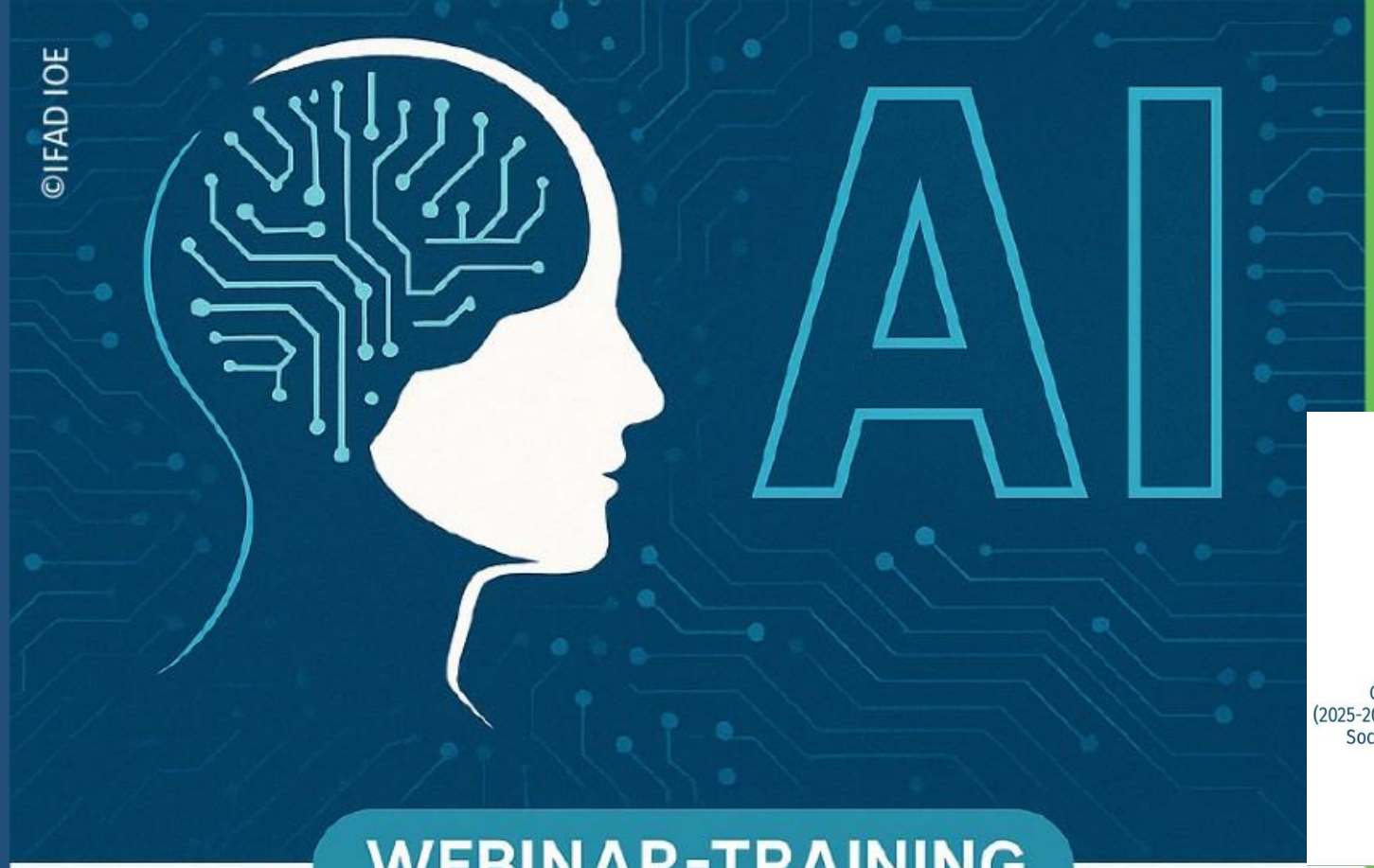
IOE | IFAD Investing in rural people

Independent Office of Evaluation




WEBINAR-TRAINING

# AI IN EVALUATION:
## Lessons from IFAD IOE and Practical Skills for Evaluators

SPEAKERS:

**Steven Jonckheere**
Senior Evaluation Officer at IFAD's Independent Office of Evaluation (IOE)

**Anoop Sharma**
Evaluation and AI Specialist at IFAD's Independent Office of Evaluation (IOE)

MODERATOR: **Lea Corsetti**
CGIAR IAES Consultant; EvalYouth Vice-Chair (2025-2027) and co-lead of the European Evaluation Society's Young and Emerging Evaluators group

