Agentic AI for Educationists: Leading Responsible AI Transformation in Education

Module 5: Measuring Impact and Value

Jasmin Jahić

07.09 - 08.09.2026

jj542@cam.ac.uk

About me

- Born in Bosnia and Herzegovina
- Master degree: Ecole Centrale de Nantes, France
- PhD: Technical University Kaiserslautern, Germany
- Director of studies in Computer Science at Queens' college, University of Cambridge
- Principal lecturer and Researcher at the Department of Computer Science, University of Cambridge
- Lecturer for Cambridge Advance Online and Professional and Continuing Education

Overview

- LO5: Analyse the impact of agentic AI initiatives using qualitative and quantitative evidence, including educational, operational and organisational metrics.
- LO6: Develop evidence-informed recommendations and policy communications for institutional and public sector stakeholders.

Overview: Topics

T1: Expected impact: AI application

- Organisational and individual goals
- Affected stakeholders

T2: Observe and predicted impact

- Measure the impact
- Side-effects
- Educational, operational and organisational metrics

T3: Measurement results

- Presentation and abstraction
- Conclusions
- Threats to validity

T4: How to use the results

- Evidence-informed recommendations
- Policy communications

Schedule

- 07.09.2026: 9:00 – 12:30
  - T1: Expected impact: AI application
  - T2: Observe and predicted impact

- 08.09.2026: 8:30 – 11:30
  - T3: Measurement results
  - T4: How to use the results

Setup

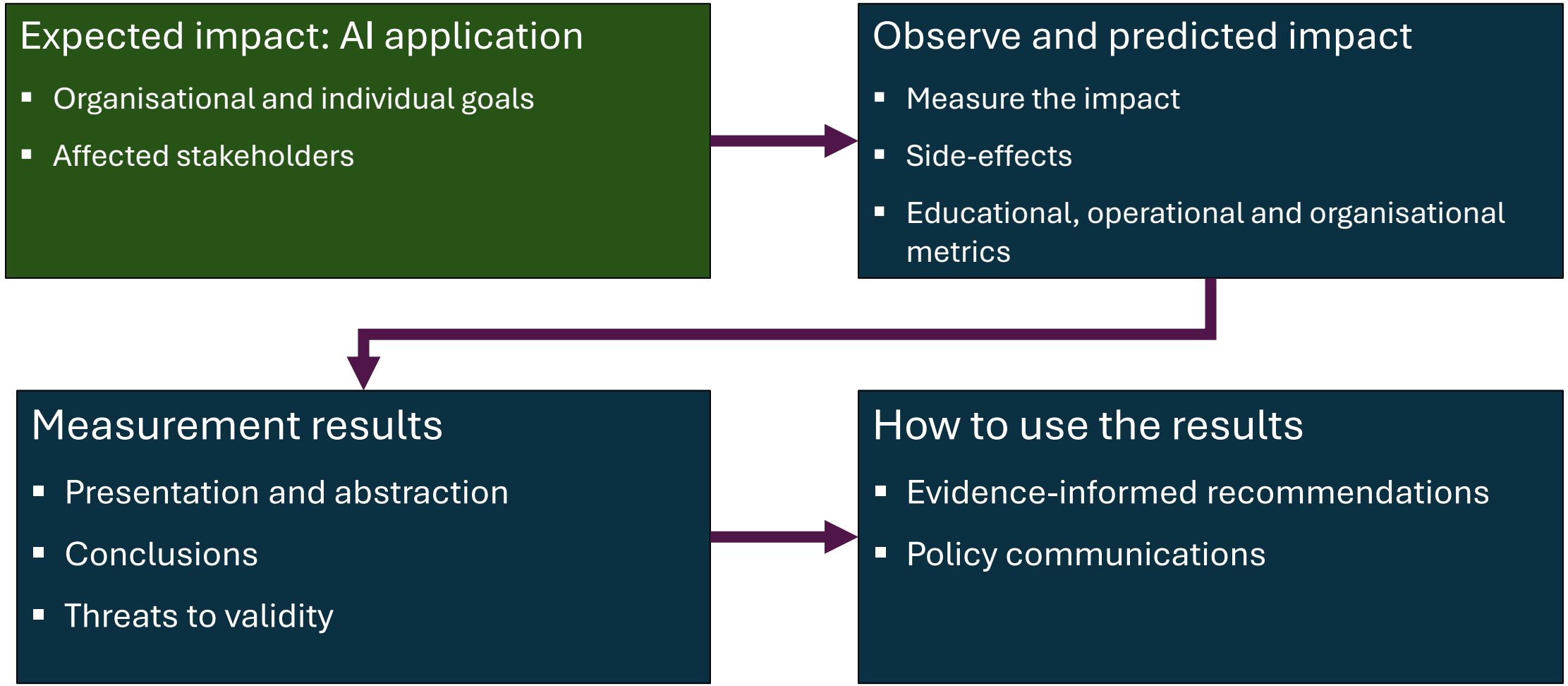
- Groups
- Working example: 1 per group
- Interactive exercises
- Take notes
- Continuous improvement of the example
- End of a topic: Prepare “presentation” (sketch) 10 – 15 minutes
- Presentation at the end of each topic (5 minutes for presentation, 2 minutes for questions and discussion)
- Immediate feedback, revisions offline

Setup

Create groups

Write notes on a board/paper sheet

Overview


Expected impact: AI application
Organisational and individual goals
Affected stakeholders
Observe and predicted impact
Measure the impact
Side-effects
Educational, operational and organisational metrics
Measurement results
Presentation and abstraction
Conclusions
Threats to validity
How to use the results
Evidence-informed recommendations
Policy communications

Previous Modules

- Go through previous modules
- Write on the board:
  - Identify goals, risks, necessary change, long-term effects
  - Identify parameters that characterize these
  - Identify quantification parameters
  - Set quantifications

Why Measuring Impact and Value matters?

Manage expectations

- Narrow the scope
- Refine goals
- Recognize pitfalls

Track progress

- Project management
  - Estimate progress
  - Manage effort and budget
  - Prioritize
- Research
  - Approve or disapprove a potential solution or a hypothesis
  - Argue trade-offs

Module 1: Understanding Institutional Transformation

- The Changing Landscape of Education in the Age of Agentic AI: global trends reshaping education, why agentic AI represents a strategic challenge. How governments and educational institutions are responding to these developments. Opportunities, risks and long-term implications for learners, educators and institutions.
  - Hype and fear
  - Ground conclusions in rational analysis and data
  - Opportunities -> Beneficial outcomes we can measure
  - Risks -> Drawbacks we can measure
  - Long-term implications -> Trade-offs we are making (need to be measurable)
    - What are opportunities and risks you have listed?
    - What are long-term implications you have mentioned?
- Anticipating the Future: A PreMortem for AI Adoption: it is the year 2035, explore a scenario of "what went wrong".
  - What went wrong?
  - How do you quantify wrong?
- Leading Institutional Transformation: AI Maturity Curve (AMC)— a practical framework designed to assess how effectively an organisation is adopting AI and translating that adoption into meaningful organisational change. Identify measures, shared measure of progress.
  - What kind of measures did you identify here?
- The Evidence for Agentic AI in Education: what current research tells us about its opportunities, limitations and impact of AI in education.
- Writing for Scholarly Publication: high-quality scholarly writing.
- Rethinking Education in the Age of Agentic AI: transform educational institutions, rethink the purpose and practice of education itself.
  - What are we transforming? Way of work? Interaction with people? Scope of education? Ways and means of education? Curriculum? Why? What do we aim to achieve?
- Preparing for the Future of Education: From Prediction to Design in the Age of Agentic AI: technological, social and economic change. Identify actions to build more responsive and future-ready systems.
  - What are the aspects of technological, social and economic change you are interested in?
  - What characterizes responsive and future-ready system?

Module 2: Leading Change and Building Trust

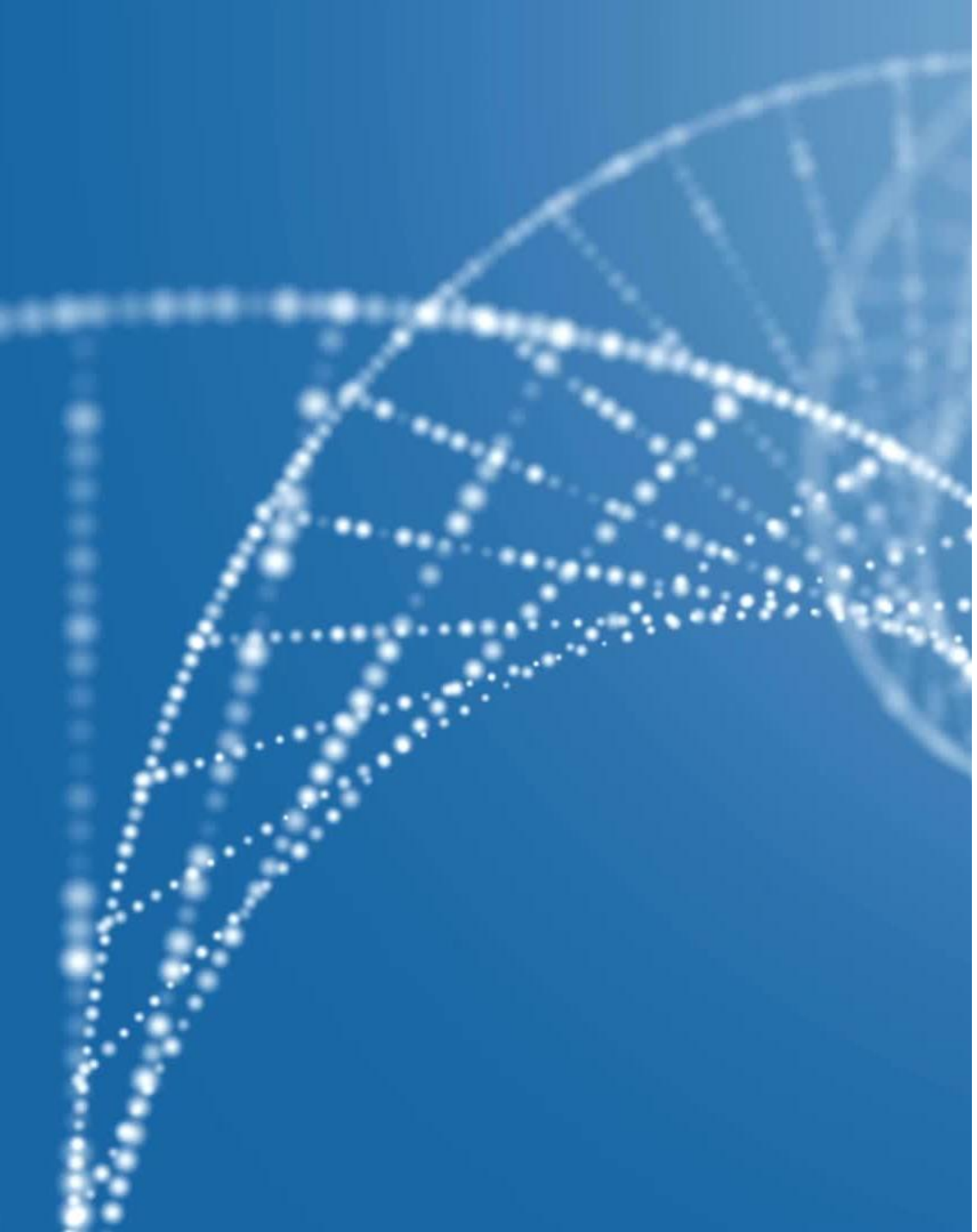
- Exploring Barriers to Change and Innovation: human factors that influence behaviour. Psychological barriers that influence decision-making.
  - What are those human factors that influence behaviour?
  - What are those psychological barriers?
- Building Engagement and Trust for AI Adoption: how leaders can encourage engagement by applying the neuroscience of trust and relationships and responding to the concerns of different stakeholder groups. Successful adoption of AI-enabled innovation.
  - What are concerns you have identified? How does success look like?
- Using Influence Without Authority: change without relying on formal authority.
- Leading Change: evaluate the strengths and limitations of different changemanagement models and consider how they can be adapted to support the responsible adoption of AI within their own educational institutions.
  - What did you evaluate?
- Creating the Conditions for AI Adoption: explore organisational values and behaviours.
- Leadership to Sustain Organisational Readiness for the Successful Adoption of AI: recognise and reinforce the behaviours that underpin long-term change.
  - How does the success look like?

Module 3: Establishing Institutional AI Governance

- From Principles to Practice: Governing AI Adoption in Public Institutions: translating high-level ethical principles into effective governance practices. Examine the gap between policy ambition and implementation capacity.
  - How to measure ambition and implementation capacity?
- Shaping Institutional Approaches to AI: Building Flexible and Coherent AI Governance Strategies: the strengths and limitations of emerging governance models, the challenge of balancing flexibility with consistency, and the practical difficulties institutions face when translating principles into policy and practice.
  - How are these characterized?
- Leading AI Adoption in Education: A Case Study from Hills Road Sixth Form College: effective AI adoption. What are practical challenges of implementation?
  - How to measure effectiveness?
- What Is AI Governance? Exploring accountability and its limits: ways in which AI can be shaped, directed and overseen within institutions. Transparency as a common governance approach.
  - How can we measure transparency?
- Governing AI in Organisations: Decisions, roles and capabilities: where important decisions arise across the AI lifecycle, who should be involved in making and reviewing them, and what roles different actors including leadership, technical teams, or ethics specialists can play.
- Reimagining Assessment: how assessment may evolve over the next two to five years
  - What do you aim to assess?
- Governance and Regulation: the implications of AI and agentic AI for assessment governance, regulation, quality assurance and standards.
  - What are quality aspects that matter?

