EVAL

AVDELING FOR EVALUERING AV NORSK UTVIKLINGSSAMARBEID

EVAL

AVDELING FOR EVALUERING AV NORSK UTVIKLINGSSAMARBEID

Henrik Nordal

Evalueringsdirektør

EVAL AVDELING FOR EVALUERING AV NORSK UTVIKLINGSSAMARBEID

Evalueringsavdelinga er eit lite, men sterkt fagmiljø. Vi jobbar uavhengig, men tett på.

Avdelinga arbeider ut frå **Instruks for evalueringsoppgaven i norsk utviklingssamarbeid**

> *«Evalueringsarbeidet har som mål å fremme læring og et kunnskapsbasert norsk utviklingssamarbeid, samt holde aktørene i norsk utviklingspolitikk forvaltningsmessig ansvarlig. Evalueringene skal holde høy kvalitet og være troverdige, uavhengige og relevante.»*

Vi er her for å løyse globale utfordringar gjennom at fleire tek betre avgjersler oftare

Fremje kunnskapsbasert utviklingssamarbeid

Læring på tvers av aktørar og program: erfaringsspreiing og institusjonell læring (UD, KLD, Norad, Norec, Norfund, partnarar)

Truverdighet: Uavhengige evalueringar med høg kvalitet

Etikk: Evalueringar med integritet i tråd med tverrgåande hensyn, «do no harm»-prinsippet, forskningsmessig forsvarlighet og andre overordna etiske prinsipper

Bredde i deltakinga: Brei deltaking i alle fasar av evalueringsprosessen

Algorithmic Insights through AI-Powered Development Evaluation

EVAL

AI ToR Tool

20 November 2025

“Any sufficiently advanced technology is indistinguishable from magic.”

Arthur C. Clarke

Agenda

Why

How

Where


Why

Policy, Power & Progress

1. Norway wants 80% of the public sector to use AI by 2025.

2. EVAL’s edge: As a new unit, it can embed AI from the start.

3. AI boosts efficiency: faster decisions & quality processes.

Evaluation Process

1 DESIGN
- Planning
- Framework
- Knowledge needs
- Stakeholder engagement
- ToR

2 PROCURE
- Tendering docs
- Tendering process
- Supplier selection
- Contract signing
- Startup meeting

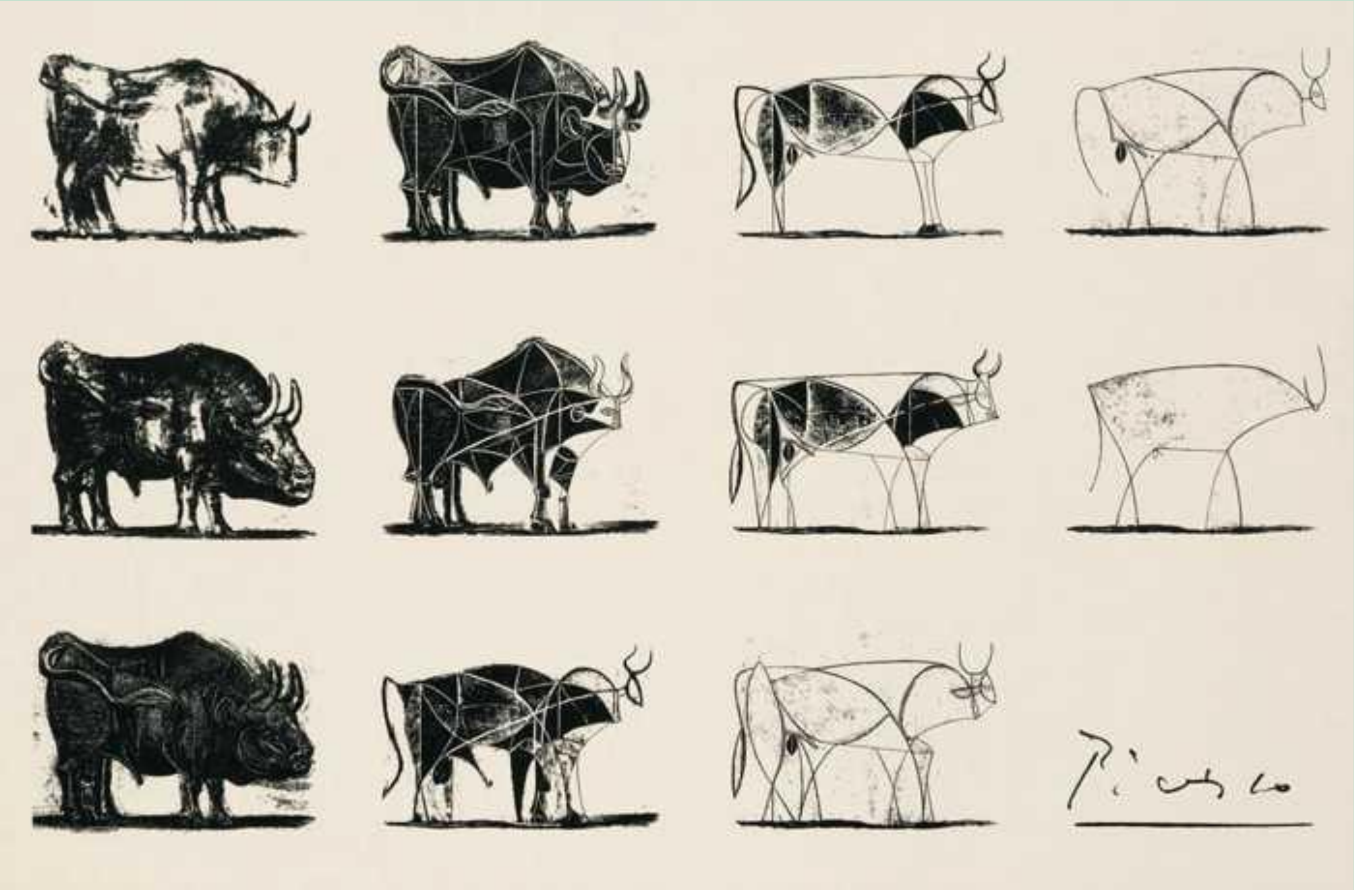
3 CONDUCT
- Inception report
- Data collection
- Data analysis
- Report writing
- Facilitation
- Quality assurance
- Draft report