HOST: **Innocent Chamisa**
EvalforEarth Coordinator

**Online,**
**22 September**
**2025**

# What can you expect in this session

**The 'Why':** Why AI? Why Now for Evaluation?

**The 'What':** IOE Use Cases (Successes & Failures)

**The 'How':** Building the Skills

**The 'So What':** Lessons Learned, Navigating Risks and Ethics

# The new way of working: Increasing use of GenAI

## Key Global Statistics (From the NBER Paper):

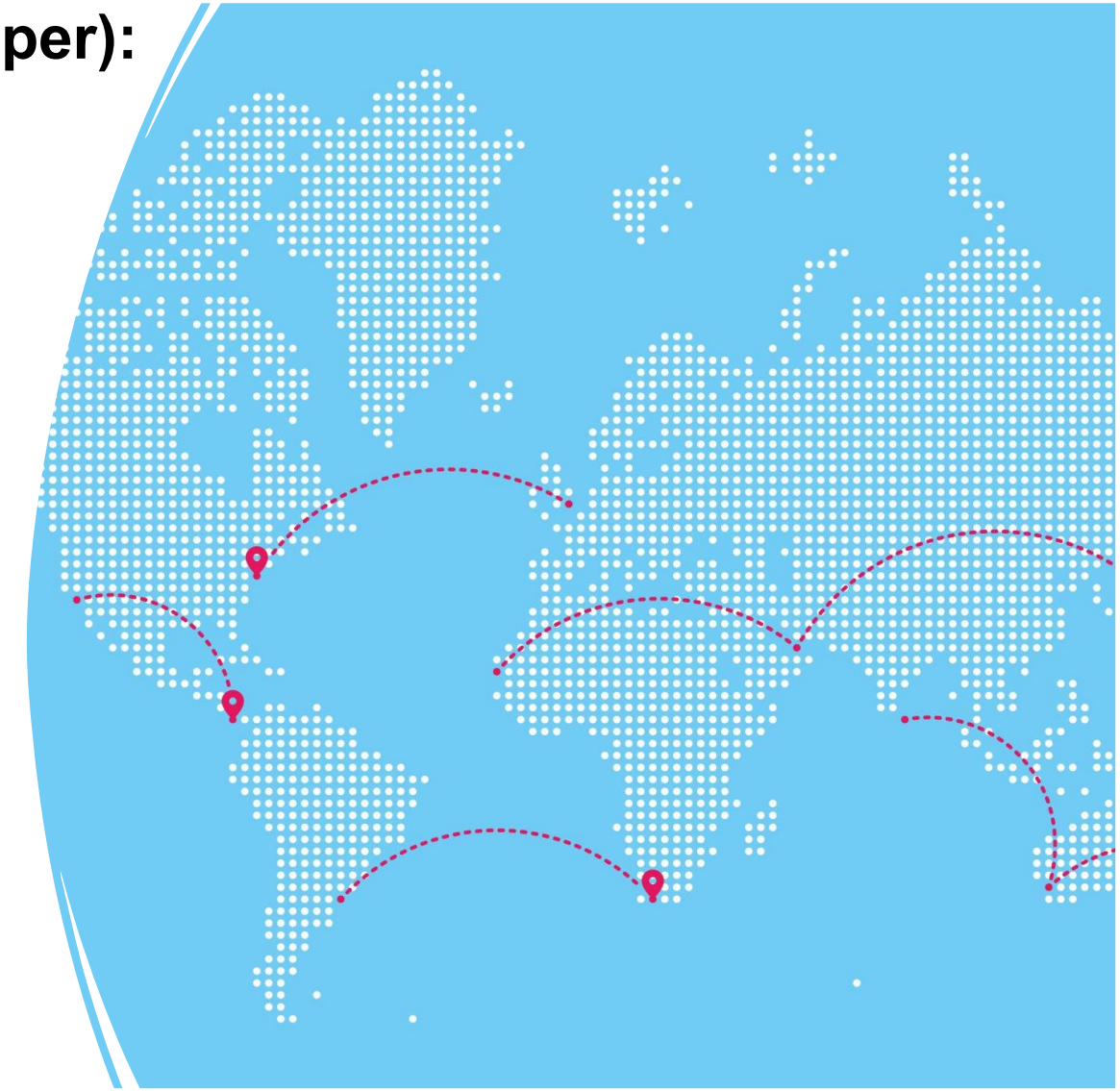
https://www.nber.org/system/files/working_papers/w34255

- **700 Million+ Users:** ~10% of the global adult population uses OpenAI models weekly.

**The Top Tasks**

- **#1 Work Use Case: Writing (40% of work messages)**
  - Emails, reports, communications, editing, summarizing.

- **#1 Overall Use Case: Practical Guidance & Decision Support**
  - Asking for advice, how-to information, and customized plans.