Module 4: Managing AI Risk, Ethics and Compliance

- Cybersecurity Leadership in the Age of Agentic AI: he nature and potential impact of cyber risk, including risks relating to data, identity, access, third-party dependencies and increasingly autonomous systems.
  - Potential consequences of a breach?
- When Things Go Wrong: Building Organisational Resilience for AI-Enabled Education: explore decision-making under uncertainty, incident response, operational continuity, communications, recovery and organisational learning.
- Ethics in Relation to Emerging AI Technologies: ethical questions raised by emerging AI technologies. Accessibility and inclusion, data privacy and security, consent, environmental impact and the implications of introducing new technologies without established
- institutional policies.
  - What would environment with full accessibility look like?
- Taking Educational Ideals and Professional Ethics Seriously in Responsible AI Leadership: the purposes of education and the professional responsibilities of those working within educational institutions.
- Building Inclusive Processes for AI Decision-Making: Opportunities and Challenges: how decisions are made and whose voices are included.
- Understanding AI Risk and Compliance in Educational Institutions: risks associated with implementing AI in educational settings, including both visible risks, such as data protection and cybersecurity, and less obvious challenges arising from staff and student use of AI, organisational capability and institutional culture.
- Managing AI Risk in Practice: Beyond Regulation and Compliance to Responsible Implementation: practical strategies to manage risk while responding to evolving regulatory expectations.
  - What are the major risks you have identified?



Module 5: Measuring Impact and Value

- LO5: Analyse the impact of agentic AI initiatives using qualitative and quantitative evidence, including educational, operational and organisational metrics.
- LO6: Develop evidence-informed recommendations and policy communications for institutional and public sector stakeholders.

Scoping the expected impact

Setup

- AI as a technology
- AI as a software-based enabler of digitalisation
- Software-enabled system in education
- Quality of the system
- Consequences of the adoption and organisational goals

Adoption of AI

- AI as a technology
- AI-enabled software system and infrastructure
- Goals to achieve with AI-enabled software system
- Challenges to solve to achieve the goals
- Informed decision: suitability of AI to solve challenges
- Deployment of AI-enabled software system
- Consequences:
  - Affected stakeholders
  - **Measure short- and long-term impact**


How does your AI-enabled Software System look like?

Software-Enabled Systems in Education

- Problem domain:
  - Organisational and business drivers
  - Functional drivers
  - Quality drivers
  - Constraints
- Solution domain
  - AI-enabled software system integrated in the education process
  - Validation: whether you are building the right product for the user's actual needs
  - Verification: whether you are building the product right against specifications

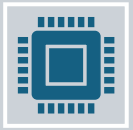
How does your AI-enabled System look like in these terms?

Software Systems in Education

Software systems (AI-enabled)

Functionalities

Constraints



Quality properties: How well software functionalities work (e.g., performance, security, scalability, usability)?

Application of software systems

Organisational goals

Impact of software-enabled solutions (e.g., higher performance of students/lecturers, reducing inefficiencies, higher quality of produced work)

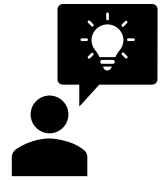
Side-effects

Long-term effects

Stakeholders: Anyone affected by or interested in the outcome of a project.

Lecturer

Student



Administration

Marketing

Parent

Researcher

Goals Scoping Table: Exercise

3.

Goals

- Goals you want to achieve by applying AI in education

1.

Challenges

- Challenges in education you want to solve by using AI

4.

Benefits for stakeholders

- Who will profit?
- How will they profit?

2.

Challenges for stakeholders

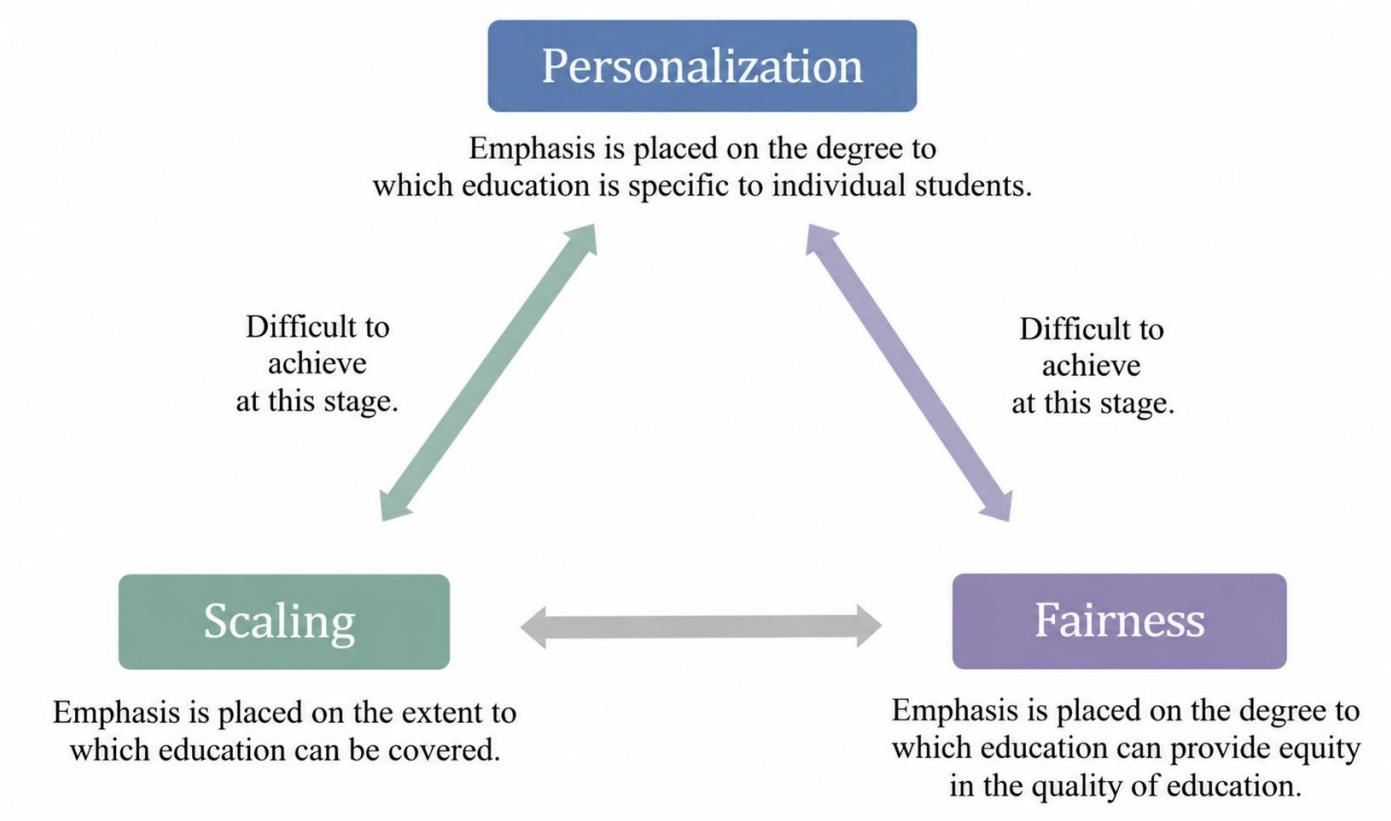
- Who are stakeholders affected by the challenges?
- How are they affected?

Solution domain

Problem domain

The Impossible Triangle

Your challenges and goals in terms of personalization, scaling and fairness - write them down



J. Zhu and L. Fu, "The Latest Innovations and Dilemmas in AI Empowering Personalized Education," 2025 5th International Conference on Artificial Intelligence and Education (ICAIE), Suzhou, China, 2025, pp. 245-249, doi: 10.1109/ICAIE64856.2025.11158694.

WHY: Organisational and Operational Drivers

What kind of organisational and operationa drivers you have in mind?

Time to deliver results	Budget and costs	Competitive pressure	Operational efficiency
Research funding	Research output	Inclusion and accessibility	Student retention
Student engagement	Student grades	Student success after education	…

Update the Goals Scoping Table (Goals) considering previously mentioned Organisational Goals

3.

Goals

- Goals you want to achieve by applying AI in education

1.

Challenges

- Challenges in education you want to solve by using AI

4.

Benefits for stakeholders

- Who will profit?
- How will they profit?

2.

Challenges for stakeholders

- Who are stakeholders affected by the challenges?
- How are they affected?

Solution domain

Problem domain

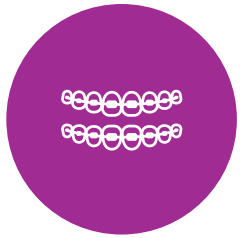
Challenges

What kind of challenges you are actually trying to solve?

COMPLEX PROBLEMS

DEADLINES

MANUAL EFFORT



LACK OF SUPPORT

LACK OF SKILLS

Update the Goals Scoping Table (Challenges) considering previously mentioned Organisational Drivers and Challenges

3.

Goals

- Goals you want to achieve by applying AI in education

1.

Challenges

- Challenges in education you want to solve by using AI

4.

Benefits for stakeholders

- Who will profit?
- How will they profit?

2.

Challenges for stakeholders

- Who are stakeholders affected by the challenges?
- How are they affected?

Solution domain

Problem domain

What kind of beneficial outcomes (considering challenges and goals) you would like to see?

BENEFICIAL OUTCOMES: Organizational Goals

Save	Save time
Improve	Improve throughput, increase the productivity of the individuals and the whole organization
Improve	Improve quality of outcomes of executed actions (precision or completeness)
Educate	Educate people by suggesting solutions outside of their knowledge (innovation, creativity)
Expose	Expose hidden and complex dependencies
Improve	Improve accessibility

Update the Goals Scoping Table (Benefits) considering previously mentioned Beneficial Outcomes

3.

Goals

- Goals you want to achieve by applying AI in education

1.

Challenges

- Challenges in education you want to solve by using AI

4.

Benefits for stakeholders

- Who will profit?
- How will they profit?

2.

Challenges for stakeholders

- Who are stakeholders affected by the challenges?
- How are they affected?

Solution domain

Problem domain

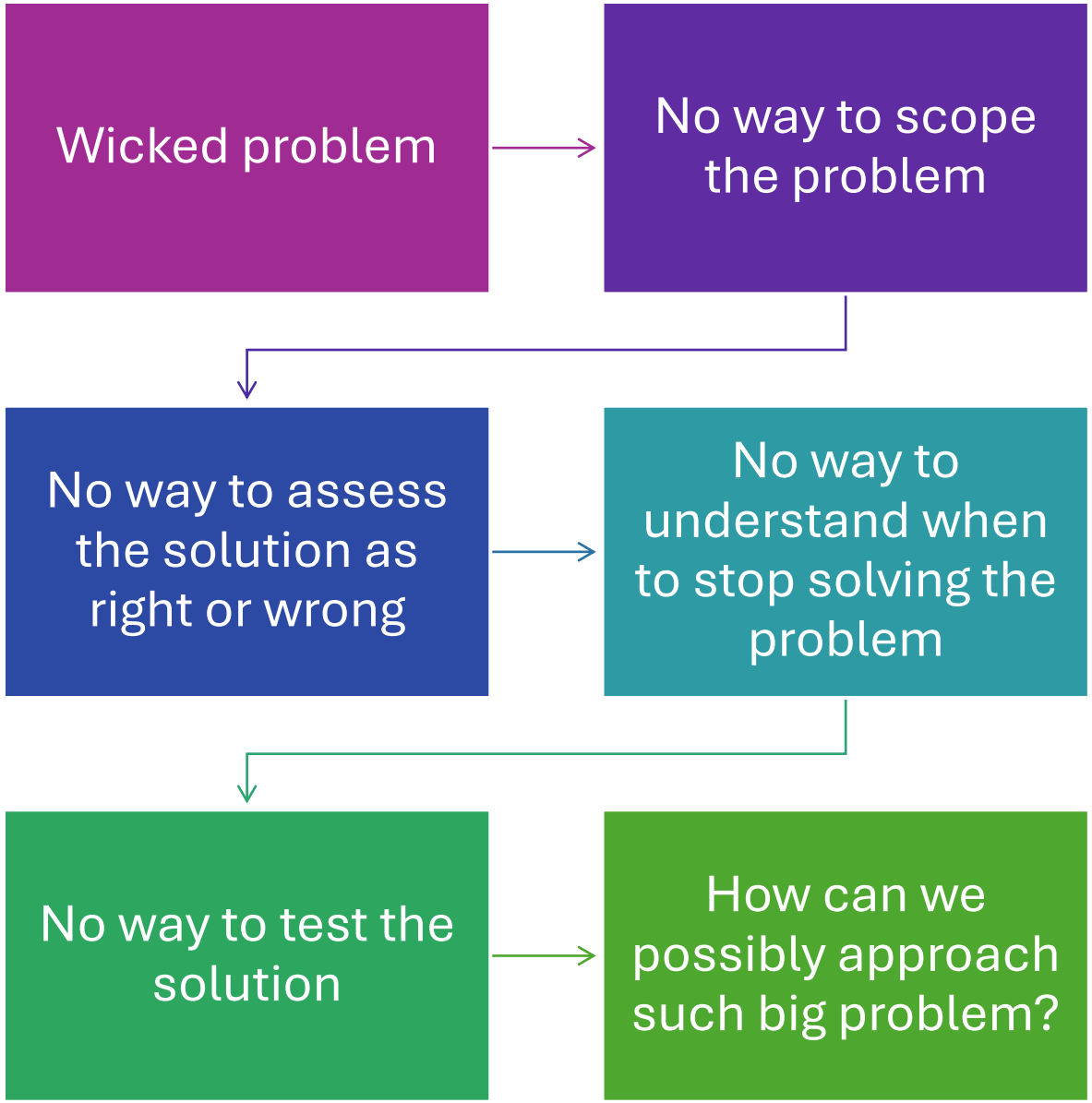
Wicked Problems

Rittel, H.W.J., Webber, M.M. Dilemmas in a general theory of planning. Policy Sci 4, 155–169 (1973). https://doi.org/10.1007/BF01405730

- Problems that are so hard that we are even cannot define them
- E.g., what would be necessary in identifying the nature of the poverty problem?