4 FINALIZE & DISSEMINATE
- Report completion
- Recommendation discussions
- Formal submission
- Dissemination
- Public launch
- Knowledge sharing & broader engagement

AI TOOL CRITERIA
- Feasibility
- Centrality
- Impact potential
- Low risk

How

Easy Yet Complex



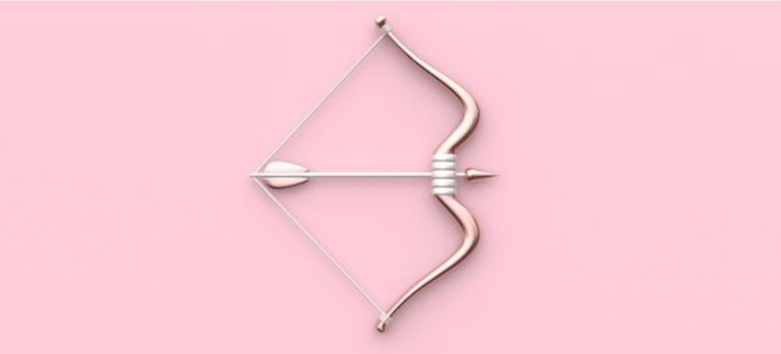
"We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.

Not known, because not looked for
But heard, half-heard, in the stillness
Between two waves of the sea.
Quick now, here, now, always--
A condition of complete simplicity
(Costing not less than everything)
And all shall be well and
All manner of thing shall be well
When the tongues of flames are in-folded
Into the crowned knot of fire
And the fire and the rose are one."

T. S. Eliot, "The Little Gidding"

Empathy for Human Needs

Task division: evaluators knew *what*, developer knew *how*



Clear goals & buy-in, maintaining human agency from ToR template to final tool

Quality testing: prioritized must-haves

User-friendly for non-AI experts & matched AI capabilities to existing expertise

Developer - Eval synergy

Strategic thinking: asked all the right ?s

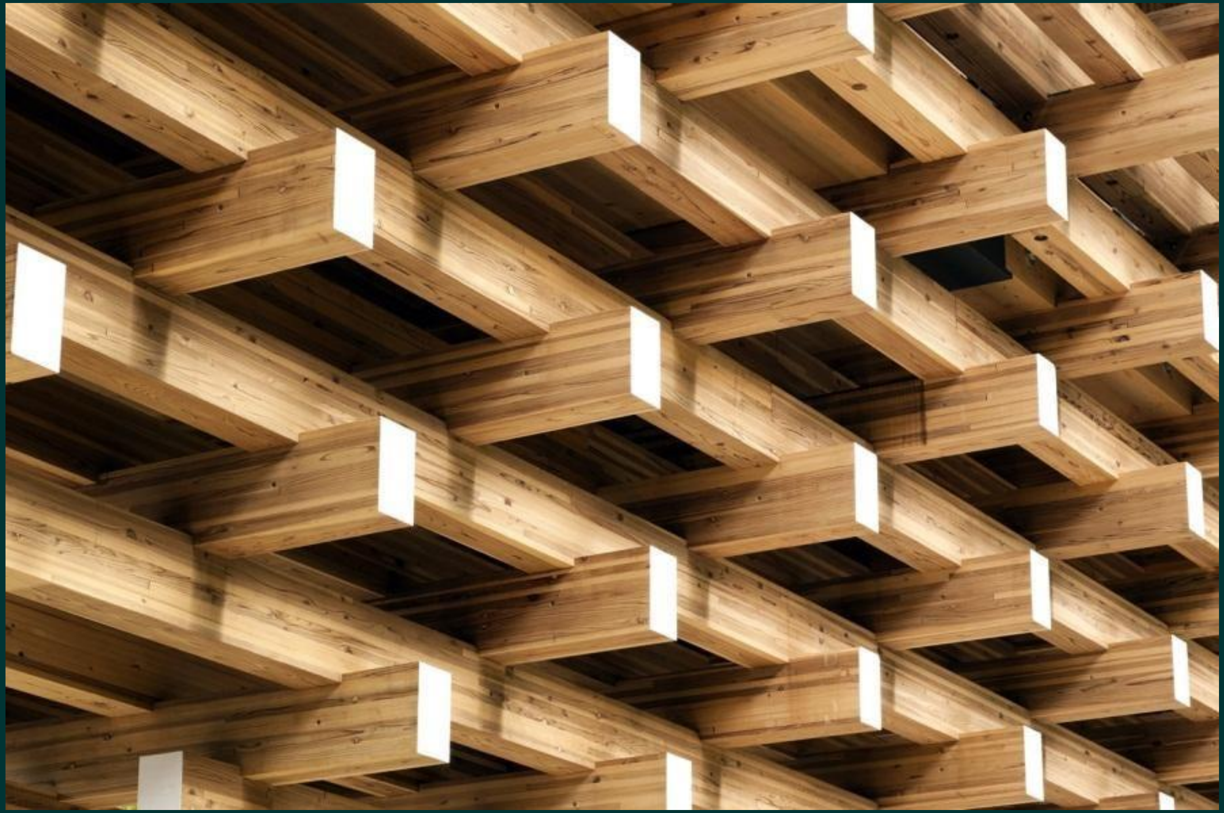
We developed a **fully functional application with a tailored frontend** in only 2 weeks with 1 developer.