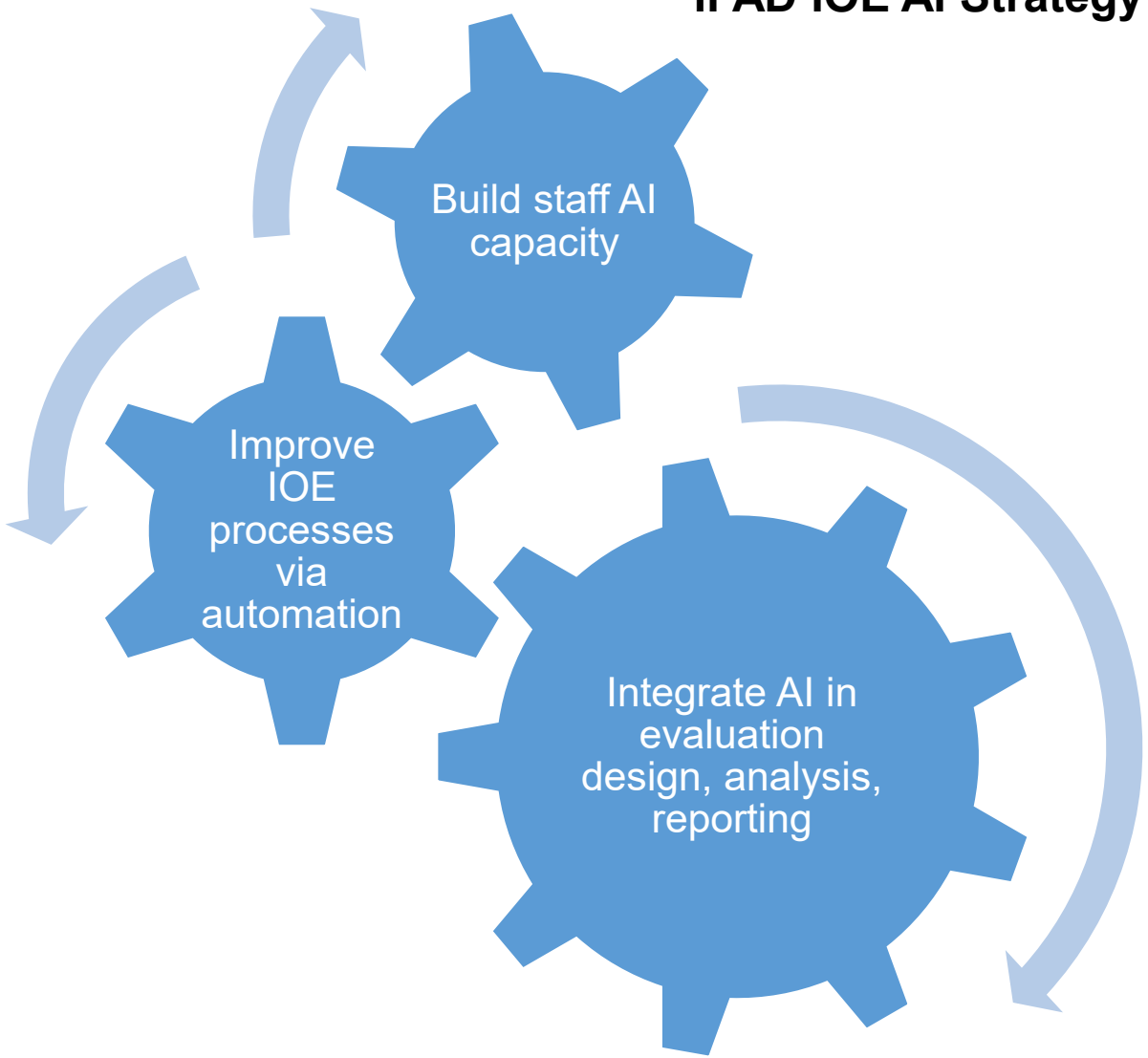
# The 'Why': Why AI? Why Now for Evaluation?

## The Evaluation Challenge and our commitment: More Data, Less Time, Deeper Questions

| | |
|---|---|
| An Expanding Evidence Universe | Growing volumes of unstructured data from interviews, reports, projects, synthesize. |
| Demand for Efficiency & Traceability | Demand to deliver robust insights faster and with clearer audit trails of the analytical process. |
| Support Evaluators in Repetitive Tasks | Automate repetitive tasks to free up human experts for high-level sense-making, judgment, and contextual understanding |
| Commitment to Adapt Technology and Innovation | Aligned with IFAD's commitment on innovation and technology as per Evaluation Policy |

# Our Response

## IFAD IOE AI Strategy 2024–2026




Partnerships with IFIs and UN

**Short-term**

- Expand pilot projects (in collaboration with ICT)
- Provide targeted training for staff