1. **There is no definitive formulation of a wicked problem**: in order to describe a wicked-problem in sufficient detail, one has to develop an exhaustive inventory of all conceivable solutions ahead of time. The problem can't be defined until the solution has been found.
2. **Wicked problems have no stopping rule**: the process of solving the problem is identical with the process of understanding its nature, because there are no criteria for sufficient understanding and because there are no ends to the causal chains that link interacting open systems, the would-be planner can always try to do better.
3. **Solutions to wicked problems are not true-or-false, but good-or-bad**: solutions are expressed as "good" or "bad" or, more likely, as "better or worse" or "satisfying" or "good enough.“
4. **There is no immediate and no ultimate test of a solution to a wicked problem**: The full consequences cannot be appraised until the waves of repercussions have completely run out, and we have no way of tracing all the waves through all the affected lives ahead of time or within a limited time span.
5. **Every solution to a wicked problem is a "one-shot operation"; because there is no opportunity to learn by trial-and-error, every attempt counts significantly**: every implemented solution is consequential. It leaves "traces" that cannot be undone.
6. **Wicked problems do not have an enumerable (or an exhaustively describable) set of potential solutions, nor is there a well-described set of permissible operations that may be incorporated into the plan**: There are no criteria which enable one to prove that all solutions to a wicked problem have been identified and considered.
7. **Every wicked problem is essentially unique**: despite long lists of similarities between a current problem and a previous one, there always might be an additional distinguishing property that is of overriding importance.
8. **Every wicked problem can be considered to be a symptom of another problem**: There is nothing like a natural level of a wicked problem.
9. **The existence of a discrepancy representing a wicked problem can be explained in numerous ways. The choice of explanation determines the nature of the problem's resolution**: the choice of explanation is arbitrary in the logical sense.
10. **The planner has no right to be wrong**: wicked problems defy efforts to delineate their boundaries and to identify their causes, and thus to expose their problematic nature.

How to Solve Challenges in Education?



How to Solve Challenges in Education?

Scope organisational and individual high-level challenges and goals

Translate big goals into individual manageable challenges

Divide and conquer: Decompose challenges

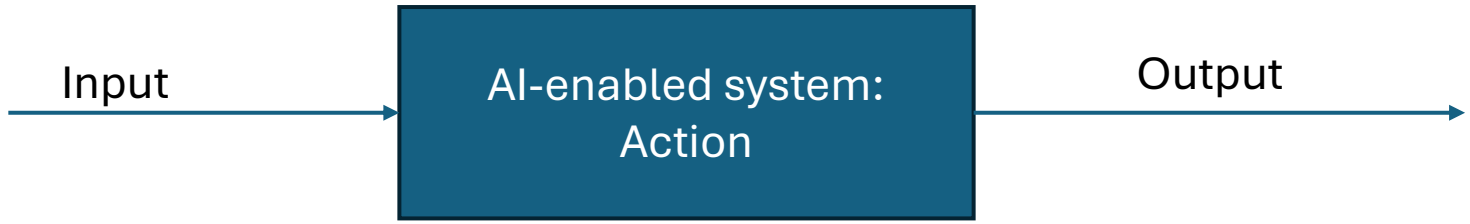
Challenges: Problem Statements

Take one challenge and walk through the table

- Background: [What is the context in which challenges in education exist? Mention stakeholders]
- Consequences: [Elaborate how the challenge influences any of the organisational goals in a negative way: costs, perceived value, and quality of the product or on stakeholders' perception of the work (if the work is stimulating or enjoyable)]
- Common challenges: [What are the actions that stakeholders execute? Elaborate on what the challenges are when executing the action, considering the given background]
- Sources of complexity: [Elaborate on the origins of the challenges – why challenges exist in the first place]

Divide Problem Statements into I/O Sets

- Input: Problem
- Action: Performed by AI
- Output: Solution



Update the Goals Scoping Table considering I/O Sets: One Table per Challenge/Goal

Goals	Challenges
▪ Goals you want to achieve by applying AI in education	▪ Challenges in education you want to solve by using AI
Benefits for stakeholders	**Challenges for stakeholders**
▪ Who will profit? ▪ How will they profit?	▪ Who are stakeholders affected by the challenges? ▪ How are they affected?
Solution domain	Problem domain

An Example from a Publication

Goals

- Improve students' learning experience.

Challenges

- Performance of undergraduate students are not adequate for the market.

Benefits for stakeholders

- Students
- Better skills, can solve real-world problems, higher marks

Challenges for stakeholders

- Students
  - Cannot keep up with industry demands
- Lecturers
  - Struggling to convey all necessary skills; struggling with management of AI solutions

Solution domain

Problem domain

J. Yoo, M. M. Lau, C. Cheng, Y. C. A. Lam and Y. Takaku, "Transforming Student Assignments through GenAI Innovation in Higher Education," 2026 14th International Conference on Information and Education Technology (ICIET), Koriyama, Japan, 2026, pp. 37-40, doi: 10.1109/ICIET69664.2026.11561565.

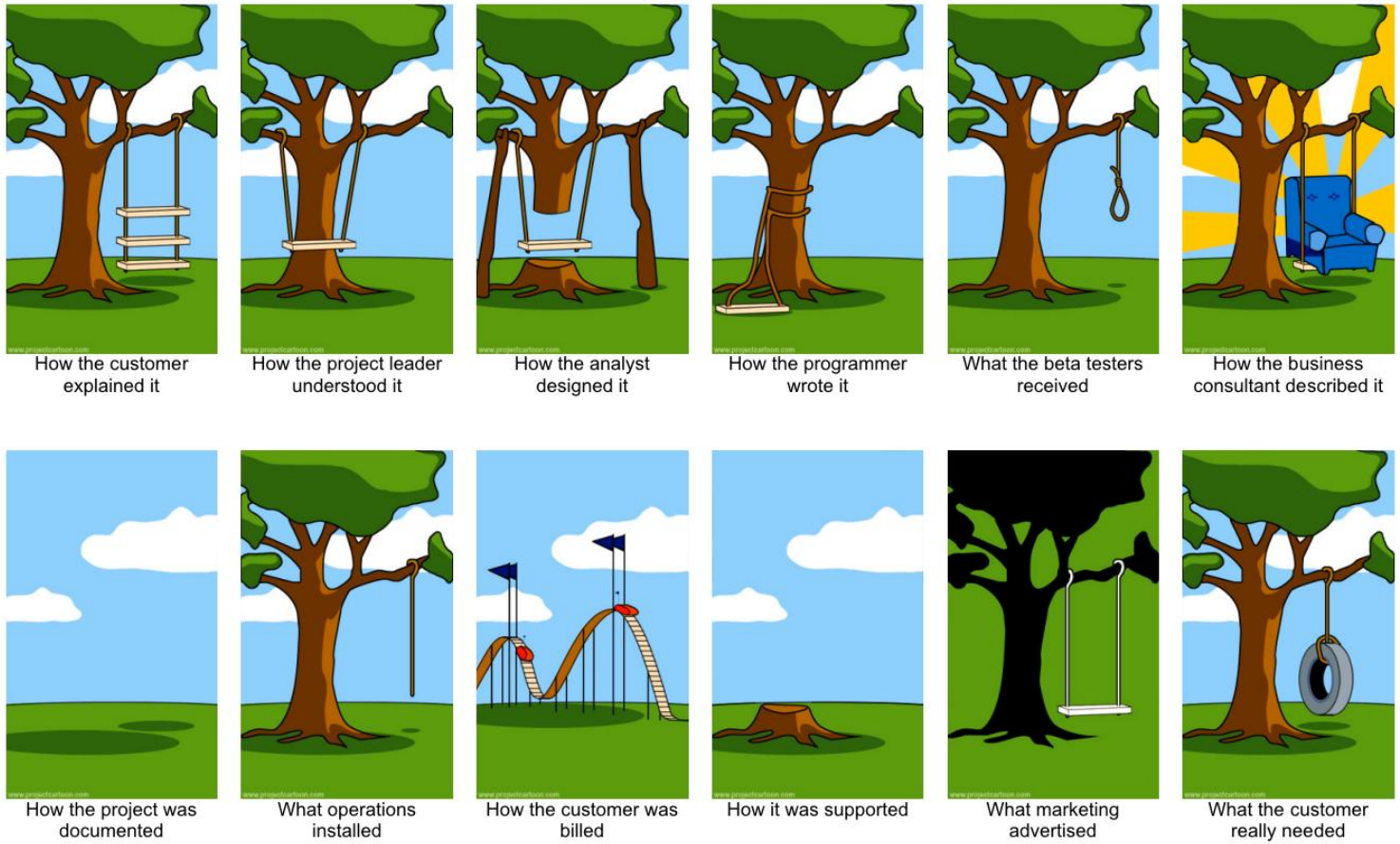
Purpose of Measurement

Problem domain

Solution domain

Understanding the Problem Domain

Challenges with capturing Organisational Goals and Drivers



Challenges with capturing Organisational Goals and Quality Drivers

Time to deliver results: E.g., faster review of essays

How to quantify faster?

Budget and costs: E.g., increase budget and reduce costs

Target budget and for what? Costs of what and by how much?

Competitive pressure: E.g., Improve competitiveness

Considering which aspects? Can it be quantified?

...

Adapted from Jens Knodel and Matthias Naab. 2016. Pragmatic Evaluation of Software Architectures (1st. ed.). Springer Publishing Company, Incorporated.

QUALITY PROPERTIES: TEMPLATE

Quality	Related quality attribute: exactly one attribute should be chosen.	**Stakeholders**	Stakeholders involved
		Quantification	
Environment	Context applying to this scenario. May describe both context and status of the system.		
Stimulus	The event or condition arising from this scenario.		
Response	The expected reaction of the system to the scenario event.		

QUALITY PROPERTIES: EXAMPLE

Quality	Time to deliver results	**Stakeholders**	Stakeholders involved
Time to review student essays		**Quantification**	
Environment	Students have completed and submitted their essays	Number of essays between 100 and 200	
Stimulus	Start reviewing essays with the assistance of AI – AI to identify common typos and missing areas	Number of typos Number of missing areas	
Response	All essays reviewed, comments provided	Reduce the time from 1h per essay review to 30 minutes per essay review	

Quantify all Challenges using Quality Properties Template

3.

Goals

- Goals you want to achieve by applying AI in education

1.

Challenges

- Challenges in education you want to solve by using AI

4.

Benefits for stakeholders

- Who will profit?
- How will they profit?

2.

Challenges for stakeholders

- Who are stakeholders affected by the challenges?
- How are they affected?

Solution domain

Problem domain

Quantify all Goals using Quality Properties Template

3.

Goals

- Goals you want to achieve by applying AI in education

1.

Challenges

- Challenges in education you want to solve by using AI

4.

Benefits for stakeholders

- Who will profit?
- How will they profit?

2.

Challenges for stakeholders

- Who are stakeholders affected by the challenges?
- How are they affected?

Solution domain

Problem domain

Side Effects: AI Properties

Strengths of GenAI

- Retrieve state-of-the-art and state-of-the-practice solutions and domain knowledge
- Suggest innovative solutions
- Identify complex dependencies and detect anomalies
- Suggest solutions with optimised trade-offs
- Personalization
- Automate or assist actions
- Abstract and summarize data sets
- Interpret complex statements
- …

Limitations of GenAI

- Hallucinations
- Precision and completeness challenges
- Bias
- Lack of data
- Reproducibility and verifiability
- Difficulty with logic and arithmetic
- Lack of contextual knowledge
- Misinterpretation and wrong summary
- Too agreeable
- Misuse
- Data confidentiality and copyright
- …

Drawbacks

- Limitations hinder beneficial outcomes.
- E.g.,: "**Save time**" -> "**Hallucinations**" -> "Introduces **overhead** to **evaluate and fix the generated faulty results before adopting them**."

Strengths of AI: Definitions

- **Retrieve solutions and domain knowledge**: GenAI is trained on large datasets, hence it can potentially retrieve adequate state-of-the-art and state-of-the-practice solutions and domain knowledge. Having a quick and summative access to such knowledge, scoped within the problem itself, enables GenAI to suggest solutions previously unknown to its users.
- **Suggest innovative solutions**: GenAI can suggest solution patterns that are new and innovative – that is, uncommon or not obvious to humans due to the patterns' complexity or lack of human creativity. GenAI does this by recombining known ideas into new patterns. One of the most famous examples of creating new solutions is Move 37 in Google DeepMind's AlphaGo game, which had a 1 in 10,000 chance of being used.
- **Identify complex dependencies and detect anomalies**: GenAI can analyse different artefacts such as code, design and reports. It also identifies dependencies and anomalies not obvious to humans. It can also identify complex patterns and unexpected or rare correlations (valuable for detecting risks and predicting behaviours) and support exploratory analysis in complex systems (focus on artefacts such as reports, design, code, tests and so on) and organisations (focus on processes, interactions, flows, toolchains and so on).
- **Suggest solutions with optimised trade-offs**: GenAI can consider many more interdependencies and plan a solution with many more steps in the future than humans can usually maintain in their heads. Therefore, GenAI can create complex and yet efficient solution patterns that are not obvious to humans. Such solutions can offer optimised trade-offs for existing or new solutions (for example, quality versus price, security versus usability). In Google DeepMind's AlphaGo game, it is worth mentioning that moves made by AlphaGo seemed at first like anomalies to some professional players. But, in hindsight and from a broader perspective, these were part of an intentional strategy.
- **Personalisation**: If individuals or organisations execute certain actions in a particular way or with a customised approach, GenAI can identify those patterns. Patterns include language (such as common phrases), sequences of actions, common logic, focus on specific parameters and so on. Pattern recognition is particularly interesting when GenAI identifies actions that always execute together and, therefore, can offer possibilities for automating them.
- **Automate or assist with manual and repetitive actions**: Many actions that SDLC stakeholders execute could be at least partially automated. GenAI can execute such actions, and if its randomness is reduced to an acceptable level, GenAI can be suitable for integration in workflows and toolchains. For a GenAI model to be suitable for automation and assistance, it must execute actions with an acceptable level of reproducibility.
- **Abstract and summarise large datasets**: GenAI is good at summarising large datasets such as code, documents and XML because of its large number of parameters and the self-attention property of transformers (Ranganathan and Abuka, 2022; Yan, 2025). Abstractive summarisation generates a new and unique summary that paraphrases the original text, similar to how a human would perform this task. On the other hand, extractive summarisation creates a summary by selecting and concatenating the most important or relevant sentences or phrases directly from the source document without any changes.
- **Interpret complex statements**: There is often a need to facilitate cross-domain communication and semantic translation in organisations (for instance, 'translating' from finance jargon to technical and vice versa). The internal knowledge representation in neural networks and the way transformers capture knowledge enable GenAI to identify and relate synonyms, and interpolate the meaning of terms and data between abstraction levels and different languages that stakeholders use. However, the interpretation of statements is highly dependent on expectations (that is, what the focus of the interpretation should be), the context of the interaction, prompt engineering and the dataset on which GenAI was trained.
- **Enhance algorithmic tools with probability**: For some actions, it is cumbersome to encode all possible combinations of conditions and solutions algorithmically (like with 'if-else'). Therefore, the presented GenAI strengths are useful for complementing algorithmic tools (such as static analysis) in executing or assisting with such actions.