The Result of Joint Effort

Clear problem definition &
rapid feedback loops

Tech stack choice allowed building fast:

- Langsmith for tracing
- LangGraph Cloud for development of agent backend
- Vercel for deployment of frontend
- Perplexity API for internet research tasks
- Azure OpenAI o4-mini as state-of-the-art reasoning model for increased intelligence

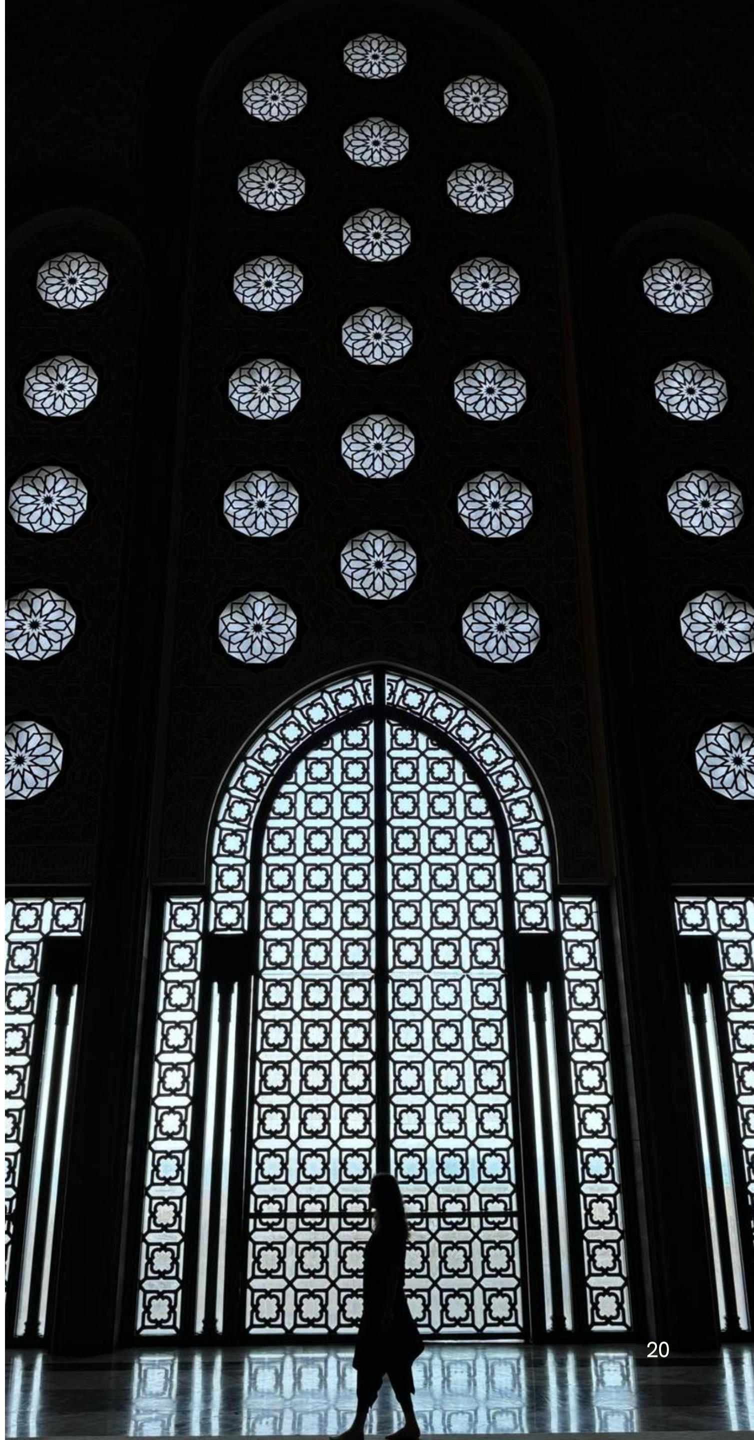


Testing Experience

- Ukraine ToR test: The AI-generated ToRs closely resembled the human-created version, also thanks to the guiding ToR database.
- Sonar Deep Research typically produced stronger outputs than Sonar Pro (context, concrete examples)
    - Some outputs were not entirely accurate, possibly due to the model generating content where data was lacking
- Increased quality when prompts highlighted aspects like geographies rather than being broad. This had a domino effect on the entire ToR, impacting all its subsequent parts. → improved instructions