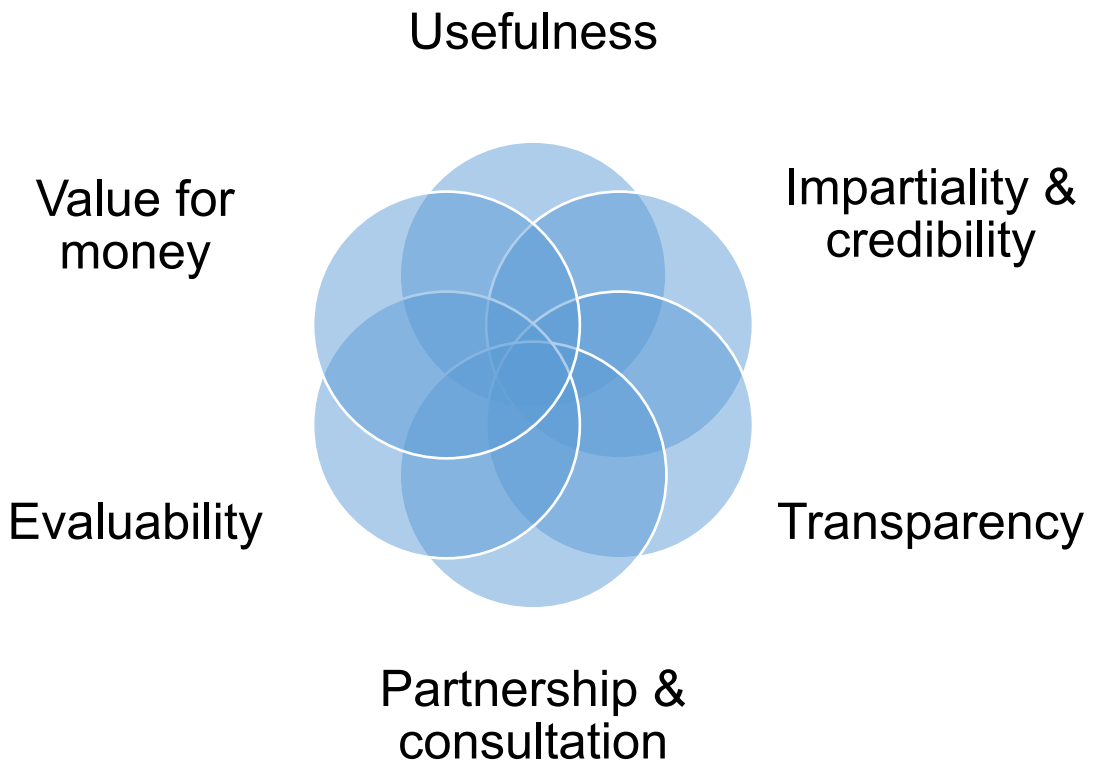
**Medium-term**

- Develop AI governance frameworks to ensure responsible use and accountability

**Long-term**

- Fully integrate AI into evaluation processes

# IOE Principles Anchoring AI




Ref- IFAD Evaluation Policy 2021

# The 'What': IOE Use Cases (Successes & Failures)

## Use Cases: Evaluation and Beyond

| | | | |
|---|---|---|---|
| Chatbot retrieval: key point extraction from HQ interview | Structured Interview analysis for country case studies: modular pipeline, synthesis across stakeholders | Document analysis: 95+ reports classified, operational parameters indexed | Triangulation: Sub-EQ matrices checked across 3+ evidence blocks |
| AI Powered dashboard for country operations analysis (Failed attempt for text analysis) | Chatbot for website – (ongoing) | AI powered newsletters – curated weekly evaluation/AI insights | Social media analysis – sentiment & reach tracking |
| Random desk allocation tool – fair workspace distribution | Smart formatting support – automated editing, consistency (failed attempt) | Report search engine – thematic indexing across corpus (Planned) | Automated meeting summaries – draft minutes from recordings |

