies



## Motley crew of providers

Often consortia of SME, evaluation methodology experts



## Public sector is a dominant procurer

Procurement, management and the practice of evaluation services have been increasingly institutionalized



## No globally dominant providers

This is seen in adjacent professional service fields such as auditing and management consulting.



## The evaluation market is segmented and multiple-layered

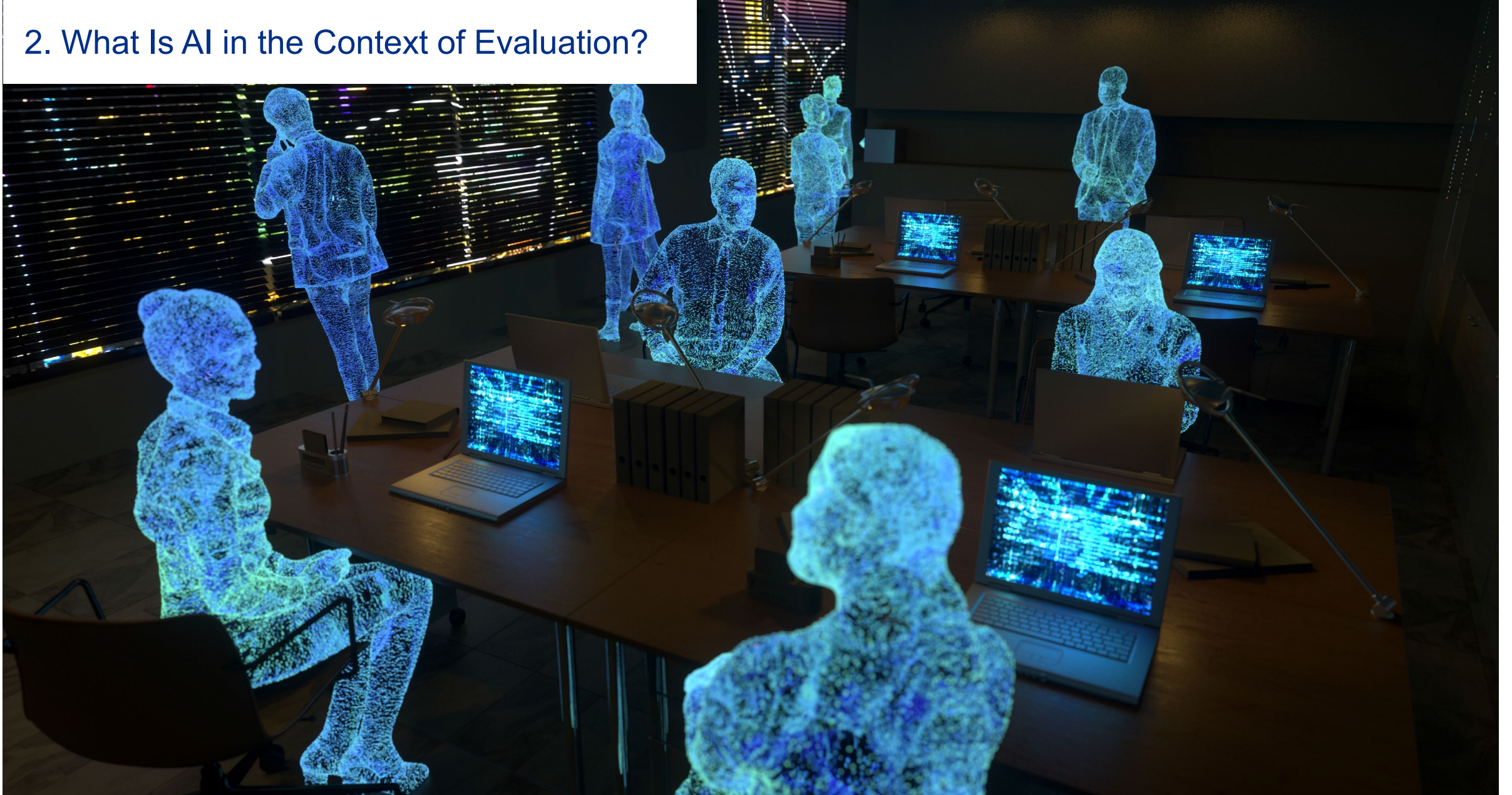
Differentiators are national, regional, type of client, domain or methodological