- The more you give it, the more it will give you

The solution covered the whole process,

not just a step.



We built a product.

Not a chatbot.

- Easier, more time-efficient
  - Less context required
  - Less switching/copying between applications
  - Close-ended: easier to know what is possible and what's not

- Users can be guided to follow an intended approach
  - Scope the solution easier
  - Fewer edge cases

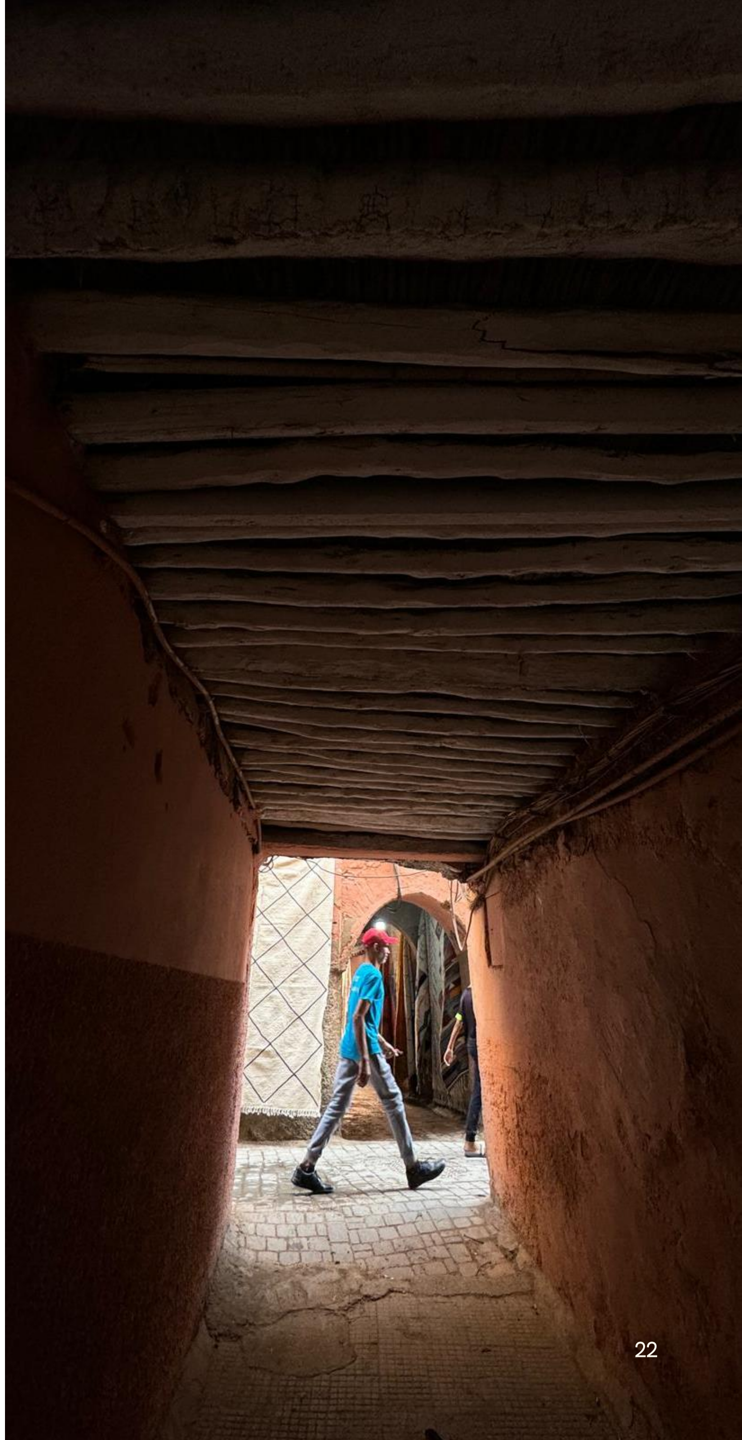
Moving Forward

Measuring efficiency gains

- Parallel testing of both models, Sonar Deep Research and Sonar Pro, comparing them against platforms like Perplexity, Claude, and ChatGPT.

MVP, yet complex

- Strong potential.
- Usability problems came from shortcuts, not missing solutions.
- Reinforcement Learning through Human Feedback (RLHF).