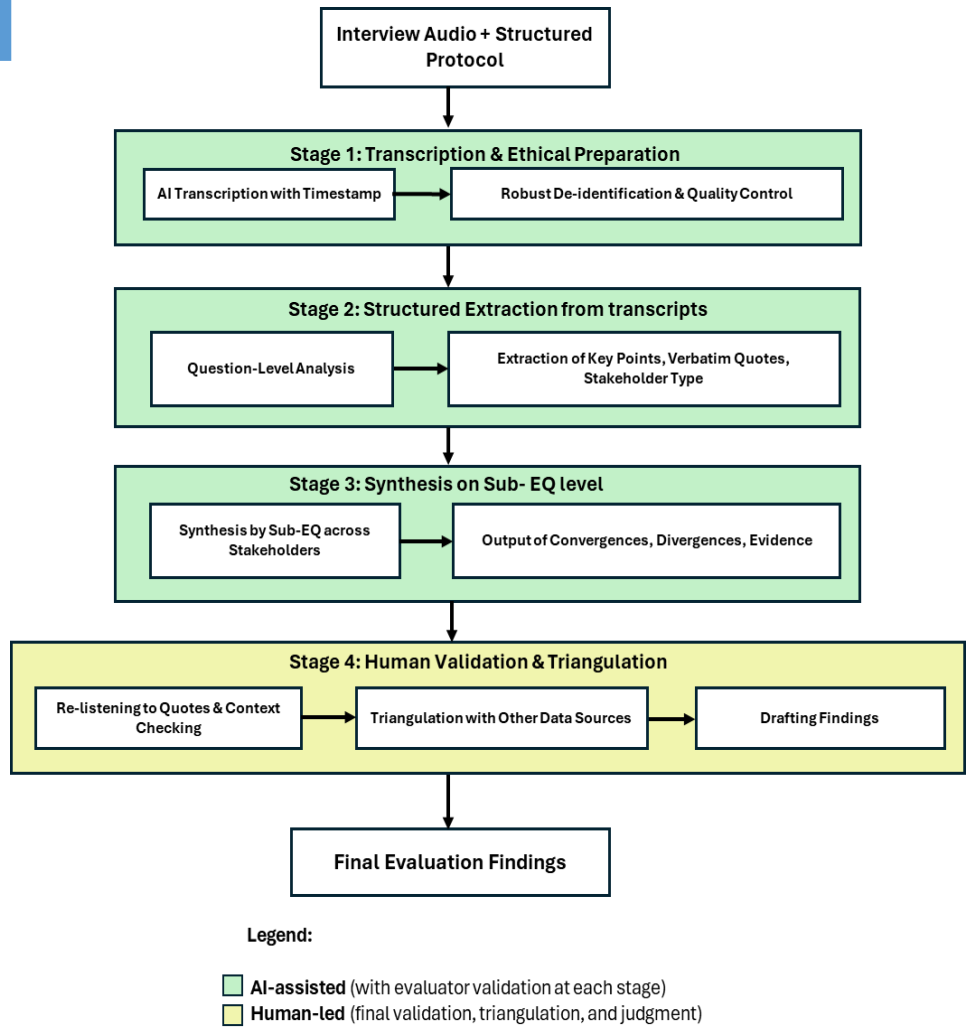
*Success was scalable; failure was educational.*

# Deepdive: AI for Structured Interview analysis

## Overview

- CLE country case studies had well defined structured protocols, and were conducted online
- 18 -25 interviews per case study, with gov't officials, beneficiaries, partners.
- ~200,000 words of transcripts to analyze manually.
- Manual coding was time consuming, and risk of inconsistency, and loss of nuance across a vast evidence base.

## Process




## Results

- Efficiency: Manual process: 2-3 weeks. AI-assisted: 2-3 days.
- Coverage: Every single transcript response was structured into an evidence database
- Traceability: A clear audit trail from a timestamped audio clip to a final finding.
- Evaluators focused on higher-order tasks

# Deepdive: Analysing non-lending activities

## Overview

- Extract information on non-lending activities from 62 COSOPs and 35 CSPEs
- 17 categories of Knowledge Management, Policy Engagement, Partnership Building.

## Process

*Step 1:* Create a database of NLA activities extracted from CSPEs and COSOPs

*Step 2:* Classify each paragraph into one of 17 categories and perform sentiment analysis on each paragraph.