## No definable market and easy access

VOPEs don't track market trends, size, and shares. Competing knowledge forms

## 2. What Is AI in the Context of Evaluation?



“

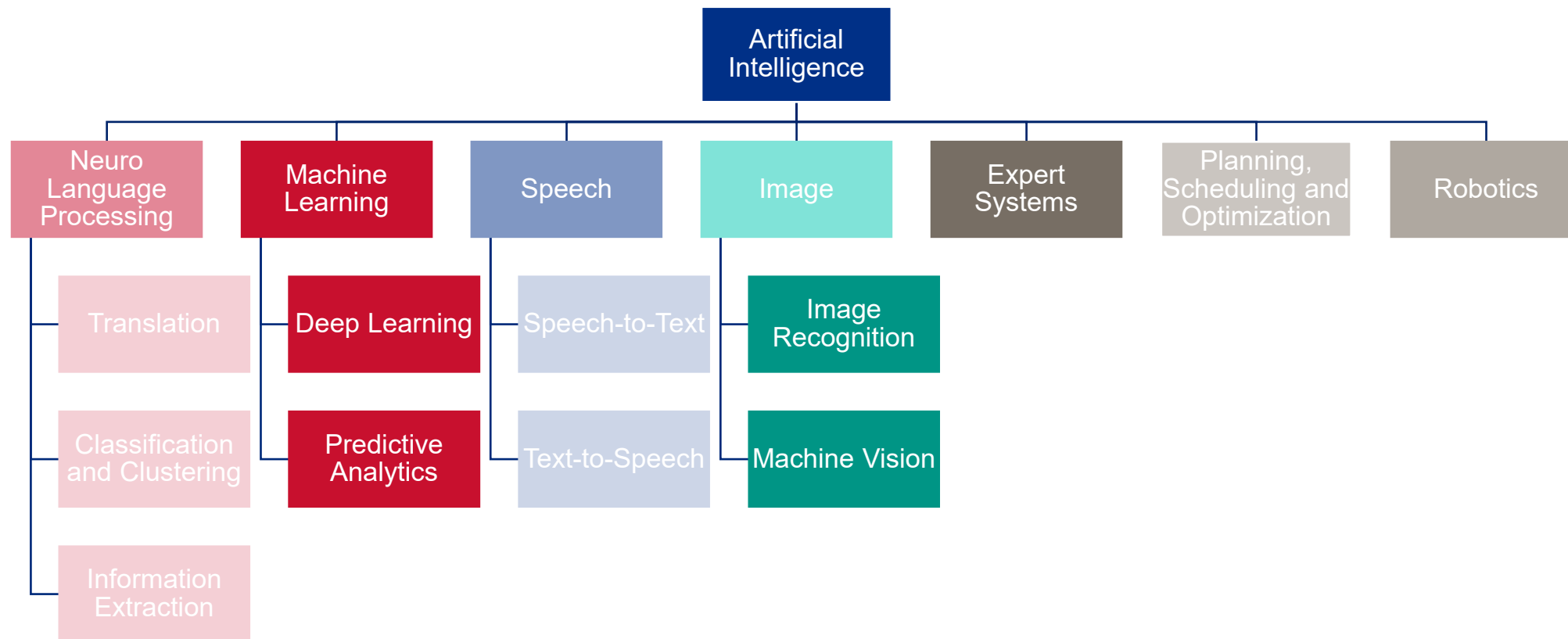
Think of Generative AI  
(currently) as a hyper  
intelligent, super creative,  
extremely knowledgeable,  
and totally unreliable  
assistant

Dr. Thomas Terney





# Types of Artificial Intelligence







# Emerging Technologies available for program evaluation

## (New) Kinds of data

- Satellites and drones
- Social Media
- Radio call-in programs
- Internet searches
- Mobile phones
- Telecom data records
- Program administration data
- Systems data
- Large-scale survey data
- Textual data
- Internet of Things (IoT)

## (New) Kinds of data storage and organization

- Distributed Ledger Technologies
- Cloud Computing
- Edge Computing

## (New) Kinds of data processing

- Machine Learning and Artificial Intelligence
  - For quantitative analysis
  - For text analysis
  - For image analysis
  - For network analysis

# Mature and Immature AI powered solutions

## QUALITATIVE DATA ANALYSIS SOFTWARE

QDA Miner is an easy-to-use qualitative data analysis software for organizing, coding, annotating, retrieving, and analyzing collections of documents and images. QDA Miner qualitative data analysis tool may be used to analyze interview or focus group transcripts, legal documents, journal articles, speeches, even entire books, as well as drawings, photographs, paintings, and other types of visual documents than any other qualitative research software on the market but also more reliably. QDA Miner's seamless integration SimStat, a statistical analysis tool, gives you unprecedented information, including numerical and categorical data.

QDA Miner offers higher levels of computer-assistance than other qualitative data analysis software on the market

### CoLoop: The AI copilot saving your qualitative

Dive deeper and use AI. Summarise, and faster than your re

## Understand people in their own words

Get insights from nuance and accuracy human capacity context specific and employee fe

## What can you do in Covidence?

- Import citations
  - Use machine learning
  - Screen titles & abstracts
- Covidence works seamlessly with your favourite reference managers like EndNote, Zotero, Refworks, Mendeley or any tool that support RIS or PubMed formats.
- Integrated with the Cochrane Randomized Controlled Trial (RCT) classifier, which allows you to quickly and accurately filter out the studies that are not RCTs.
- Breeze through screening with keyword highlighting & a lightning quick interface. Covidence keeps full records of who voted and also supports single or dual screeners.
- Covidence also automatically uploads open access studies in full text review.



### Academic Analyzer

Analyzes academic articles, summarizing research topics, methods, and conclusions.

By community builder



### ResearchGPT

AI Research Assistant. Search 200M academic papers from Consensus, get science-based answers, and draft content with accurate citations.

By [consensus.app](https://consensus.app)

PRACTICAL GUIDE TO BOOST RESEARCH			
Use Cases & Prompting Techniques			
<b>Research Proposal Writing</b> - Assists in writing research proposals by suggesting a structure, providing example texts. - Helps to write certain parts of the proposal.	<b>Data Analysis</b> - Generates comments/prompts for data analysis in SPSS/Excel/Python/R. - Helps to design or optimize existing scripts.	<b>Literature Review</b> - Summarizes key insights from large volumes of scientific literature or databases. - Helps to keep up with the latest research.	<b>Grant Writing</b> - Assists in writing and editing grant applications. - Helps in developing a compelling narrative, budget justification, and exploration of potential impact of research.
<b>Public Health Communication</b> - Drafts clear, concise, and understandable public health communication materials. - Translates complex medical jargon into layman's terms.	<b>Epidemiological Modelling</b> - Drafts the logic behind the model. - Assists with coding, or in transforming potential applications or hypotheses.	<b>Reviewing and Critiquing</b> - Helps understand, analyze, and critique epidemiological studies, including their methodologies, results, and conclusions.	<b>Learning Complex Topics</b> - Explains complex methodologies or concepts from papers in an easy-to-understand manner for beginners. - Simplifies complex theories for quick understanding.
<b>Title Brainstorming</b> - Provides title suggestions for your research paper or presentation based on the provided topic. - Stimulates creativity & helps overcome writer's block.	<b>Presentation Preparation</b> - Assists in creating presentation outlines. - Suggests ways to visually represent complex data.	<b>Writing Improvement</b> - Proofreads, checks grammar and spelling, and suggests improvements for clarity. - Can mimic user's writing style to maintain consistency.	<b>Generating Literature Review Flow</b> - Assists in structuring review logically & comprehensively. - Ensures all key themes are covered, providing a coherent overview of the subject.