Limitations of AI: Definitions

- **Hallucinations**: cause plausible-sounding but false or non-existent content. Examples include false statements, false citations, or source code that may look syntactically and stylistically correct but is semantically or factually incorrect, or those with application programming interfaces (APIs) that do not actually exist. GenAI can fabricate sentences and citations, reference books and authors that also do not exist, discuss hallucinated definitions and import non-existent software libraries.
- **Challenges with precision and completeness**: Unlike in the case of hallucinations, where GenAI 'invents facts', GenAI can be optimised to provide fluent and confident answers over complete or fully accurate ones. The facts may be correct, but the logic to provide an output sequence leads to low precision or incomplete results. Probability-driven decoding favours the 'most likely' phrasing rather than the 'most precise' explanation. No properties of 'right and wrong' or 'true and false' exist in these models by default (guardrails can help here), as it lacks intrinsic moral or factual grounding. In other words, it does not 'know' truth, ethics or correctness. It's a statistical model, and it maps inputs to probable outputs.
- **Bias**: Training data can contain societal, cultural and gender biases that GenAI models may unintentionally reproduce or amplify.
- **Lack of data**: For domains that are underrepresented in the training data (for instance, either an industrial domain like finance or mobile apps, or disciplines like software architecture and business management), GenAI's performance drops drastically.
- **Reproducibility and verifiability**: Because outputs are probabilistic and a product of randomness, they are hard to reproduce. Because of this, it is also hard to reproduce the 'line of arguments' that led to the output.
- **Difficulty with logic and arithmetic**: GenAI struggles with consistent logical reasoning and exact numeric operations – GenAI's 'reasoning' is pattern-based, not traditionally algorithmic.
- **No memory/Limited Long-Term Memory (LTM):** GenAI models do not have a concept of (long-term) memory. This can be mitigated to a certain degree by expanding input tokens to include previous results or through the concept of the context window. However, this limitation presents a challenge for continuous operations. Limited memory usually means less contextual precision since, at some point, the model is forced to 'forget' previous outputs.
- **Dependency on prompt engineering**: GenAI models are often packaged as tools to improve their usability (such as in chatbots and agents). These tools often offer commands and configurations to tailor the results. Therefore, users of these tools need to have proper prompting skills as changes in input can greatly affect output (due to the probabilistic nature of GenAI).

Side Effects – Institutional Perception

- Perceived Usefulness
- Perceived Ease of Use
- Social Influence
- Institutional Readiness
- Technological Fit
- Creativity Enhancement
- Peer Support
- Teacher’s Attitudes
- External Influence
- AI Readiness
- AI Intrinsic Motivation
- Data Privacy Concerns
- Perceived Control
- Peer Influence
- Art Design Relevance
- AI Confidence
- AI Anxiety
- User Interface
- Content
- Learner-Interface Interactivity
- Innovation Characteristics
- Psychological Needs
- Facilitating Conditions (FC)
- Trust
- Ethical Concerns
- Performance Expectancy
- Effort Expectancy
- Hedonic Motivation
- Price Value
- Habit
- Subject Norms
- Trialability
- Perceived Compatibility
- Relative Advantage
- Feedback Quality
- Perceived Assessment Quality

I. Zualkernan, "Adoption of Generative AI and Large Language Models in Education: A Short Review," 2025 International Conference on Electronics, Information, and Communication (ICEIC), Osaka, Japan, 2025, pp. 1-5, doi: 10.1109/ICEIC64972.2025.10879632.

Add Side Effects and Quantify them using Quality Properties Template

3.

Goals

- Goals you want to achieve by applying AI in education

1.

Challenges

- Challenges in education you want to solve by using AI

4.

Benefits for stakeholders

- Who will profit?
- How will they profit?

2.

Challenges for stakeholders

- Who are stakeholders affected by the challenges?
- How are they affected?

Solution domain

Problem domain

So far, we have focused on Application of Software

What about quality of software?

AI-enabled software solution

Software Solution Quality Drivers

Internal: How well software functionalities work (e.g., performance, security, scalability, usability)?

External: What is the impact of software-enabled solutions? (e.g., higher performance of students/lecturers, reducing inefficiencies, higher quality of produced work)?

Quality of performed work – reflects in organisational goals

Time to deliver results	Budget and costs	Competitive pressure	Operational efficiency	Research funding	…

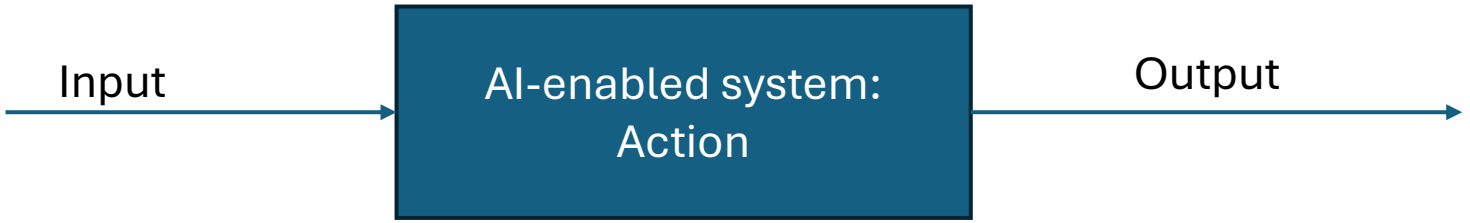
Systems and software Quality Requirements and Evaluation

https://www.iso.org/obp/ui/#iso:std:iso-iec:25023:ed-1:v1:en

- **Functional suitability**: how well a system or software product provides functions that meet stated and implied needs.
  - **Functional correctness**: the degree to which a system or product provides the correct, expected results or outputs for specific inputs.
- **Reliability**: the degree to which a system, product, or component performs specified functions under specified conditions for a specified period of time.
  - **Maturity**: measures are used to assess the degree to which a system, product, or component meets the needs for reliability under normal operation.
- **Usability**: quantitative measures to assess how well a system or software product enables specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use. E.g., how well outputs align with user needs (e.g., output accuracy)
  - **Appropriateness recognizability**: Users have to be able to select a system/software product that is suitable for their intended use. The quality measures for appropriateness recognizability are used to assess the degree to which users can recognize whether a product or system is appropriate for their needs.
- **Performance efficiency**: the capability of a system or software product to provide appropriate performance relative to the amount of resources used under stated conditions
- **Security**: a measurable quality characteristic that evaluates the degree to which a system or software product protects information and data so that people or other products have the degree of access appropriate to their types and levels of authority. E.g., how to mitigate risks like sensitive data leakage.

AI-Enabled Software System

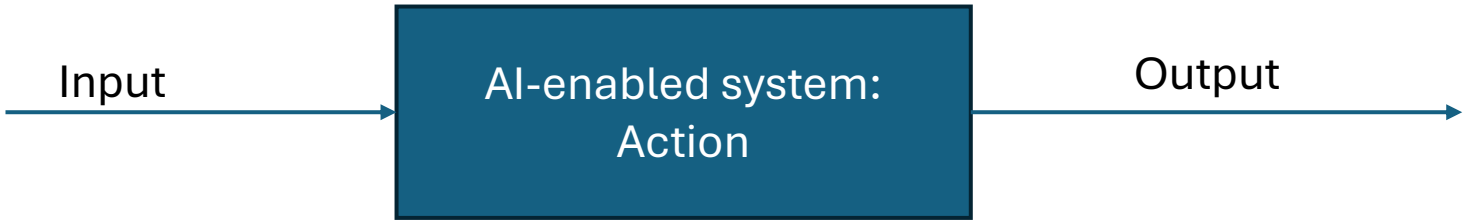
- Main functionalities



- Deliverables (functionalities) that your AI-enabled software system provides: artefacts and actions
  - For start, in the context of your selected challenges and goals
  - What are deliverables and actions?
  - What are functionalities behind deliverables and actions?

AI-Enabled Software System

- Main functionalities



- Quality of functionalities
  - Functional correctness, reliability, maturity, usability, performance efficiency, security, etc.
- Quantify all relevant qualities of functionalities using Quality Properties Template that enable those deliverables

Adapted from Jens Knodel and Matthias Naab. 2016. Pragmatic Evaluation of Software Architectures (1st. ed.). Springer Publishing Company, Incorporated.

QUALITY PROPERTIES: TEMPLATE

Quality	Related quality attribute: exactly one attribute should be chosen.	**Stakeholders**	Stakeholders involved
		Quantification	
Environment	Context applying to this scenario. May describe both context and status of the system.		
Stimulus	The event or condition arising from this scenario.		
Response	The expected reaction of the system to the scenario event.		

Research Paper

Goals + stakeholders = scope

Challenges + expected benefits = research questions

Expected benefits + side-effects = Assessment scope

Presentations

- Prepare 10 - 15 minutes
- 5 minutes presentation
- 2 minutes questions
- No need for slides
- Organic presentation on a flipchart or a board

Exercise: 1

- Motivation
- Goals
- Relevant stakeholders
- Expected benefits
- Expected challenges

- Parameters relevant for
  - Motivation
  - Goals
  - Expected benefits
  - Expected challenges

- Quantification of parameters relevant for
  - Motivation
  - Goals
  - Expected benefits
  - Expected challenges

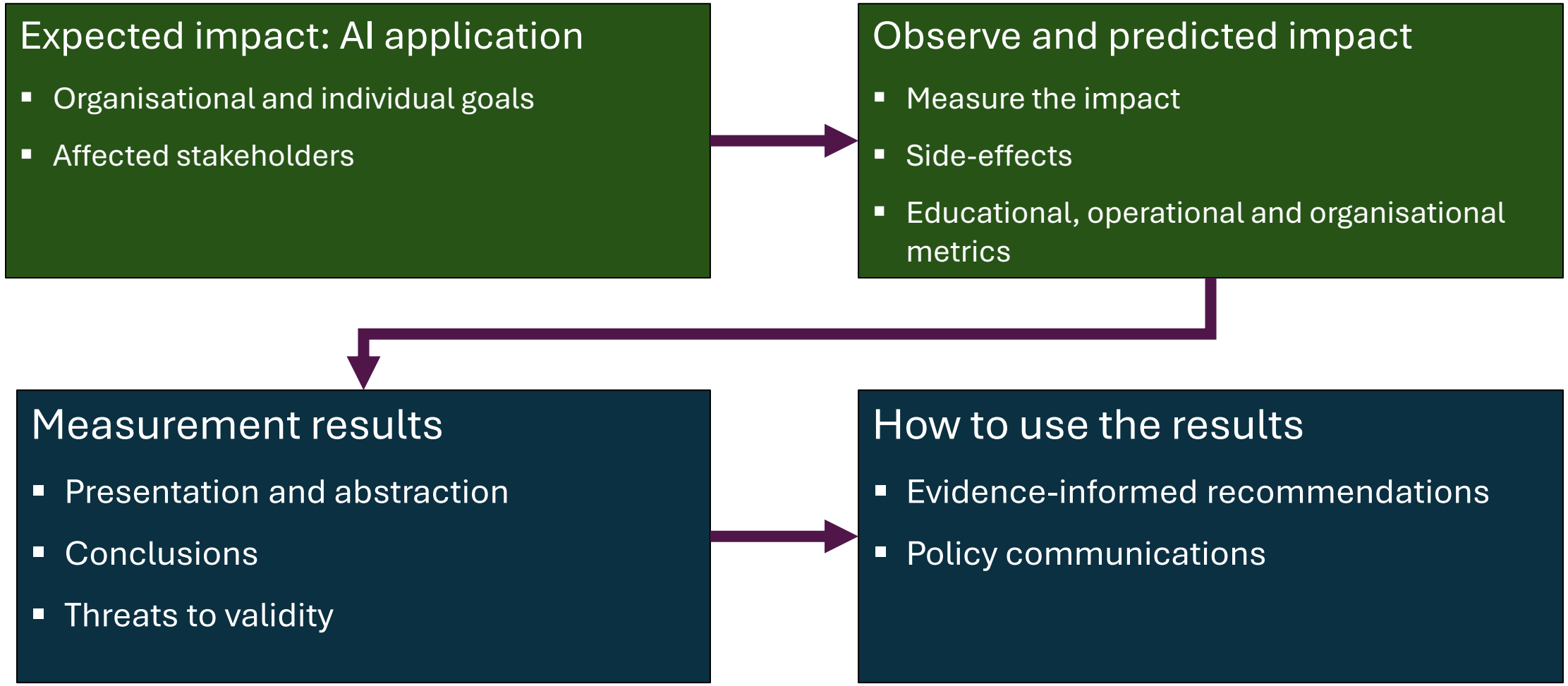
- Side-effects and quantification

Outcomes

- Goals scoping tables
  - Scope of goals
  - Relevant challenges
  - Benefits for stakeholders
  - Challenges for stakeholders
- Quantification templates
  - Quantification of goals
  - Quantification of challenges
  - Quantification of side effects
  - Quantification of software quality properties