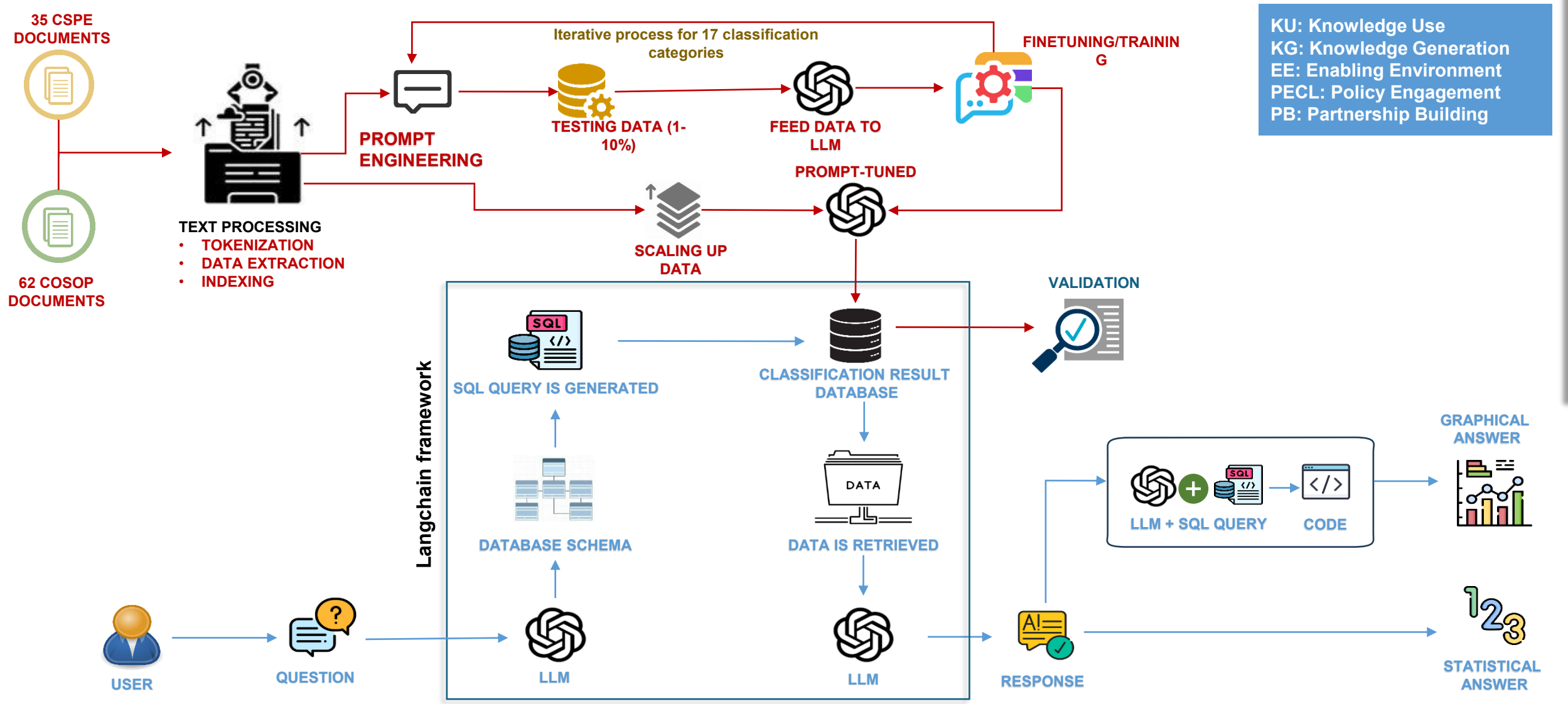
*Step 3:* Create a User Interface to generate necessary results and analysis

Validation

## Results

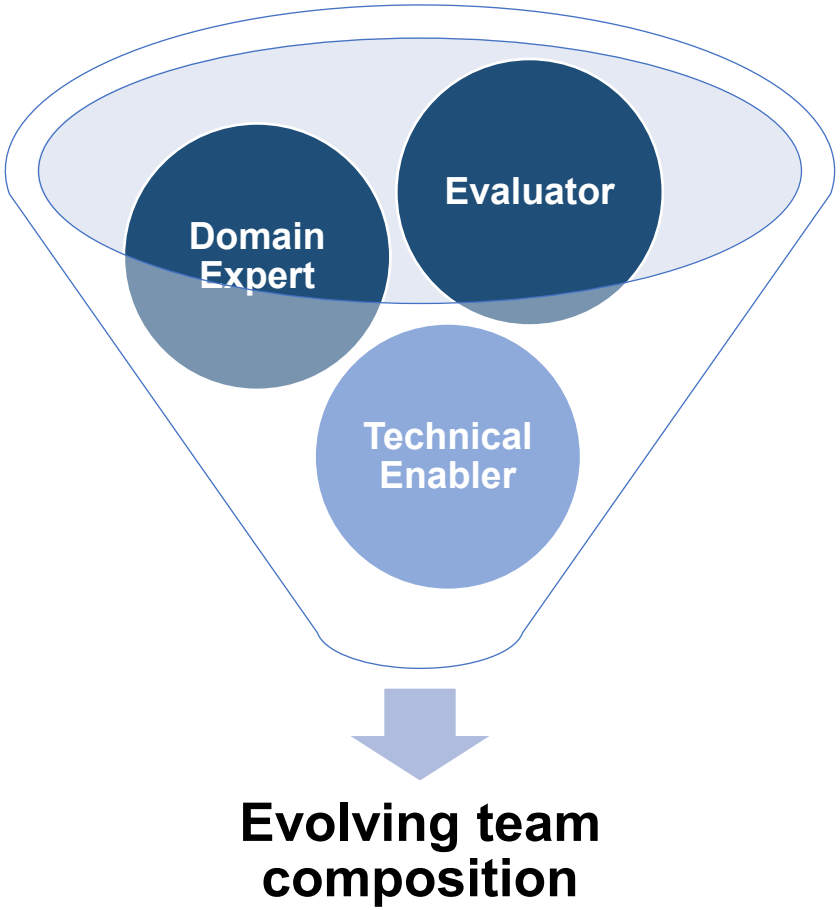
- Generate necessary graphs and tables of non-lending descriptive statistics
- 1500 Validated with F1 Score 0.82