If you have used ChatGPT, specify the exact AI model and its version date: [Model Name], [Specific Version/Model Variant], [Date of Version], [Additional Parameters if applicable]  
 Example: "GPT-4, ChatGPT, April 22, 2023, Temperature 0.8"

**Research Plugins**

- Academic Paper Search: Scholarly, Scholar Assist, Consensus Search, Scholar AI, Xpapers
- Extracts Data from Papers: AskYourPDF, Mixerbox ChatPDF, ChatWithPDF, AI PDF
- Diagram Generation: Mixerbox Diagrams, Whimsical Diagrams, Skrive, Show Me Diagrams
- Bibliography Management: Bibliographic Crossref, Literature Mapping, Litmaps

### 3. How Will Evaluation Be Affected?



# Key features in the Proliferation of AI in the Evaluation Industry



## Evaluation providers' competitive strategies

Value propositions: Quality, time, price



## Appropriateness of the technology

Evaluation question - Sources, types of data, analyses



## Nature of the evaluation service

M&E systems, evaluation studies, evaluation capacity building



## Size and duration of evaluation contracts

Larger vs. Smaller. ROI

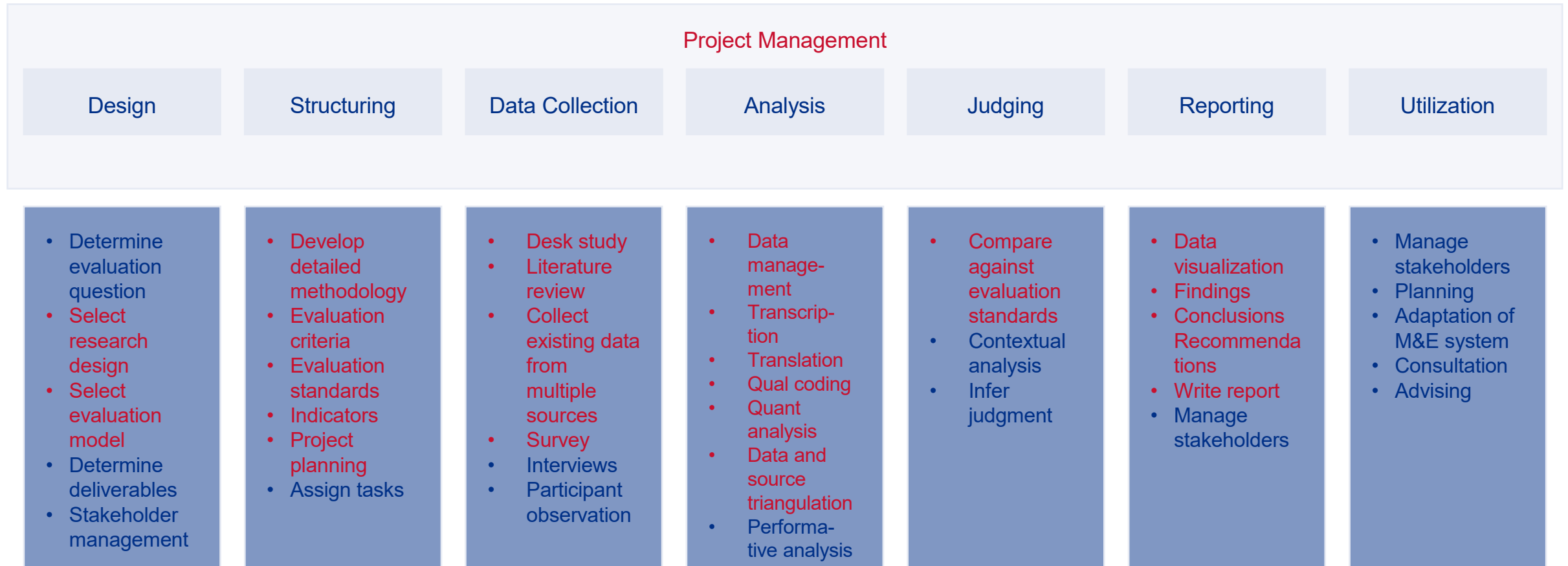


## Breadth and depth of the evaluation provider's capability

Different methodologies and disciplines

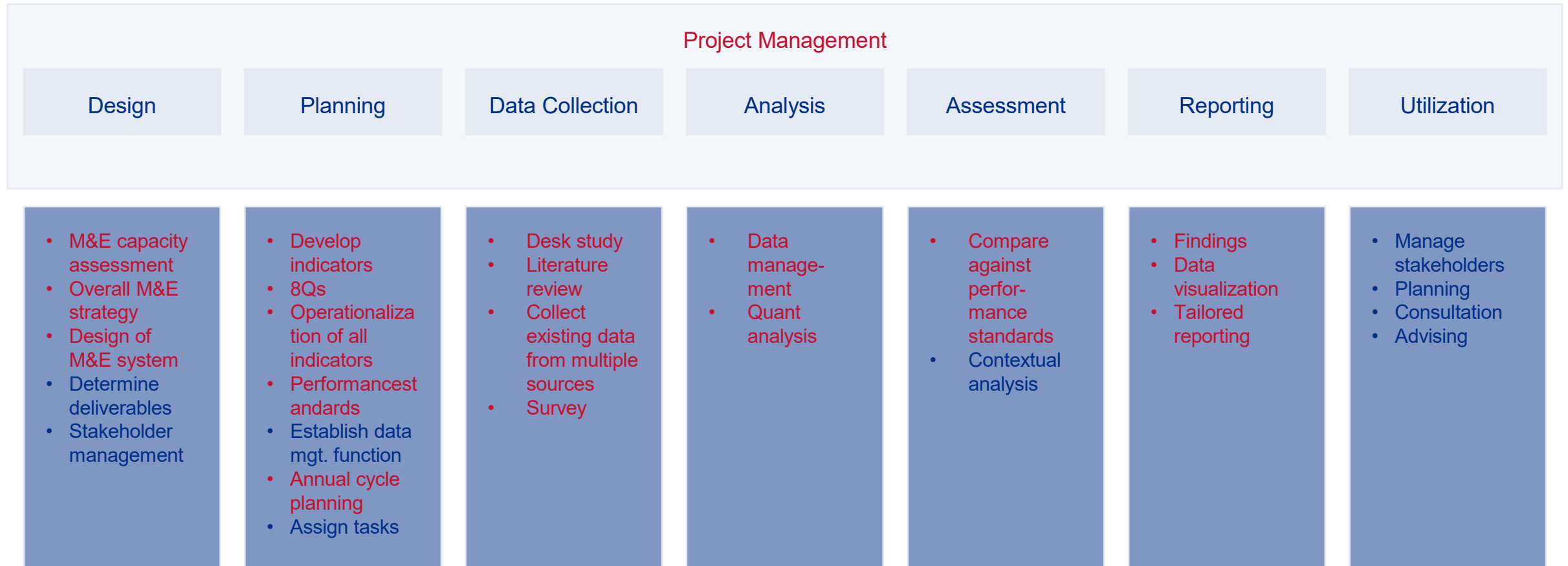


# Evaluation Studies – Task level breakdown





# Monitoring and Evaluation System – Task level breakdown

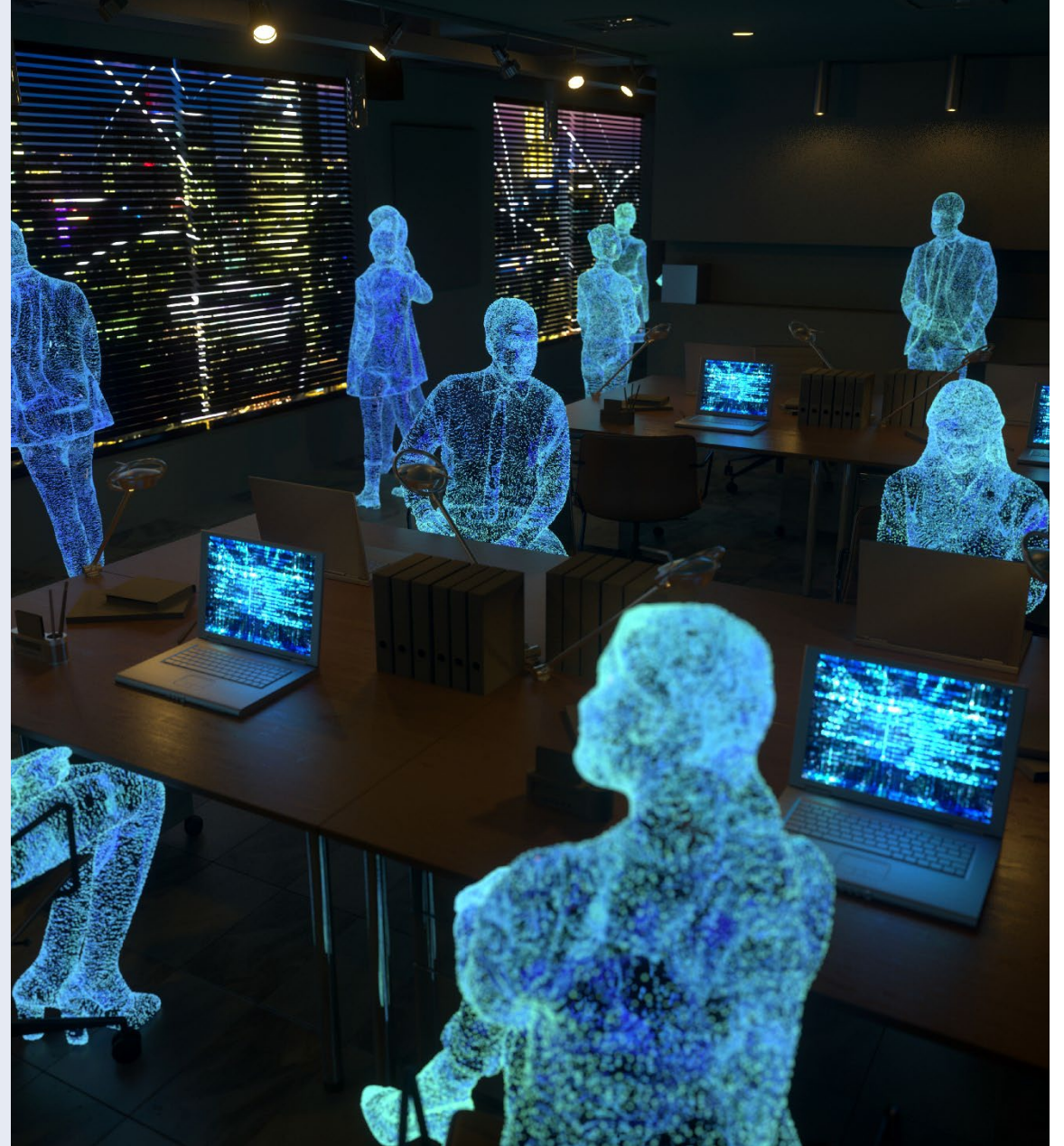


“

## Thematic coding of qualitative data

Using CoLoop to analyze 10 transcripts produced remarkably similar subthemes in 10–20 seconds compared to nearly a dozen hours spent coding, reviewing, and summarizing data manually.