- Motivation for applying AI
  - Challenges
  - Who is suffering?
- Expected impact
  - Goals and concrete quantifications
  - Who will profit?
- Deliverables (functionalities) of AI-enabled system: artefacts and actions
- How to measure success?
  - Quantification expectations
- What could possibly go wrong?
  - Side-effects
  - Quality of the software solution

Overview


Expected impact: AI application
Organisational and individual goals
Affected stakeholders
Observe and predicted impact
Measure the impact
Side-effects
Educational, operational and organisational metrics
Measurement results
Presentation and abstraction
Conclusions
Threats to validity
How to use the results
Evidence-informed recommendations
Policy communications

By now, we understand the problem domain

- Challenges worth solving
- Who is affected by the challenges and how
- Goals to achieve by solving the challenges
- Who will profit from solutions and how
- Side-effects that can compromise beneficial outcomes

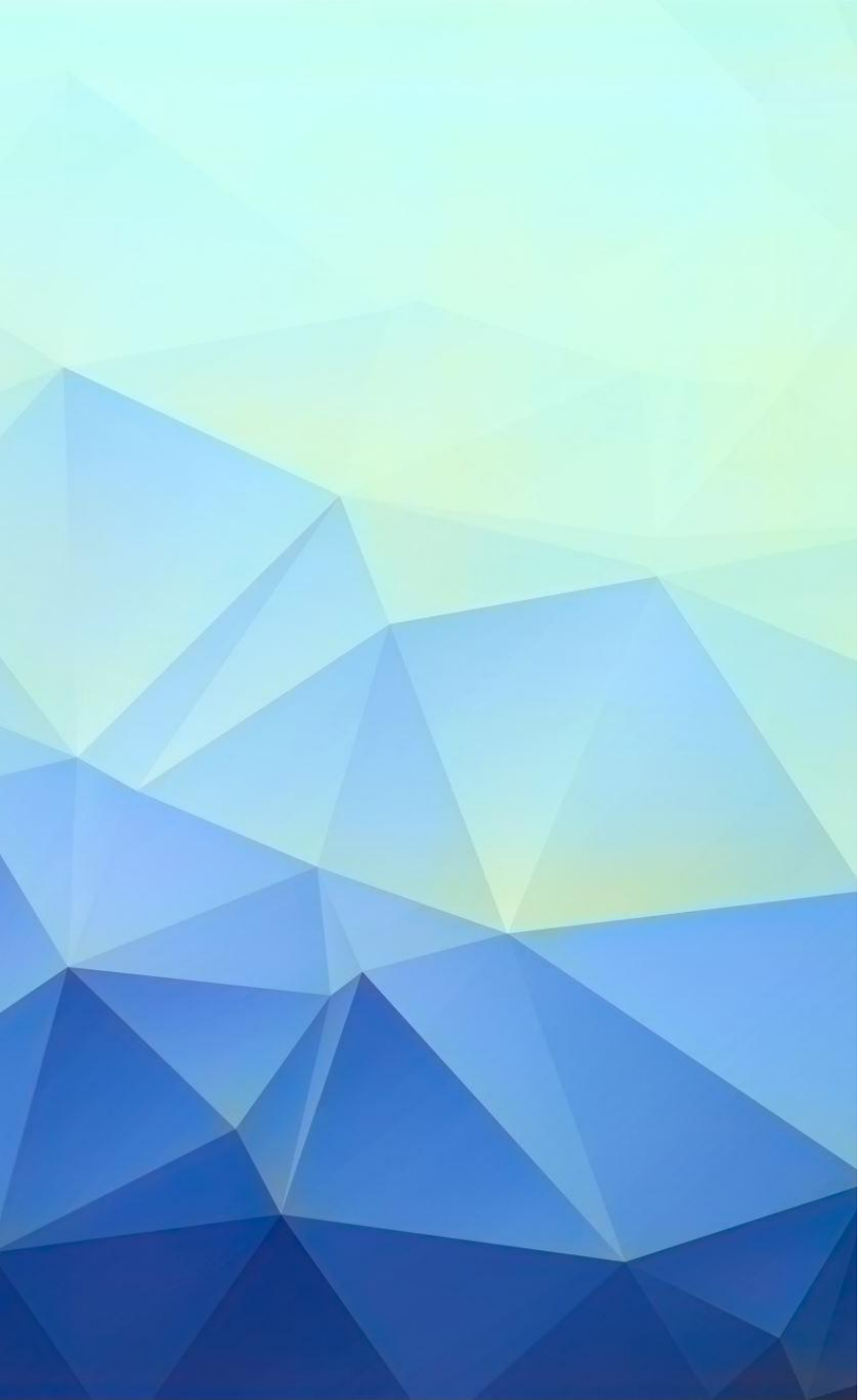
How to measure progress?

Short-term

Long-term

Immediate effects

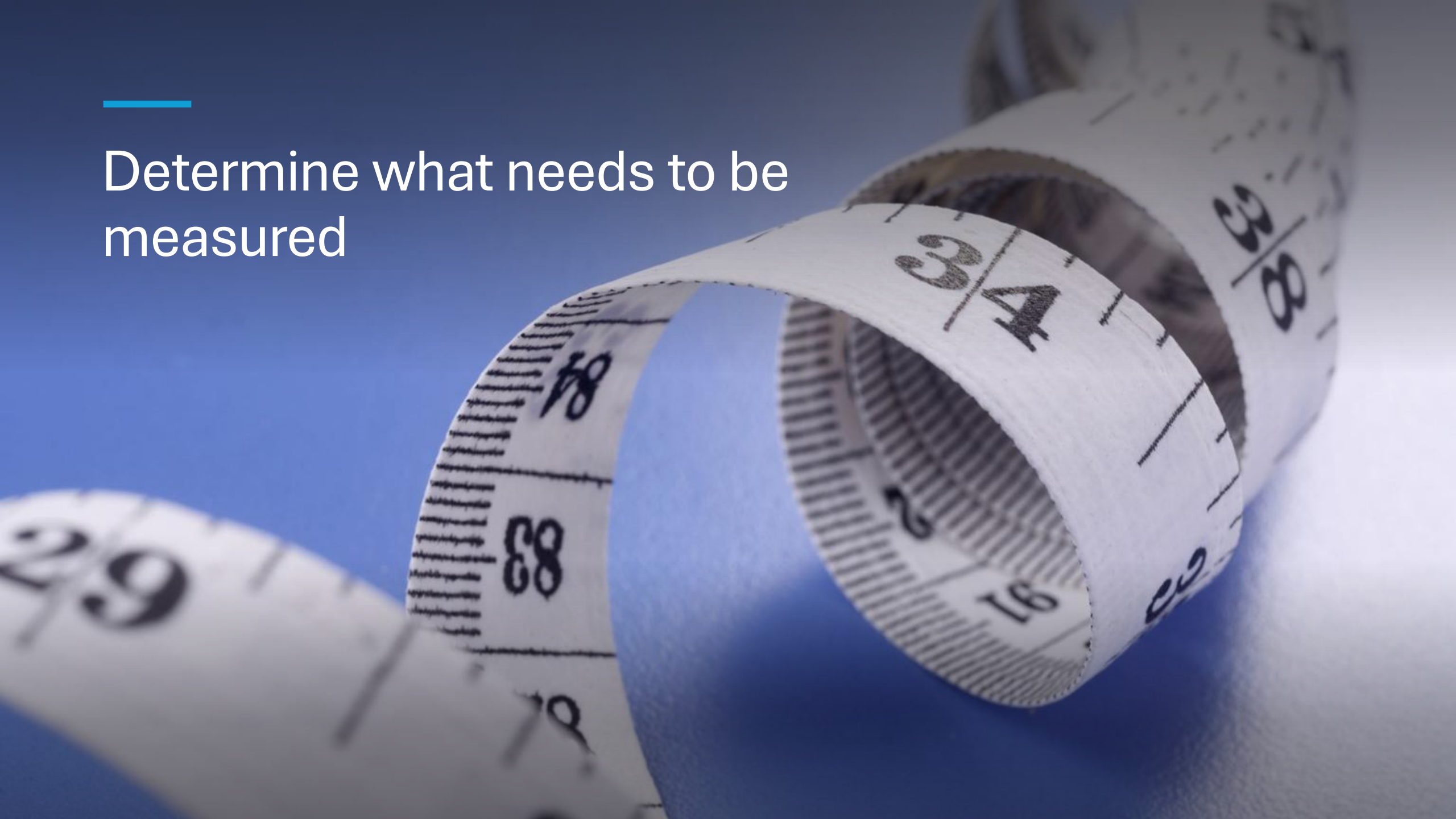
Goals



Measurements

- Metrics
- Measurement techniques
- Measurement protocol
- Intrusivness
- Threats to validity

Determine what needs to be measured



Metrics

IMPORTANT: Quality of AI-enabled solutions for individual tasks

Educational

- Education course specific metrics
- Grades
- Learning process quality
- Learning progress
- Developed skills
- Expanded educational scope
- Practical implementations
- Quality of projects
- Competitions
- Reported innovations
- Performance after graduation
- …

Operational

- Save time
- Improve throughput
- Improve quality of outcomes
- Learn about new topics
- Handle complex tasks
- Knowledge management
- Information sharing
- …

Organisational

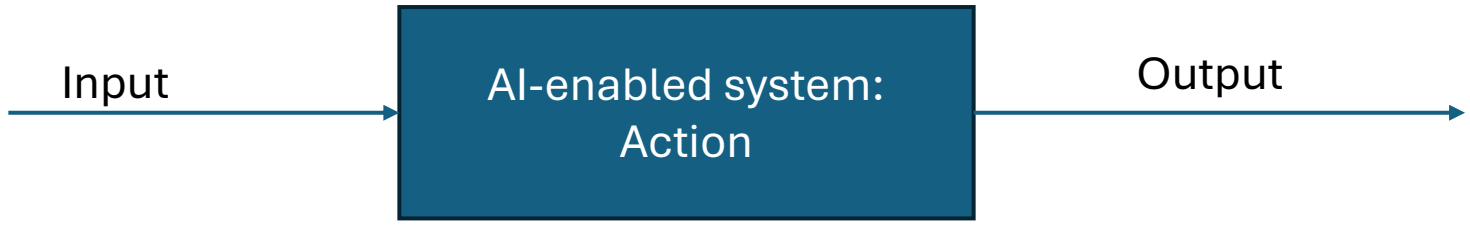
- Lower costs
- Improve reputation
- High-quality research
- Reduced stress
- Improvements in working conditions
- Remove inefficiencies
- Facilitate communication between different groups
- …

But how to measure these?

- Many educational, operational, and organisational metrics
  - Educational metrics:
    - Universal
    - Course, domain, geography, culture, or education level (e.g., high education) specific
  - Operational
    - Universal
    - Educational level (e.g., university) and stakeholder specific
  - Organisational
    - Universal
    - Educational level (e.g., university) and leadership style specific
- They map to a handful quantifiable and descriptive context-independent metrics

I/O Sets

- Input: Problem
- Action: Performed by AI
- Output: Solution



What are the functionalities that your system performs (AI-enabled)?
What are the inputs?
What are the expected outputs?

Common AI Metrics

Which one of these is relevant? For which functionality/artefact/action?

Make notes!

- Correctness = number of accurate output elements / number of total output elements
  - E.g., the accuracy of interactions and structural elements in a diagram
- Clarity: the absence of redundant elements to ensure ease of understanding
- Repetition: Is the presented information free of redundant or duplicated information?
- Similarity: the overlap between the generated outputs and the reference solution.
- Edit similarity: the minimal number of edit operations required to transform the generated output into the reference solution, normalised by the length of the reference.
- Levenshtein distance: calculate how different two texts are. E.g., cat → cut; Only one substitution is needed: a → u So: Levenshtein distance = 1

Common AI Metrics

Which one of these is relevant? For which functionality/artefact/action?

Make notes!

- Completeness = (number of) expected output elements/ (number of) total output elements; E.g., When generating a diagram with visual elements, is anything missing from a diagram?
  - 0: The explanation is insufficient and omits essential details.
  - 1: The explanation includes some relevant details but remains incomplete, hindering full comprehension.
  - 2: The explanation is complete, coherent and provides a comprehensive explanation, facilitating the understanding.
- Relevance: Are the pieces of information in each section to the question relevant and directly addressing its needs?
  - 0: The explanation includes unrelated or incorrect information, impeding comprehension.
  - 1: The explanation demonstrates partial relevance but lacks clarity and contains flaws.
  - 2: The explanation is relevant, clear and logical, effectively aiding in the understanding.
- Conciseness: evaluates whether the generated explanation is clear and focused, avoiding unnecessary details.
  - 0: The explanation includes superfluous details or excessive elaboration, hindering comprehension.
  - 1: The explanation is mostly concise in length and structure, but contains occasional instances of wordiness or redundancy.
  - 2: The explanation is precise, succinct and conveys the intended information clearly and coherently.

Confusion Matrix

Which one of these is relevant? For which functionality/artefact/action?

Make notes!

- True Positive (TP): The model correctly predicted a positive outcome. E.g., AI says Dog → it really is a dog.
- True Negative (TN): The model correctly predicted a negative outcome. AI says Not Dog → it really isn't a dog.
- False Positive (FP): The model incorrectly predicted a positive outcome. E.g., AI says Dog → but it isn't a dog.
- False Negative (FN): The model incorrectly predicted a negative outcome. E.g., AI says Not Dog → but it actually is a dog.

- Accuracy = (TP+TN)/(TP+TN+FP+FN)
  - how many predictions the model got right out of all the predictions.
- Precision = TP/(TP + FP)
  - How many of the "positive" predictions were actually correct - quality of the model's positive predictions
- Recall = TP/(TP+FN)
  - The proportion of true positives detected out of all the actual positive instances
- F1 Score = 2*Precision*Recall/(Precision + Recall)
  - balances trade-off between precision and recall, penalises their imbalance.