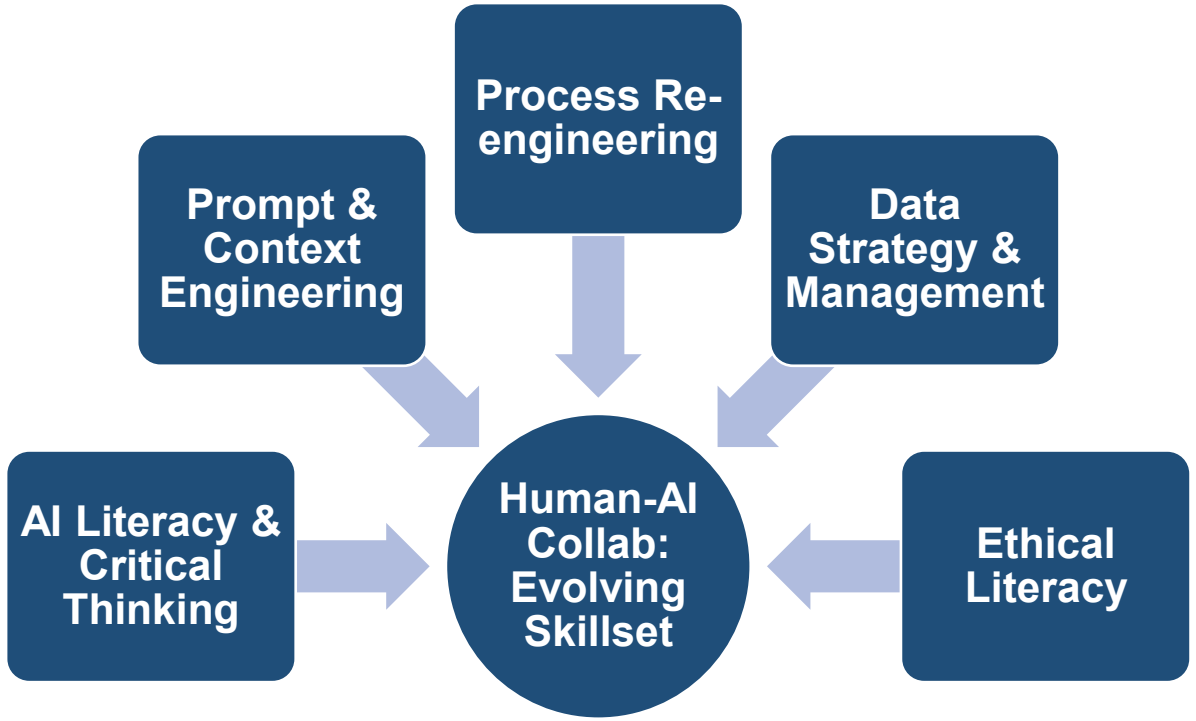
# Deepdive: Analysing non-lending activities