Sabarre, Beckmann, Bhaskara & Doll, 2023:63





“

## Coding open ended survey questions

...Avalanche produced more granular themes than those identified manually. While less frequently cited, most of these could be appropriately grouped as subthemes under our manually produced themes. There were only a few instances where Avalanche did not identify a manually generated theme...”

Sabarre, Beckmann, Bhaskara & Doll, 2023:65







“

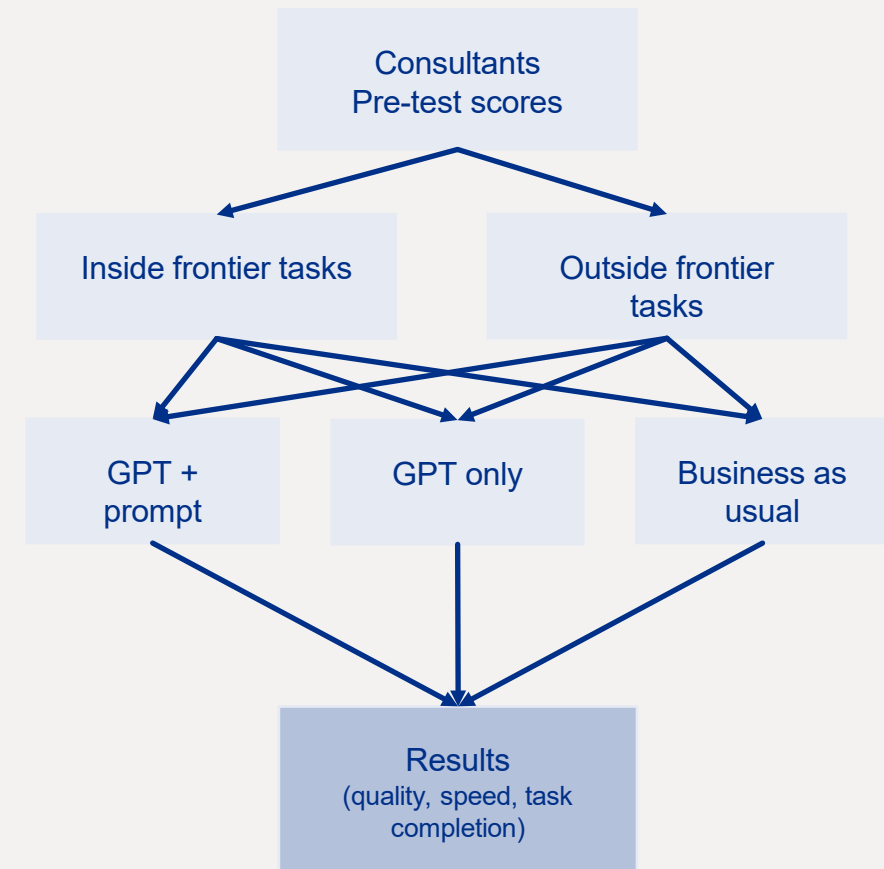
Competencies that are highly social and highly creative and strategic—which may allow us to retain our specialized expertise as evaluators

Sarah Mason, 2023: 20



# Study on Generative AI interaction in management consulting

- Collaboration experimental study: Harvard, MIT and BCG.
- Study population: Boston Consulting Group study (N=758)
- Participants incentivized
- Five hours
- Experimental design. Control group, two intervention groups; GPT only and GPT with prompt overview
- Pre-test on assessment tasks (realistic tasks)
- Random Assignment based on: Initial assessment task score, Big 5 personality traits, and demography
- Two experimental assignments. One inside frontier and one outside (18 realistic tasks)
- Assessed on quality (human graders and AI grader), task completion and speed