High-level to Implementation

	Challenge	Goal	Side-effect	AI-enabled software system
Expectations	Drivers and challenges (time to deliver, budget and costs, manual effort, ...);	Goals (save time, improve throughput, ...)	Hallucinations, bias, reproducibility, ...	Functional suitability, reliability, usability, ...
Metrics	Accuracy, precision, recall, F1, completeness, relevance, conciseness, correctness, clarity, repetition, similarity, edit similarity	Accuracy, precision, recall, F1, completeness, relevance, conciseness, correctness, clarity, repetition, similarity, edit similarity	Accuracy, precision, recall, F1, completeness, relevance, conciseness, correctness, clarity, repetition, similarity, edit similarity	Functional suitability, reliability, usability, ...

To achieve a goal, AI-enabled system produces an artefact or performs an action, solving a challenge, with potential side-effects

	Goal	Challenge	Side-effect	AI-enabled software system
Expectations	Goals (save time, improve throughput, ...)	Drivers and challenges (time to deliver, budget and costs, manual effort, ...);	Hallucinations, bias, reproducibility, ...	Functional suitability, reliability, usability, ...
Deliverables	Artefacts and actions by AI			Delivering artefacts and executing actions
Metrics	Accuracy, precision, recall, F1, completeness, relevance, conciseness, correctness, clarity, repetition, similarity, edit similarity	Accuracy, precision, recall, F1, completeness, relevance, conciseness, correctness, clarity, repetition, similarity, edit similarity	Accuracy, precision, recall, F1, completeness, relevance, conciseness, correctness, clarity, repetition, similarity, edit similarity	Functional suitability, reliability, usability, ...

Decompose previous Quantifications and Map them to AI Metrics

Goals

- Goals you want to achieve by applying AI in education

Challenges

- Challenges in education you want to solve by using AI

Benefits for stakeholders

- Who will profit?
- How will they profit?

Challenges for stakeholders

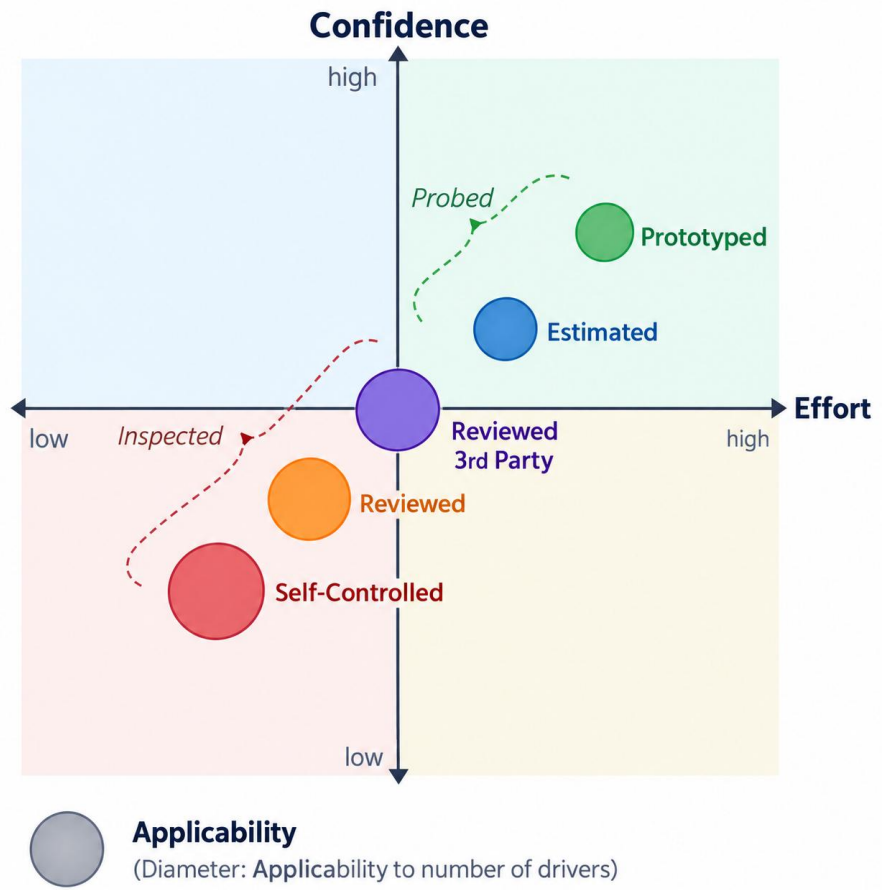
- Who are stakeholders affected by the challenges?
- How are they affected?

Solution domain

Problem domain

Have we understood the Problem Domain?

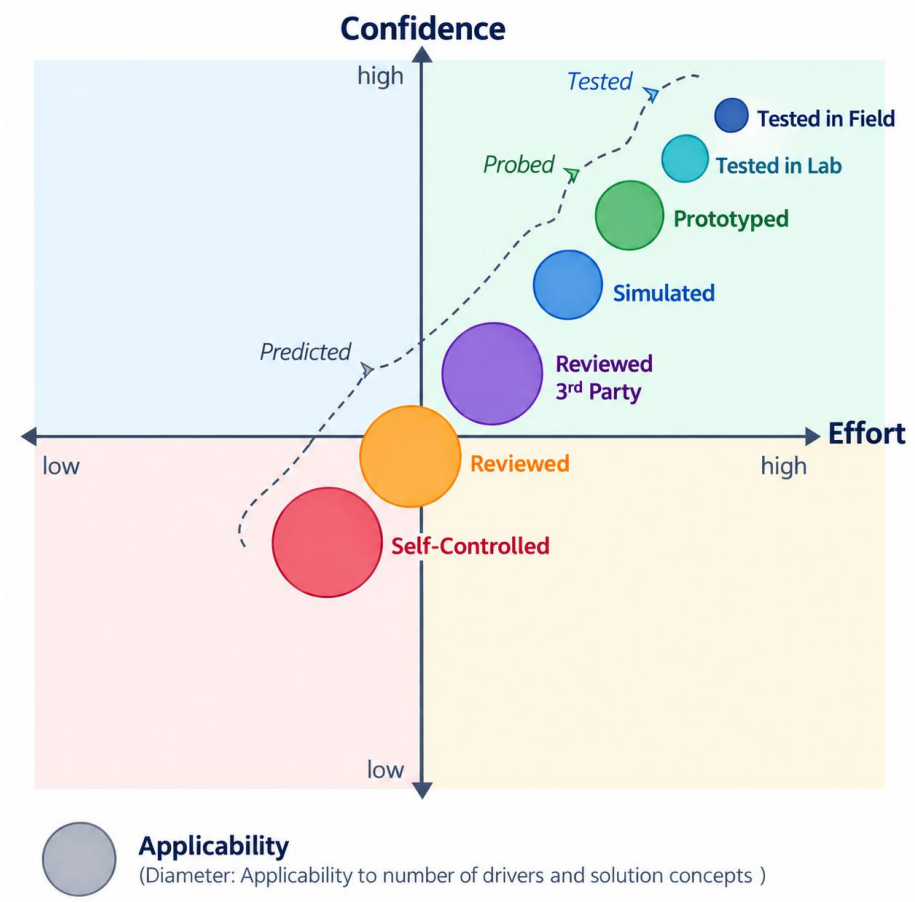
- Involved stakeholders
- Stakeholder concerns
- Areas of interest, recurring items, hot spots, disagreements, and potential conflicts
- Trade-offs
- Merge and unify terminology
- Document all relevant problem-domain aspects



Fraunhofer RATE - Driver Integrity Check (DIC) - Jens Knodel and Matthias Naab. 2016. Pragmatic Evaluation of Software Architectures (1st. ed.). Springer Publishing Company, Incorporated

Have we understood the Solution Domain?

- Motivation
- Goals
- Artefacts and actions
- Benefits
- Risks
- Trade-offs
- Boundaries



Fraunhofer RATE - Driver Integrity Check (DIC) - Jens Knodel and Matthias Naab. 2016. Pragmatic Evaluation of Software Architectures (1st. ed.). Springer Publishing Company, Incorporated

Measurement in Practice

- Challenges and goals
- AI-system in practice
- AI metrics
- Organise measurement: measurement methodologies
  - Data collection
  - Data analysis

Measurement Methodology

Quantitative

- Quantification
- Numbers
- Measurable values
- How much, how often, how many?

Qualitative

- Descriptions
- Meaning
- Opinions
- Experiences
- Relationships
- Dependencies
- Correlations
- Coupling
- Trend
- Indications
- …

Quantitative Analysis

- Surveys: closed-ended questions
- Experiments
- A/B Testing
- Financial Ratio Analysis
- Statistical Forecasting
- Regression Analysis
- The Randomised Control Trial (RCT)
- Mann-Whitney U tests
- T-test
- Bootstrapping: quantitative statistical resampling technique
- Mean Score Rate (MSR)
- Net Promoter Score (NPS)
- Cluster analysis
- Meta-analysis

Surveys

Example?

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

- Survey: a research method/process of collecting information from a sample, often using a questionnaire
  - Questionnaire: set of questions used within that method
  - Closed-ended questions
  - Target audience
  - Opinion/experience based

- Advantages:
  - Easy to standardise
  - Can reach a large number of participants
  - Can be analysed statistically and automatically
  - Can provide anonymity
  - Easy to repeat in future studies
  - Can be done offline

- Disadvantages:
  - Limited ability to explore “why” behind an answer
  - Responses may be superficial
  - Participants may misunderstand questions
  - Self-reported answers may not reflect actual behaviour
  - How to select audience?

Experiments

- Input to output mapping
- Test how changing one factor causes a change in another factor

- Controlled experiment
- Randomised controlled experiment

- Laboratory experiment
- Field experiment

Example?

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

- Advantages
  - Results can be measured objectively
  - Can test cause-and-effect relationships (e.g., we introduced AI and grades improved by 30%)
  - Can provide strong evidence when well designed
  - Can test different configurations

- Disadvantages
  - Can be time-consuming and expensive
  - Artificial settings may not reflect real-world behaviour
  - Ethical constraints may limit what can be tested
  - Experiments might need repetition to remove external disturbances

A/B Testing

Example?

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

- Compare two versions of something to see performance differences
- Group A receives Version A
- Group B receives Version B
- The groups are compared using a predefined outcome or metric

- Advantages
  - Simple to understand and implement
  - Directly compares two alternatives
  - Arbitrary change can be compared
  - Useful for continuous improvement and optimisation
- Disadvantages
  - Potential privacy issues
  - Usually compares only a limited number of changes
  - Requires enough participants to detect meaningful differences
  - Can take time if the effect is small
  - Can be complicated to implement

Financial Ratio Analysis

Example?

- Evaluates organisation's financial performance and position by comparing relationships between figures in its financial statements
- Usually based on secondary financial data, such as company reports and financial statements.

- Advantages
  - Easy to calculate and understand
  - Uses readily available financial data
  - Helps compare performance over time
  - Useful for comparing companies or business units
  - Can identify financial strengths and weaknesses
  - Supports investment and management decisions
- Disadvantages
  - Depends on the quality and accuracy of financial statements
  - Historical data may not reflect current or future performance
  - Sampling rate hard to choose

Statistical Forecasting

Example?

- Predict future values or trends based on historical data and statistical models

- Collect historical data
- Identify patterns or trends
- Select a forecasting model
- Estimate future values

- Advantages
  - Can support planning and decision-making
  - Helps identify trends and seasonal patterns
  - Provides quantitative and reproducible predictions

- Disadvantages
  - Predictions are never completely certain
  - Unexpected events can make forecasts inaccurate
  - Past patterns may not continue into the future
  - Some models require large amounts of data
  - Complex models can be difficult to interpret
  - Models may need frequent updating as conditions change

Regression Analysis

Example?

- Examine the relationship between a dependent variable and one or more independent variables
- Types: linear regression, multiple regression, etc.
- E.g., a university might examine whether students' exam scores are related to GenAI usage - A simple linear regression models the relationship between one predictor and an outcome.

- Advantages
  - Can identify relationships between variables
  - Useful for prediction and forecasting
  - Supports hypothesis testing
  - Widely used and well supported by statistical software
- Disadvantage
  - Does not automatically prove causation
  - Results depend on the quality of the data
  - Relationships may change over time or across different populations

The Randomised Control Trial (RCT) Example?

- Used in medicine - https://pmc.ncbi.nlm.nih.gov/articles/PMC6235704/
- Select the population (e.g., randomised)
- Comparison criteria, measurement criteria
- The number of participants needed to reliably determine the outcome

Other Quantitative Analyses

- Mann–Whitney U test: randomly selected values X and Y from two populations have the same distribution.
- T-test: whether the difference between the response of two groups is statistically significant or not.
- Bootstrapping: quantitative statistical resampling technique - estimate how uncertain or stable a statistical result is, by inserting variations (resampling the existing observations with replacement).
- Mean Score Rate (MSR): express the average score as a proportion or percentage of the maximum possible score.
- Net promoter score (NPS): based on a single survey question asking respondents to rate the likelihood that they would recommend a company, product, or a service.
- Cluster analysis: quantitatively groups similar observations
- Meta-analysis: statistically combines quantitative results from multiple studies

Qualitative Analysis

- Open-ended surveys
- Interviews
- Focus groups
- Case studies
- Decision and solution templates
- SWOT analysis
- Thematic analysis
- The word cloud analysis: visualise which words appear most frequently in a body of text.
- Diaries and journals: Personal logs kept by participants to track thoughts, motivations, and experiences over time.
- Projective techniques: Participants are presented with ambiguous stimuli (like an abstract image or an incomplete sentence) and asked to interpret them. This bypasses conscious filters to reveal subconscious attitudes and emotions.
- Ethnographic fieldwork: Researchers immerse themselves directly into a community or culture for an extended period. They live alongside participants to understand their lifestyle, rituals, and unwritten social rules from the inside.
- Netnography: A specialized form of ethnography adapted for the internet. Researchers observe and analyze the organic behaviours, language, and interactions of online communities, forums, and social media groups.

Surveys

Example?