# The 'How': Building the Skills

## Evolving Team Composition and Evaluator's Skillset



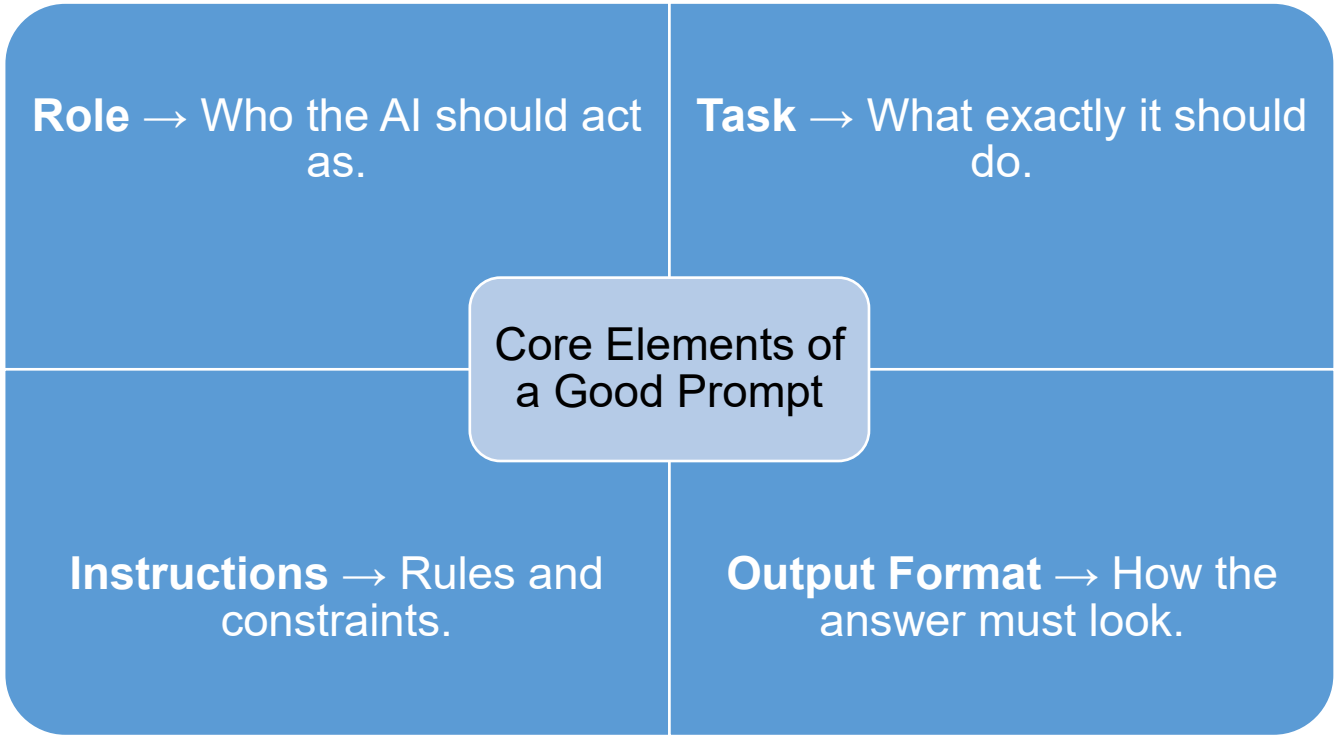



# Prompt/Context Engineering

## Prompt/Context Engineering

OpenAI (2022–2023). *OpenAI Cookbook & Best Practices for Prompting*.

**Prompt engineering**, *Prompt engineering is the art and science of designing effective instructions for AI models so they generate accurate, reliable, and useful outputs.*



# Good prompt vs bad prompt

📄 **Example Paragraph:** *"The project team worked closely with national stakeholders and local organizations to expand outreach to rural farmers. Regular meetings were held to coordinate inputs and share lessons, which also informed ongoing dialogue with ministries."*

❌ **Bad Prompt** *"Read this paragraph and tell me what kind of partnership it is."*

**Output:** *"This is a partnership between government, NGOs, and farmers, focused on capacity building and knowledge sharing."*

✅ **Good Prompt :** [Role]You are a partnership specialist.
[Task] Classify the paragraph into one or more of the following categories:
1.Financial Institutions/donors
2.Rome-Based Agencies & UN Country Teams
3.Civil Society Organizations
4.Farmer Organizations
5.Research Organizations
6.Private Sector
7.Policy Engagement Partnerships
8.Knowledge Partnerships
[Instructions] Respond only with the category name(s) and number(s). If none apply, respond with [no partnership detected].
Output format: "Category (number)"

**Output**
** Farmer Organizations (4)*
*•Policy Engagement Partnerships (7)*

# Prompt Engineering

## Types of Prompt Engineering

- **Zero/Few-Shot** = Examples for training.
- **Chain-of-Thought** = Improves reasoning.
- **Role + Context** = e.g Aligns with frameworks.
- **Output Constraints** = Makes answers usable.
- **Iterative Refinement** = Improves reliability.

# Gains

- **Efficiency**
- **Consistency**
- **Traceability**
- **Coverage**

# Challenges

- **Context-sensitive interpretation still requires human expertise**
- **Heavy validation & benchmarking workload**
- **Staff adoption challenges**
- **Nuance & ambiguity remain**

# Risks & Mitigation Measures

| Risk Category | Specific Risks | Mitigation Measures (Technical + Process) |
|---|---|---|
| **Bias** | - Data bias<br>- Algorithmic bias | - Diverse, balanced datasets<br>- Fairness checks & bias audits<br>- Human validation across diverse sources |
| **Quality & Reliability** | - Mis/disinformation –<br>- Hallucination<br>- Validation gaps | - Retrieval-Augmented Generation (RAG)<br>- Parameter tuning (temperature, top-p)<br>- Fine-tuning / grounding with evaluation corpora<br>- Human-in-the-loop validation<br>- Cross-checking & benchmarking (RACCA, F1) |
| **Other Risks** | - Dependency on ICT infrastructure<br>- Mistrust of technology<br>- Data privacy & evaluation ethics | - Local/on-prem LLM hosting & backup workflows<br>- Training & literacy for staff adoption 🛡<br>- Secure/local processing, anonymization<br>- Ethical & data-sharing guidelines |

# Lessons Learned

- AI is most powerful in evaluator-led, question-driven workflows
- Human validation is non-negotiable (cross-checking, RACCA, F1 Score triangulation)
- Documentation ensures transparency, replicability, and learning
- AI supports both evaluations & operations

AI outputs = inputs only, evaluators in charge

Responsible and Ethical Use (Secure, local processing & anonymization)

# Thank You!

We value your time and look forward to your questions.

## Q&A

Questions?
We're here to discuss.

# We'd love your feedback!

Go to www.menti.com
and enter the code

**8831 9213**

# What is key takeaway you are leaving with from todays' session?