“When you build a system like ChatGPT, there’s an important step that comes after training it on text, called Reinforcement Learning from Human Feedback. Imagine thousands of people reading AI responses and rating them like restaurant reviews. What gets high ratings? Responses that sound confident, helpful, and agreeable. Not necessarily truthful, just… pleasant. And this is what reinforces and shapes the model’s future responses.

So ChatGPT is basically trained to tell you what you want to hear.

Understanding this training process has shaped the way I use AI tools. **I never ask “how should we approach this product launch?” Instead, I ask “what assumptions could cause this launch to fail spectacularly?”** The difference is huge - instead of getting validation, I get genuine analysis that challenges my assumptions and reveals blind spots.

Same system, completely different outcomes.”

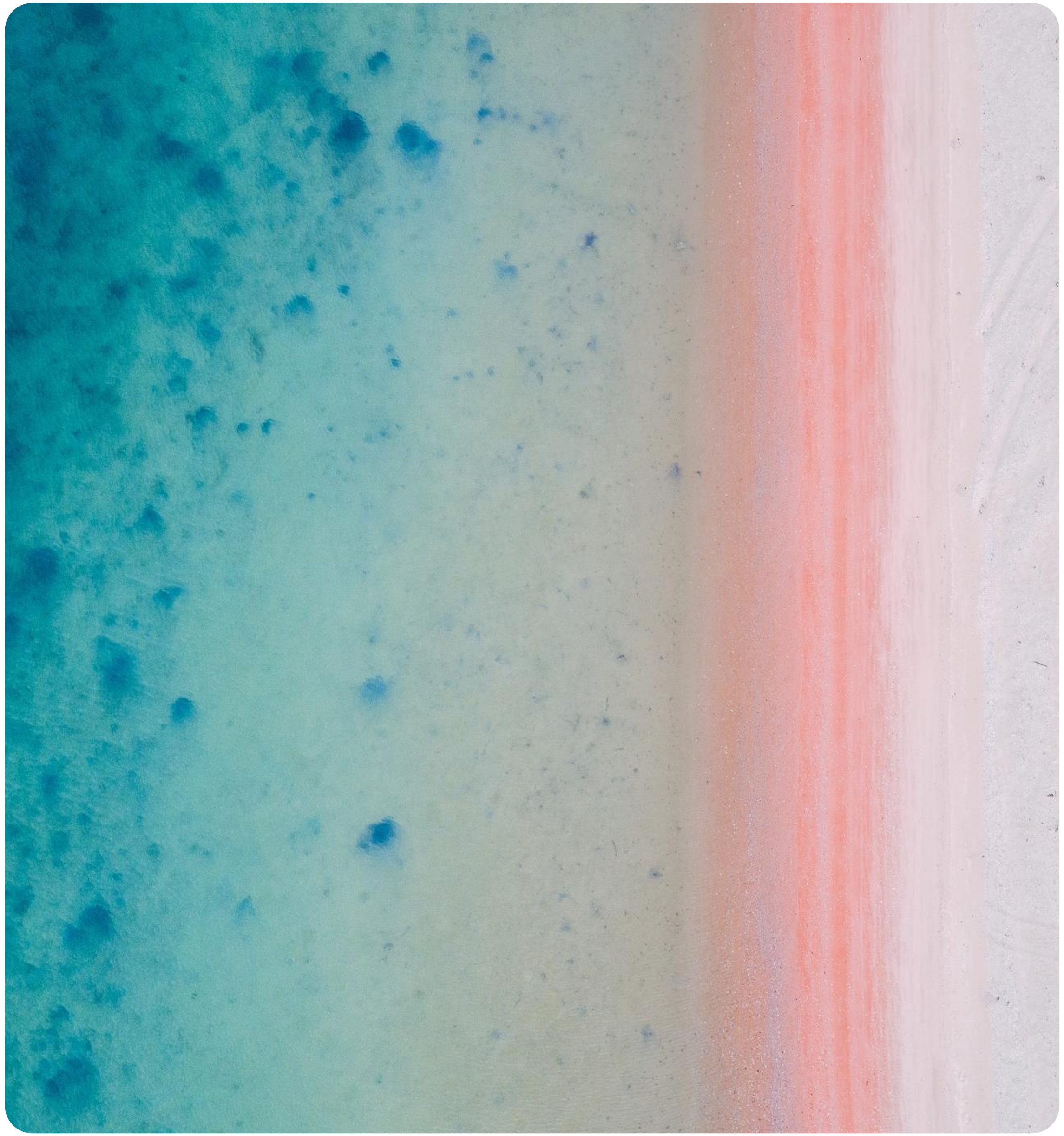
Tey Bannerman, AI Strategy, Product & Design Leader at McKinsey London

A value-driven, human-centered approach must lie at the heart of responsible AI use, enhanced by guidelines, training and risk-based planning.

It's important not just to ask what AI can do, but what it should do.

AI *in Flow*

What AI Can Do



Advantages

Efficiency and Quality

▪ Quantitative optimization and predictive power. Cuts bias, boosts accuracy, and speeds evaluations.
▪ Creates more time for strategic and creative work.

Depth of Knowledge and Critical Thinking

▪ Pushes evaluators to consider new perspectives.
▪ Reveals alternative insights and methods.
▪ Redirects focus to under-researched areas for deep-dives and exploration.