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

- Survey: a research method/process of collecting information from a sample, often using a questionnaire
  - Questionnaire: set of questions used within that method
  - Open-ended questions
  - Target audience
  - Opinion/experience based

- Advantages
  - Allows participants to answer in their own words
  - Can reveal unexpected ideas or issues
  - Useful for understanding opinions, experiences, and motivations
  - Can explain why participants think or behave in a certain way
  - Avoids forcing respondents into predefined answer categories
  - Can generate themes for further research
  - Useful when the topic is exploratory or not yet well understood
  - Can be done offline

- Disadvantages
  - Responses can be time-consuming to analyse
  - Responses can be off the topic
  - Participants may provide very short or unclear answers
  - Usually requires more effort from respondents
  - Large datasets can become difficult to manage
  - Results are harder to summarise statistically
  - Prone to analysis bias (due to interpretation)

Interviews

Example?

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

- One-on-one conversations using structured, semi-structured, or open-ended questions to explore individual perspectives

- Structured interviews — every participant is asked the same predefined questions in the same order. This makes responses easier to compare.

- Semi-structured interviews — the researcher prepares a set of core questions but can ask follow-up questions and explore interesting responses in more depth.

- Unstructured/open-ended interviews — the conversation is much more flexible. The researcher introduces broad topics and allows the participant to guide much of the discussion.

- Advantages
  - Provide rich and detailed information
  - Allow researchers to explore individual experiences in depth
  - Can reveal motivations and reasons behind behaviour
  - Allow follow-up and clarification questions
  - Useful for complex or sensitive topics

- Disadvantages
  - Require presence of both interviewer and interviewee
  - Time-consuming to conduct
  - Time-consuming to transcribe and analyse
  - Researcher interpretation can introduce bias
  - Interviewer behaviour may influence participants' responses
  - Participants may give socially desirable answers
  - Results can be difficult to generalise to a larger population

Focus Groups

Example?

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

- Guided group discussions to observe collective views and group dynamics.
- Semi-structured
- Explore:
  - Experiences
  - Opinions
  - Attitudes
  - Perceptions
  - Agreements and disagreements
  - How participants react to each other's ideas

- Advantages
  - Collects rich and detailed qualitative data
  - Allows several participants to be studied at the same time
  - Participants can build on each other's ideas
  - Can reveal agreements and disagreements
  - Can generate unexpected insights
  - Often faster than conducting many individual interviews
- Disadvantages
  - Some participants may dominate the discussion
  - Quieter participants may contribute less
  - Group pressure may influence responses
  - Participants may avoid expressing sensitive opinions
  - The moderator can influence the discussion
  - Data can be difficult and time-consuming to analyse
  - It may be difficult to attribute ideas clearly
  - Privacy and confidentiality issues

Case Studies

Example?

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

- An in-depth, multi-faceted investigation of a single "case" (such as one person, an organization, a classroom, a school, or a specific event – e.g., technology implementation).
- It uses multiple data sources to provide a highly detailed, holistic view. Often combine other approaches. E.g., interviews and focus groups.
- E.g., perform experiments and then discuss results in a focus group.

- Advantages
  - Can combine multiple data sources
  - Rich and detailed understanding
  - Can capture organisational and contextual factors
  - Useful for generating new hypotheses or theory
- Disadvantages
  - Can be time-consuming
  - Data collection and analysis can be complex
  - Researcher interpretation can introduce bias
  - It can be hard to separate the effect of one factor from the wider context
  - Difficult to replicate exactly (complex dependencies between different methodologies)

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

Decision-making Template

Rationale (pros and advantages)	**Assumptions and risks (constraints)**
…	…
Scaling factors	**Trade-offs**
…	…

Adapted from Jens Knodel and Matthias Naab. 2016. Pragmatic Evaluation of Software Architectures (1st. ed.). Springer Publishing Company, Incorporated.

Solution Template

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

Steps	… The steps necessary to fulfil the scenario….	
Decisions	Accepted	Rejected
Rationale (Pros)	**Assumptions and Risks (Cons)**	
…	…	
Scaling Factors	**Trade-offs**	
…	…	

Adapted from Jens Knodel and Matthias Naab. 2016. Pragmatic Evaluation of Software Architectures (1st. ed.). Springer Publishing Company, Incorporated.

SWOT Analysis

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

https://github.com/joelparkerhenderson/architecture-decision-record - Templates for decisions

Strengths	**Weaknesses**
Characteristics of the project (GenAI adoption): • helpful in achieving objectives • provide an advantage considering the existing solution (and competition).	Characteristics of the project (GenAI adoption): • hindering objectives • create a disadvantage considering the existing solution (and competition).
Opportunities	**Threats**
The elements that the project (GenAI adoption) could use to its advantage: • external influences • elements in the business environment.	The elements that could hinder the project (GenAI adoption): • external influences • elements in the business environment.

Thematic Analysis

Example?

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

- Identify, organise, and interpret recurring patterns or themes in textual or verbal data.

- Commonly used with:
  - Interview transcripts
  - Focus-group discussions
  - Open-ended survey responses
  - Documents

- Advantages
  - Flexible and widely applicable
  - Can handle large amounts of textual data
  - Helps organise complex qualitative information
  - Can reveal recurring patterns and important themes
  - Suitable for exploratory research

- Disadvantages of thematic analysis
  - Can be subjective
  - Researcher bias may influence coding and theme development
  - Different researchers may identify different themes
  - Can be time-consuming with large datasets
  - Quality depends heavily on careful coding
  - Themes can become too broad or superficial

Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology.
Qualitative Research in Psychology 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

Other Quality Analyses

- The word cloud analysis: visualise which words appear most frequently in a body of text.
- Diaries and journals: Personal logs kept by participants to track thoughts, motivations, and experiences over time.
- Projective techniques: Participants are presented with ambiguous stimuli (like an abstract image or an incomplete sentence) and asked to interpret them. This bypasses conscious filters to reveal subconscious attitudes and emotions.
- Ethnographic fieldwork: Researchers immerse themselves directly into a community or culture for an extended period. They live alongside participants to understand their lifestyle, rituals, and unwritten social rules from the inside.
- Netnography: A specialized form of ethnography adapted for the internet. Researchers observe and analyze the organic behaviours, language, and interactions of online communities, forums, and social media groups.

GQM (Goal–Question–Metric)

Is this adequate for the type of metrics you need? Type of challenges/goals you are addressing? Take notes!

- Goal - define what you want to evaluate or improve
- Question - formulate questions that determine whether the goal is being achieved
- Metric - identify measurable indicators that answer those questions
- Usually organised in a table. One goal can have several questions and several metrics per question are possible
- E.g.,
  - Goal: Evaluate whether GenAI improves student learning.
  - Question: Do students using GenAI achieve better learning outcomes?
  - Metrics: Exam score, Assignment score
- Analyse answers (e.g., open-ended) to identify codes
  - Open Coding: Break the raw data into small meaningful pieces and give them labels.
  - Axial Coding: Connect related open codes and group them into broader categories.
  - Selective Coding: Identify the main or core category that connects the major categories and explains the overall phenomenon.
  - Theory development: explain how the main categories relate to each other.

- Advantages
  - Builds explanations directly from empirical data
  - Helps explain how and why a phenomenon occurs
  - Connects individual themes and categories into a coherent model
  - Can reveal relationships that were not obvious initially
  - Useful for complex or poorly understood topics
  - Produces findings that go beyond simple description
  - Can guide future research
  - Useful for emerging areas

- Disadvantages
  - Can be highly time-consuming
  - Requires extensive and careful coding
  - Researcher interpretation can introduce bias

Mixed-Methods Studies

- John W. Creswell and Vicki L. Plano Clark. 2007. Designing and Conducting Mixed Methods Research. Sage Publications, Inc., Thousand Oaks, CA

Primary vs Secondary Studies

- A primary study collects or analyses original data to answer a research question.

- Experiment
- Survey
- Interview study
- Focus group study
- Case study

- A secondary study analyses or synthesises existing primary studies rather than collecting new empirical data.

- Systematic Literature Review (SLR)

- Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, Stephen Linkman, Systematic literature reviews in software engineering – A systematic literature review, Information and Software Technology, Volume 51, Issue 1, 2009, Pages 7-15, ISSN 0950-5849, https://doi.org/10.1016/j.infsof.2008.09.009.

Measurement Protocol

- Selecting audience/selecting experiments
- Benchmarks
- Preparation
  - Shared info
  - Setup experiments
- Conducting experiment protocol
  - Data storing
- Data analysis

Intrusiveness

- Does conducting a measurement change the reality?

- The Hawthorne Effect
- A psychological phenomenon where people change their behaviour because they know they are being watched or studied
  - 1 experiment working room
  - 1 regular working room

  - Exp.1: Increase light level in the experiment room -> Productivity improves in both rooms
  - Exp.2: Decrease light level in the experiment room -> Productivity improves in both rooms

Threats to Validity

- Internal Validity: Internal validity refers to the cause-effect reasoning within the study.
- External Validity: External validity relates to the generalizability of our findings.
- Reliability: Reliability concerns the repeatability of the study procedures.

C. Wohlin, P. Runeson, M. Hst, M. C. Ohlsson, B. Regnell, and A. Wessln, Experimentation in Software Engineering. Springer Publishing Company, Incorporated, 2012

Formative vs Summative

Formative: Evaluates and refines ideas

Summative: Tests and evaluates systems

Which tools to use?

- Online forms
- Miro/draw.io boards
- Flip charts
- Software programs
- AI analysis?

Presentations

- Prepare 10 - 15 minutes
- 5 minutes presentation
- 2 minutes questions
- No need for slides
- Organic presentation on a flipchart or a board

Exercise: 1

- Motivation
- Goals
- Relevant stakeholders
- Expected benefits
- Expected challenges

- Parameters relevant for
  - Motivation
  - Goals
  - Expected benefits
  - Expected challenges

- Quantification of parameters relevant for
  - Motivation
  - Goals
  - Expected benefits
  - Expected challenges

- Side-effects and quantification

Outcomes – Exercise: 1

- Goals scoping tables
  - Scope of goals
  - Relevant challenges
  - Benefits for stakeholders
  - Challenges for stakeholders
- Quantification templates
  - Quantification of goals
  - Quantification of challenges
  - Quantification of side effects
  - Quantification of software quality properties
- Motivation for applying AI
  - Challenges
  - Who is suffering?
- Expected impact
  - Goals and concrete quantifications
  - Who will profit?
- Deliverables (functionalities) of AI-enabled system: artefacts and actions
- How to measure success?
  - Quantification expectations
- What could possibly go wrong?
  - Side-effects
  - Quality of the software solution

Exercise: 2

- Parameters relevant for
  - Motivation
  - Goals
  - Expected benefits
  - Expected challenges
- Quantification of parameters relevant for
  - Motivation
  - Goals
  - Expected benefits
  - Expected challenges
- Side-effects and quantification

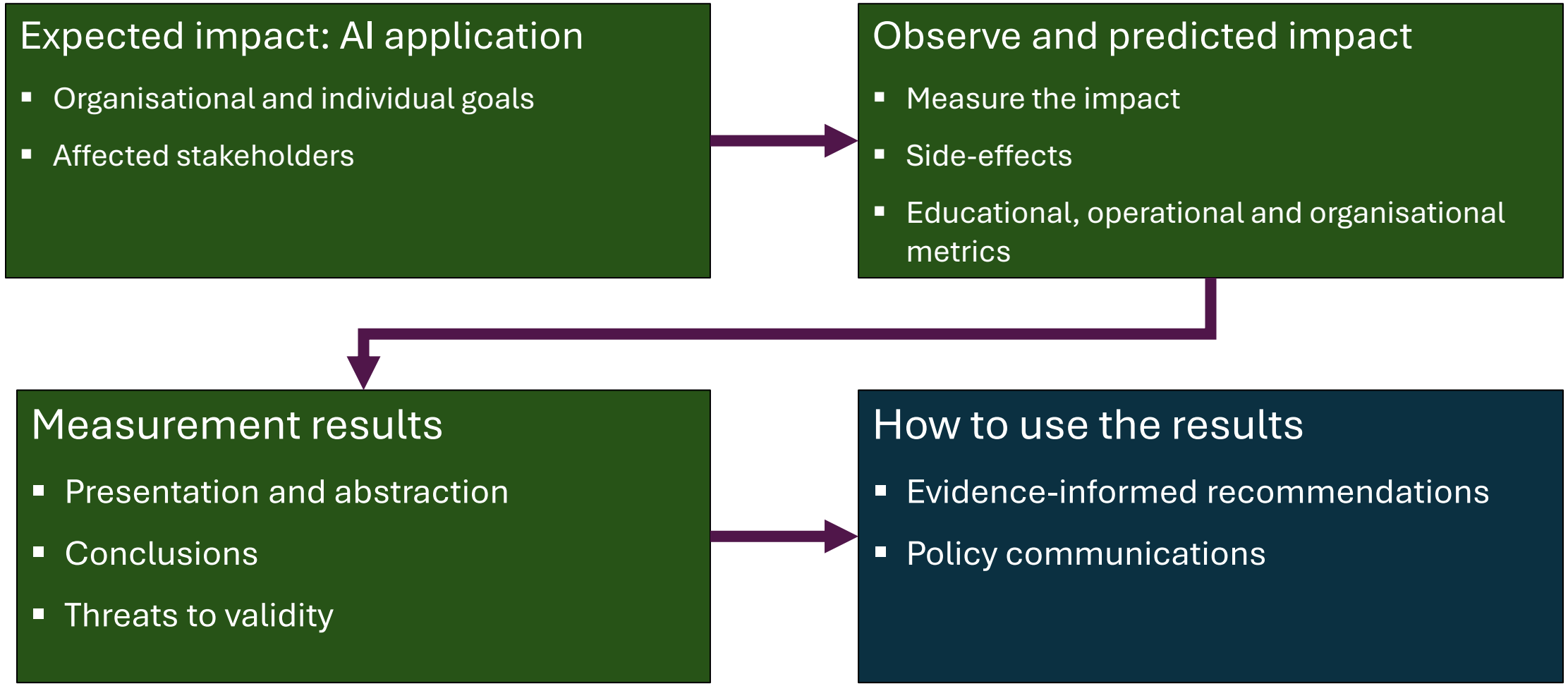
- Further decomposition into measurable parameters
- Further decomposition into confusion, common, and AI metrics
- Approach to collect the data (quantitative or qualitative; concrete techniques)
- Elaboration why the approach is adequate
- Note drawbacks and how to counter them
- Protocol to collect and store the data