# Findings from trial

- **Inside the frontier tasks**

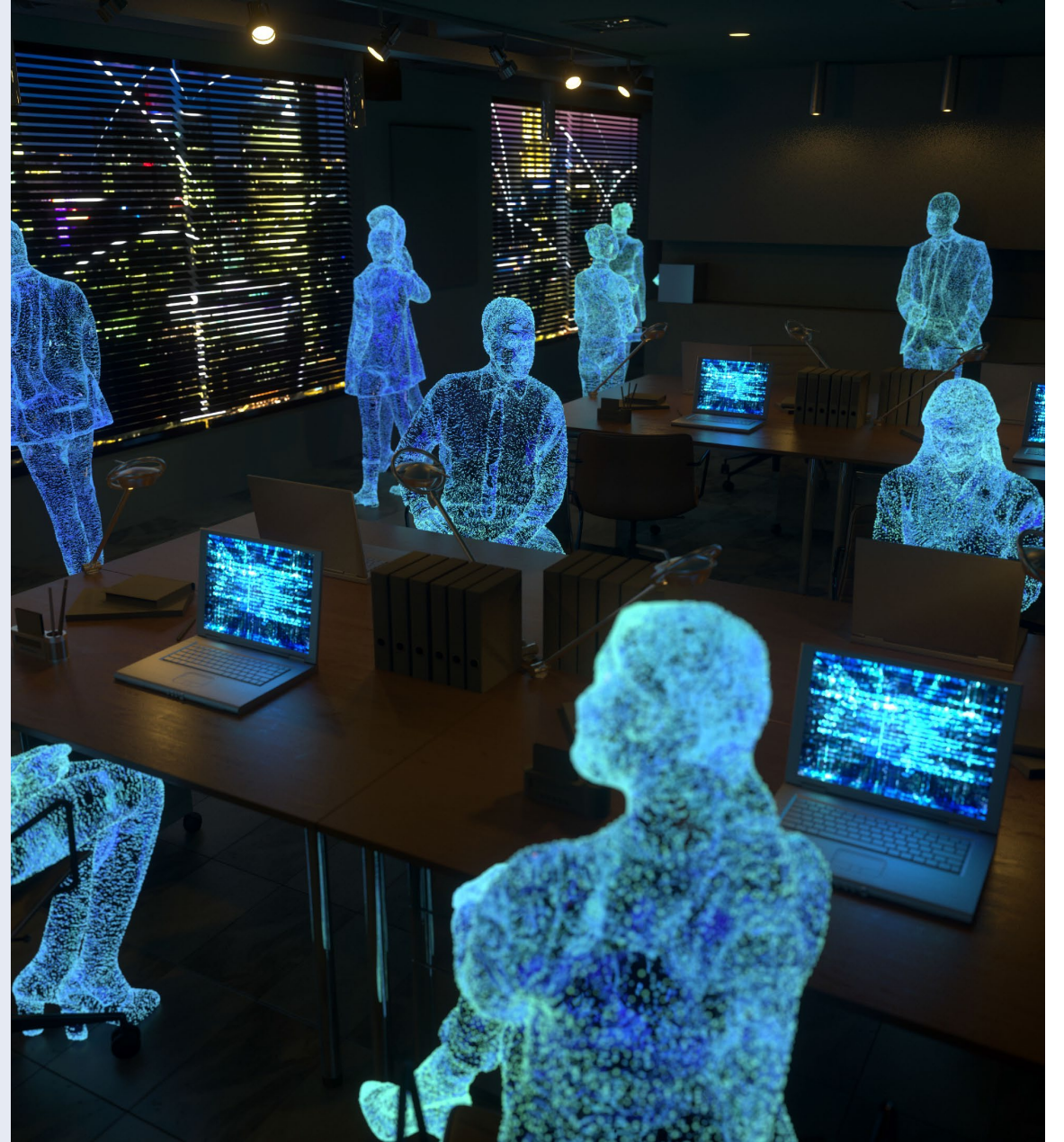
- Across 18 realistic business tasks, AI significantly increased performance and quality for every model specification
- **Effectiveness:** Performance (as rated by humans) increased by more than 40%
- **Efficiency:** Task completion increased by more than 12%
- **Expedience:** Increased speed by more than 25%

- **Outside the frontier tasks**

- **Effectiveness:**
  - Correctness: Control group was correct about 84.5%, while the AI conditions scored at 60% and 70%
  - Quality recommendation, the treatment GPT + Overview (25% increase over the control). GPT Only increased 18%
- **Efficiency.** Not reported
- **Expedience.** GPT + Overview increased speed 30% and GPT Only increased by 18%

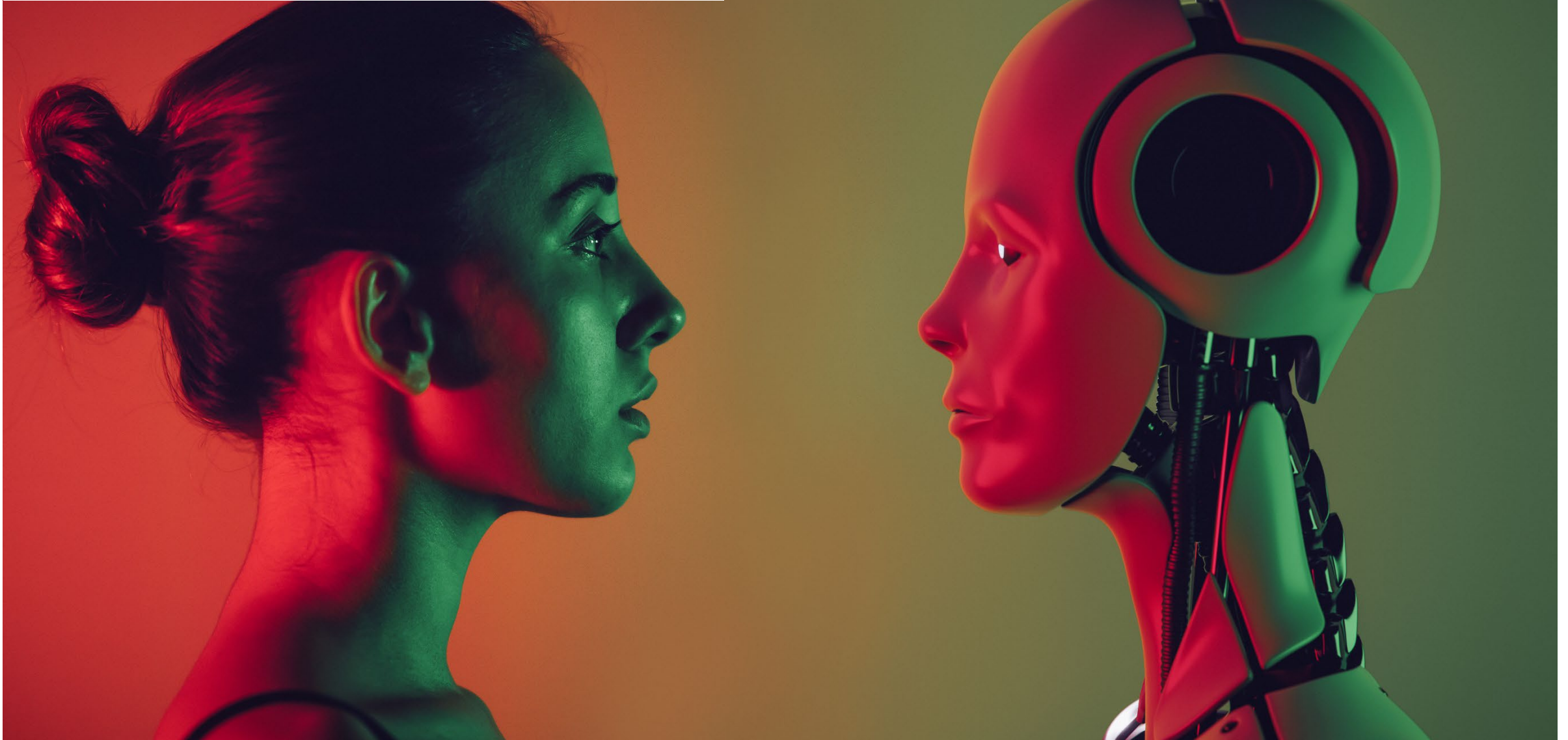
## In summary:

- **Inside the frontier tasks:** Benefitted bottom-half performers the most, although all users benefitted from AI.
- **Outside the frontier tasks:** It was only when tasks were outside the frontier that we saw performance decreased as a result of AI.
- “On those tasks, this study highlights the importance of validating and interrogating AI and of continuing to exert cognitive effort and experts’ judgment when working with AI.” (Dell Aqua et al, 2023:15)





## 4. What Does the Future Hold?





# What Evaluators Evaluate Will Change

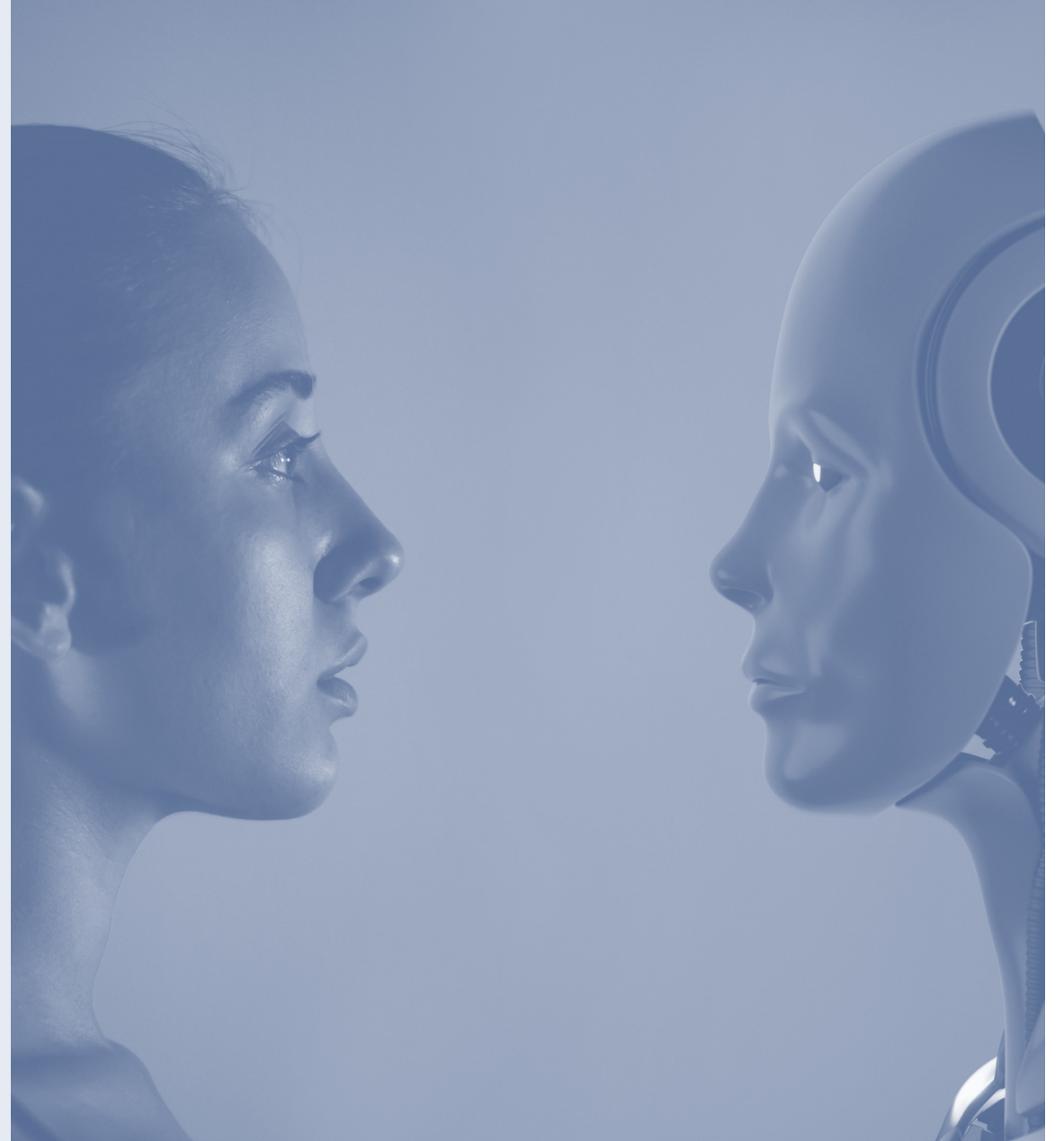




“

The evaluation community is well-positioned to provide leadership on the evaluation of and use of AI, including what criteria ought to be used.