Learning, Coaching & Design

▪ Guides evaluators on methods, ToRs, best practices.
▪ Strengthens design: flags risks, gaps, improvements.

Horizon Scanning & Future Forecasting

▪ Detects emerging trends, risks, and opportunities.
▪ Informs evaluation planning with future-focused insights.

What AI Can't Do

Challenges

Unplanned, Ad Hoc Implementation
- Without a structured approach, AI use is ineffective.

Uncritical Use & Bias Risk
- AI reflects training data biases and lacks contextual awareness.
- Outputs require critical validation to avoid poor research.
- Over-automation may limit learning and judgment.
- AI lacks intuitive grasp of what it knows and doesn’t.

Privacy, Security & Ethics
- Sensitive (open) evaluation data may face data risks.
- Potential environmental impacts from AI use.
- Equity issues: strong access for governments, weaker for partners and civil society.

Quality Gaps in the ToR database
- Consider supplementing with additional ToRs.
- Ensure effectiveness even with limited data availability.

Risks

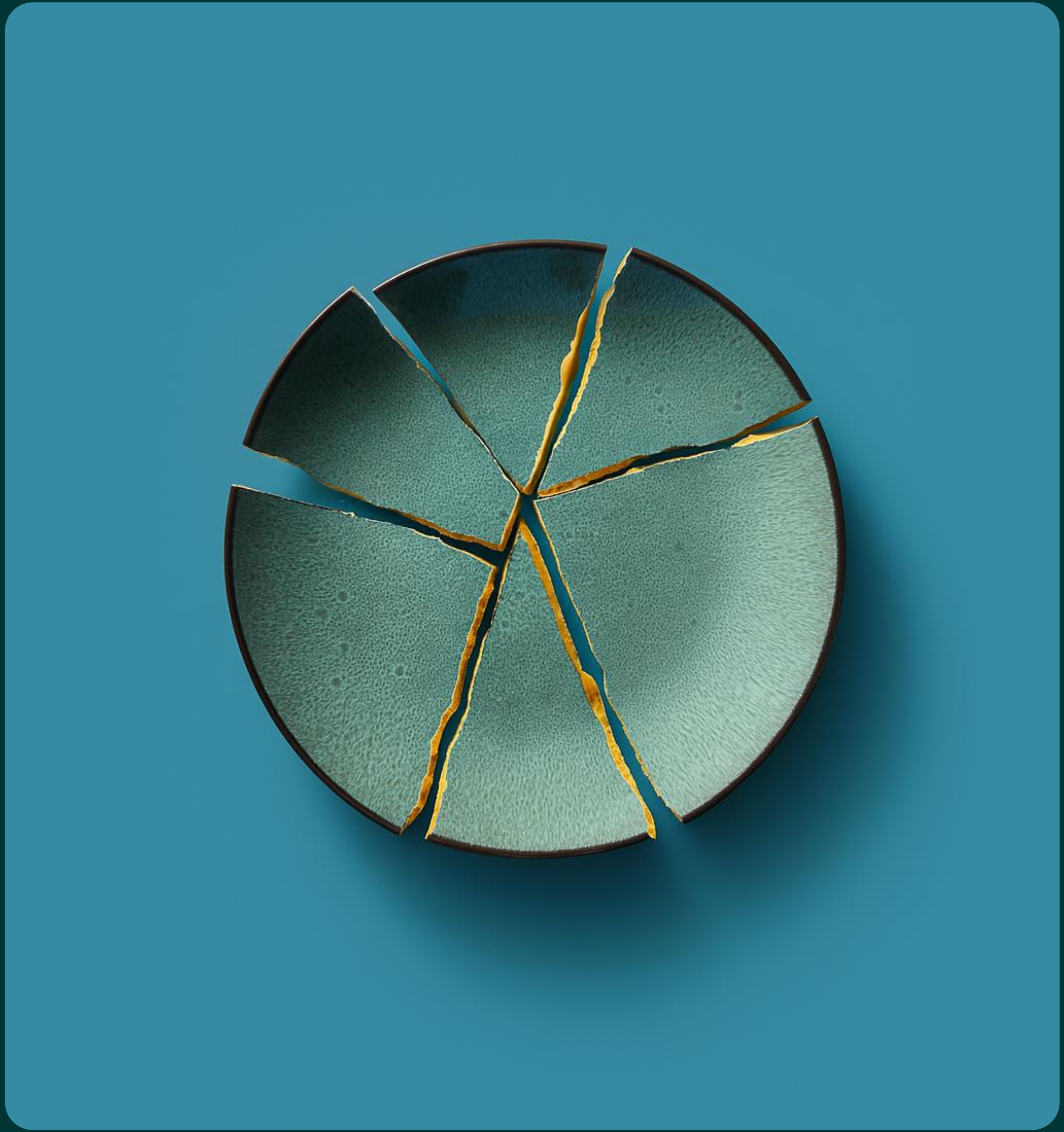
Human "autopilot"
- AI is ineffective without user expertise and critical thinking.
- Process: stakeholder engagement, drafting, feedback, finalization.
- Misalignment with EVAL's qualitative priorities.

Over-reliance on AI outputs
It is an iterative process requiring regular feedback and updates. Once roles are clear, it can be cost-effective to wait for multiple feedback rounds before updating the knowledge base.

Saves time but not effort
- All references must be double-checked, as errors are common.

Counter-intuitive to use
- Best results come from using LLM-based AI in smaller, iterative doses with interdisciplinary review.

Limits with sensitive data
- Certain data can't be shared with AI due to privacy and security risks.
- In areas like conflict-related decision-making, classified data gaps mean AI insights remain partial and require human intuition to fill.
- Weak regulations and human errors can expose sensitive data, sometimes resulting in serious legal or reputational harm.



#Human

The Human Advantage

1. **Creativity**: create, conceptualize or plan strategically, thinking across domains or applying common sense.

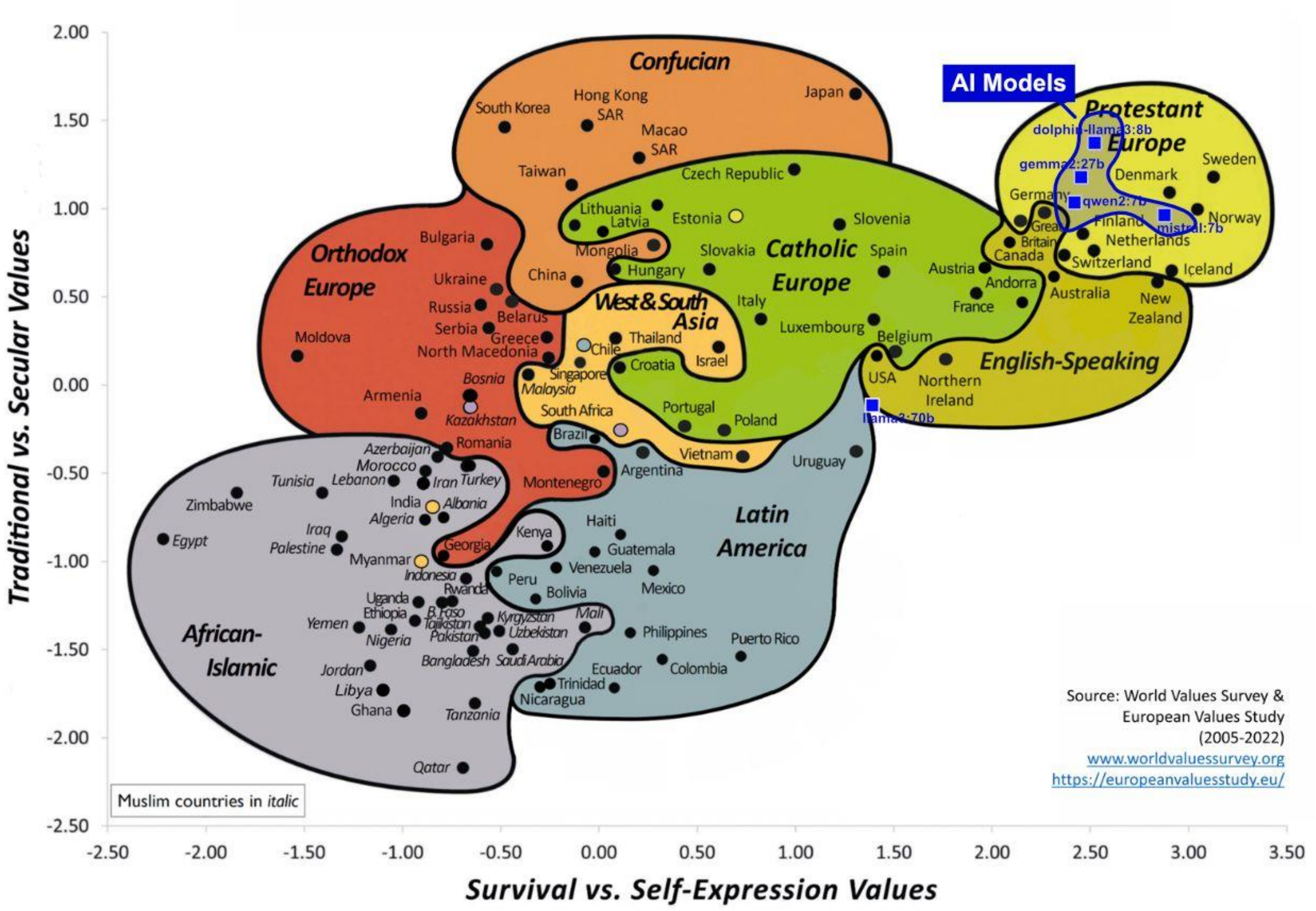
2. Empathy & compassion: make people feel understood or cared for → AI has no “human touch”.

3. Adaptive sensemaking: AI can’t iteratively navigate ambiguity and uncertainty to reach understanding.

Disadvantages stemming from the overuse and lack of nuanced use of AI can enhance global flat knowledge: while the population at large can easily understand and access information, complex, layered knowledge derived from deeper cross-sectoral, cultural and contextual awareness can be diminished when AI is not a springboard for further refining data, but a final aim.