Montrosse-Moorhead, 2023: 124





# Team Expertise Composition Will Change



## Subject Matter Expert

Provides in depth knowledge about the subject matter evaluated



## Evaluation Expert

Provides evaluation methodology and competencies



## Data Scientist

Provides technical expertise in data capture, storage and processing



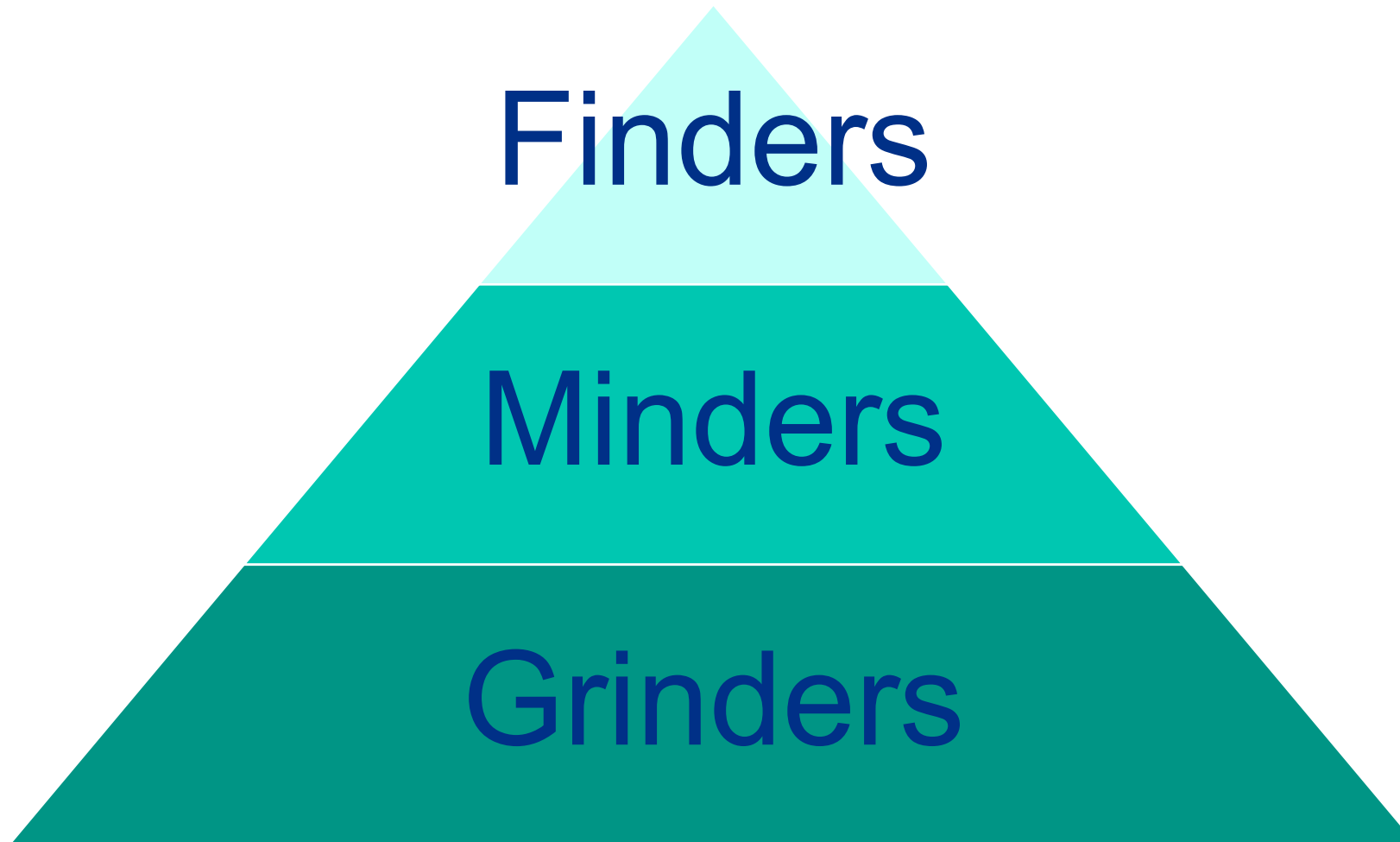
## AI solutions

Work alongside one or several technologies





# Team Experience Composition Will Change





# What Different Actors Need to Do



## **EVALUATORS**

Upskilling, AI literacy

## **EVALUATION PROVIDERS**

Grow talent, hire talent,  
collaborate to develop AI  
capabilities

## **VOPES**

Evaluator competencies,  
upskilling programs, advocacy  
policy-makers

## **EDUCATIONAL INSTITUTIONS**

Adapt education curriculum,  
develop competencies



## Some Resources

- Sabarre et al. (2023). Using AI to disrupt business as usual in small evaluation firms.  
<https://doi.org/10.1002/ev.20562>
- Nielsen, S.B. (2023). Disrupting evaluation? Emerging technologies and their implications for the evaluation industry. <https://doi.org/10.1002/ev.20558>
- Mason, S. (2023). Finding a safe zone in the highlands: Exploring evaluator competencies in the world of AI. <https://doi.org/10.1002/ev.20561>
- Dell'Acqua et al (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.  
<http://dx.doi.org/10.2139/ssrn.4573321>
- <https://merltech.org/>

# Cyber Society, Big Data, and Evaluation

Comparative Policy Evaluation  
Volume 24

Gustav Jakob Petersson and  
Jonathan D. Brouil, editors

With a Foreword by Caroline Hedder