AI Models and their cultural alignment

This graph visualises the cultural alignment of various LLMs by plotting them on the Inglehart-Welzel Cultural Map - comparing their cultural values with those of different countries & cultural regions.



"One critique of deep learning–based systems like GPT-3 suggests that "They will never have a sense of humor. They will never be able to appreciate art, or beauty, or love. They will never feel lonely. They will never have empathy for other people, for animals, or the environment. They will never enjoy music or fall in love, or cry at the drop of a hat."

ChatGPT-3 offering a critical take on itself

"Does the technology's ability to make such an accurate critique contradict the critique itself?"

Kai-Fu Lee, AI2041

Where

Walking backwards into the future

Practical & Rapid Solutions

Develop small, scalable, adaptable AI prototypes.
Weigh short, medium, long-term goals.

Energy Efficiency

Choose the right approach and model size for the task, such as small language models (SLMs) and lightweight solutions.

Learning from AI

- AI-assisted learning, coaching & evaluation design.
- Future-ready evaluations.

1 What are you optimising for? →

2 What's at stake? ↓

What's at stake?	SPEED/VOLUME *Maximum efficiency and scale*	QUALITY/ACCURACY *Ensuring precision & correctness*	COMPLIANCE *Meeting regulatory requirements*	INNOVATION *Fostering breakthrough thinking*
Irreversible consequences	**Circuit-breaker protocols** Human authority to halt system immediately. Multiple verification steps. Clear escalation chains.	**Expert override systems** System explains reasoning, human validates. Domain experts have full authority to counteract AI outputs.	**Mandatory human approval** AI recommends, human decides. No automation without explicit human authorisation.	**Collaborative validation** AI and human analyse independently, flag discrepancies, resolve together.
High-impact failures	**Monitored automation** AI acts autonomously with real-time human monitoring. Alerts on anomalies.	**Regular expert review** Expert periodically audits AI decisions. Identifies patterns, adjusts parameters.	**Approval workflows** AI processes routine cases, escalates edge cases to human reviewers.	**Collaborative ideation** AI generates options, human provides creative direction and strategic context.
Recoverable setbacks	**Batch processing** AI processes large volumes, human reviews samples and overall patterns.	**Spot checking** Random sampling of AI decisions for quality assurance and calibration.	**Rule-based guardrails** AI operates within predefined parameters. Human oversight for exceptions.	**Creative collaboration** AI generates initial concepts, human refines and adds strategic insight.
Low-stakes outcomes	**Full automation** AI operates independently with minimal human intervention. Performance monitoring only.	**Feedback learning** AI learns from user interactions and feedback to improve over time.	**Template automation** System handles routine, standardised processes with minimal variation.	**Generative assistance** AI provides creative suggestions and variations for human consideration.

- **Active control:** Human authority at all times
- **Human augmentation:** Human-led with AI support
- **Guided automation:** AI with human oversight
- **AI autonomy:** Minimal human involvement

REAL-WORLD EXAMPLES

E-commerce: email marketing

Goal: Send personalised emails to millions (optimising for speed/volume *and* quality).

Risks: Poor copy, spam risk (recoverable setbacks).

→ Batch processing + Spot checking:
AI generates emails, marketing team reviews random samples pre-sending, and monitors overall engagement metrics, intervening when unusual patterns seen.

Recruitment: applicant screening

Goal: Support recruitment/HR team by processing hundreds of applications efficiently (volume, quality, compliance)

Risks: Systemic bias, missing great candidates or hiring poor fits (high-impact failures).

→ Monitored automation + regular expert review + approval workflows
AI screens applications for requirements and fit indicators, flagging top candidates and clear rejections. Recruiters review candidates and can override any AI decision. Hiring managers get AI analysis alongside resumes for final interviews.

What type of intentional human-AI oversight design works best?

How can AI be used in summarizing/collecting and using already existing knowledge?

How to drive the learning from the AI process further?

What should we teach AI to unlearn?

How not to get caught in forgetting the power of one's own intuition?

How can AI deepen human connection and creativity rather than create distance between us?

"You see things:
and you say, 'Why?'
But I dream things that never were;
and I say 'Why not?'"

George Bernard Shaw,
Back to Metuselah

THANK YOU.

norec.no/evaluation

EVAL

EVAL

AVDELING FOR EVALUERING AV NORSK UTVIKLINGSSAMARBEID

Hvordan styrke strategisk evaluering og kunnskapsanvendelse i norsk utviklingssamarbeid?

EVAL

Jens Frølich Holte

Leder politisk avdeling, Høyres stortingsgruppe. Tidligere statssekretær i UD, KLD og Samferdselsdepartementet.

Ottar Mæstad

Professor, spesialrådgiver, direktør for development learning lab (DLL)

Henrik Nordal

Direktør for Avdeling for Evaluering av Norsk Utviklingssamarbeid (Eval), Norec

Moderert av **Anne-Thora Vardøy-Mutale,** Eval