Outcomes – Exercise: 2

- Educational, organisational, operational metrics
- Mapping to AI metrics
- Adequate data collection methodology
- Adequate data analysis methodology

- Threats to validity

Overview


Expected impact: AI application
Organisational and individual goals
Affected stakeholders
Observe and predicted impact
Measure the impact
Side-effects
Educational, operational and organisational metrics
Measurement results
Presentation and abstraction
Conclusions
Threats to validity
How to use the results
Evidence-informed recommendations
Policy communications

Evaluation and Assessment in Scientific Papers

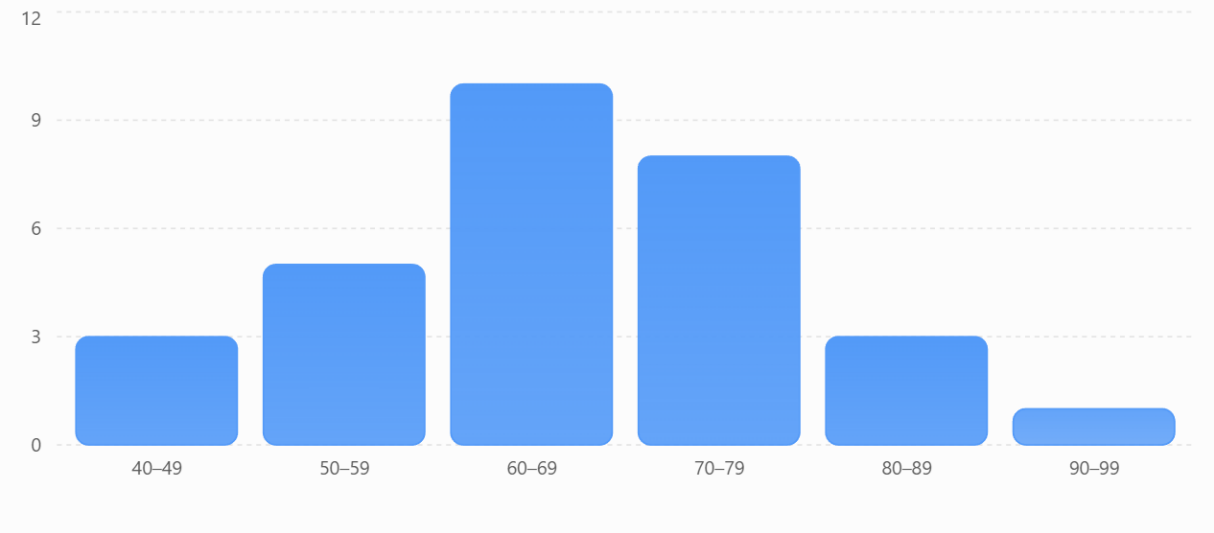
Key factor: REPRODUCIBILITY!	▪ Other researchers MUST BE able to reproduce results following the same setup
Setup	▪ Evaluation goals ▪ Benchmarks ▪ Assessment and measurement methods ▪ Why all these are adequate for the evaluation
Results	▪ Raw results
Discussion	▪ Abstracting, merging, analysing results ▪ Answering research questions
Threats to validity	▪ Critical evaluation of weaknesses of the evaluation

Representation of Results

- Numerical values
- Percentages and proportions
- Minimum and maximum values
- Graphs: e.g., Pie charts
- Tables
- …

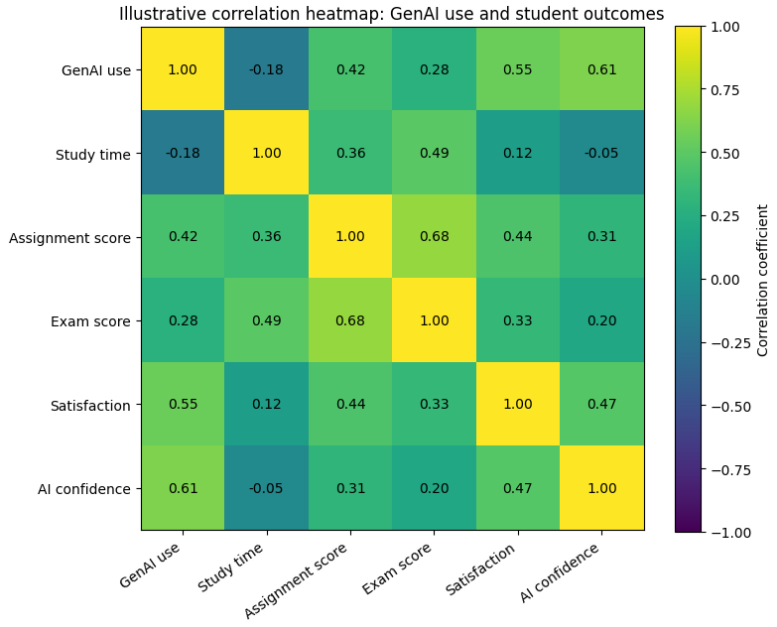
Histograms

- Histograms show the distribution of numerical data. They group numerical values into intervals, called bins, and show how many observations fall within each interval.



Heatmaps

- Use colour intensity to show the magnitude or frequency of values across a matrix or grid.



Add Measurements, make Conclusion

Rationale (pros and advantages)	**Assumptions and risks (constraints)**
Points in favour of adopting GenAI as an adequate solution in this context, experienced GenAI strengths and beneficial outcomes	Assumptions made for selecting GenAI, risks of the adoption, experienced drawbacks and limitations
Scaling factors	**Trade-offs**
What are the factors that will influence the decision? What can we expect to significantly increase?	As GenAI solves one problem, does it create others?

SWOT Analysis

https://github.com/joelparkerhenderson/architecture-decision-record - Templates for decisions

Strengths	**Weaknesses**
Characteristics of the project (GenAI adoption): • helpful in achieving objectives • provide an advantage considering the existing solution (and competition).	Characteristics of the project (GenAI adoption): • hindering objectives • create a disadvantage considering the existing solution (and competition).
Opportunities	**Threats**
The elements that the project (GenAI adoption) could use to its advantage: • external influences • elements in the business environment.	The elements that could hinder the project (GenAI adoption): • external influences • elements in the business environment.

Exercise: 2

- Parameters relevant for
  - Motivation
  - Goals
  - Expected benefits
  - Expected challenges
- Quantification of parameters relevant for
  - Motivation
  - Goals
  - Expected benefits
  - Expected challenges
- Side-effects and quantification

- Further decomposition into measurable parameters
- Further decomposition into confusion, common, and AI metrics
- Approach to collect the data (quantitative or qualitative; concrete techniques)
- Elaboration why the approach is adequate
- Note drawbacks and how to counter them
- Protocol to collect and store the data

Outcomes – Exercise: 2

- Educational, organisational, operational metrics
- Mapping to AI metrics
- Adequate data collection methodology
- Adequate data analysis methodology

- Threats to validity

Exercise: 3

- Goals
- Expected benefits
- Side-effects

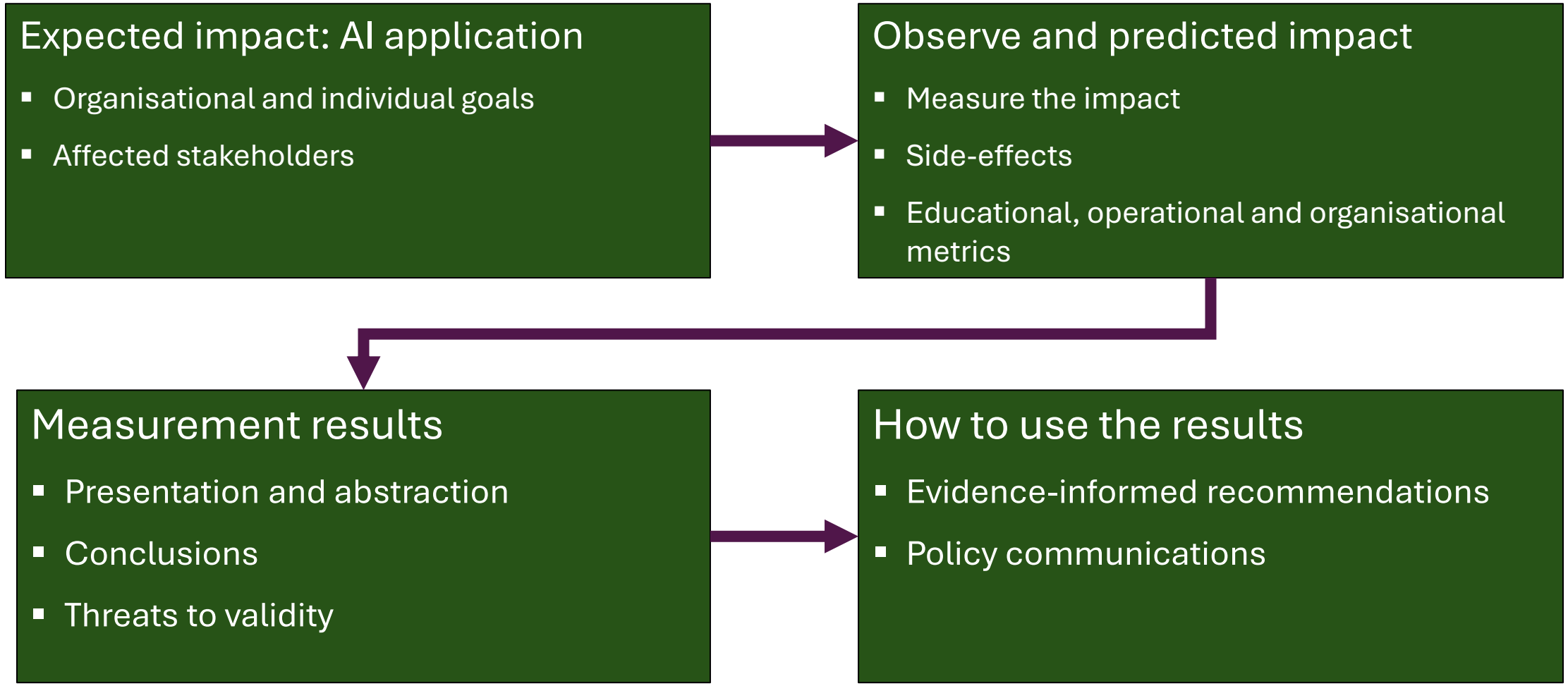
- Measurable parameters
- Approach to collect the data

- Audience/experiments
- Benchmarks/reference values
- Data representation

Outcomes – Exercise: 3

- Measurement framed in a context of a publication
- Results representation

Overview


Expected impact: AI application
Organisational and individual goals
Affected stakeholders
Observe and predicted impact
Measure the impact
Side-effects
Educational, operational and organisational metrics
Measurement results
Presentation and abstraction
Conclusions
Threats to validity
How to use the results
Evidence-informed recommendations
Policy communications

Review of the existing Studies – Take Notes!

- Goals
- Challenges
- Metrics
- Data collection techniques
- Assessment approaches
- Results presentation
- Results abstraction
- Conclusions
- Things to avoid

Practical Educational Applications

- Learning Assessment and Feedback
- Personalized Learning Support
- Instructional Design and Planning
- Intelligent Teaching and Tutoring
- Collaborative Learning and Interaction
- Teacher Professional Development
- Educational Management and Decision-making

J. Zhang and D. Sun, "A Systematic Review of Generative Artificial Intelligence in Education," 2025 7th International Conference on Computer Science and Technologies in Education (CSTE), Wuhan, China, 2025, pp. 552-556, doi: 10.1109/CSTE64638.2025.11092288.

Influence of AI on Grades in Higher Education

- Focus: How does the use of AI affect quality of students' work?
- 46 students, third-year of Finance and Accounting undergraduate studies at a Hong Kong higher education institution during
- 8 groups to ***complete a group project***
- Free use of GenAI (but must be reported)
- **Grades** were assessed based on project quality and presentation, not AI proportion
- The prompt quality (precision, sophistication, alignment with domain knowledge) was assessed as part of the grading
- Structured Q&A probed conceptual mastery, ensuring that GenAI-enabled work was accompanied by human understanding.

Group	Grade	Types of GenAI tools used	GenAI usage %
1	A	GPT Deep Research, ChatGPT o4-mini-high	100
2	A	ChatGPT	80
3	A	ChatGPT, DeepSeek, Grammarly	30
4	B	ChatGPT, DeepSeek	80
5	B	ChatGPT, DeepSeek	70
6	C	ChatGPT	70
7	C	ChatGPT, DeepSeek	60
8	C	ChatGPT	55
Average			68

J. Yoo, M. M. Lau, C. Cheng, Y. C. A. Lam and Y. Takaku, "Transforming Student Assignments through GenAI Innovation in Higher Education," 2026 14th International Conference on Information and Education Technology (ICIET), Koriyama, Japan, 2026, pp. 37-40, doi: 10.1109/ICIET69664.2026.11561565.

Influence of AI on Grades in Higher Education

- The students who received higher scores in group project tend to receive the higher scores in individual exams.
- Students that have higher scores in individual exams tend to have yielded better scores in the group project that uses the GenAI.
- Students who do not have enough subject knowledge have lower chance to receive the good scores in the group project, even though they can rely on GenAI.
- Prompts: Higher-performing groups posed questions that targeted core concepts
- Students were asked to provide short reflective feedback using open-ended questions (results report representative comments):
  - What did you learn from this project?
  - What are the benefits and limitations of using GenAI tools compared to your own research?

The statistics of students' mid-term and final test results in each group-project **scores** of A, B, and C

Group project score	Statistic	A	B	C	t-test (A>C)	Mann–Whitney U	Bootstrap
# students		16	11	17			
Mid	Avg	66	54	55	1.4*	0.072; U=177	0.921; 95% CI [-3.89, 24.10]
Mid	Std	22	25	20			
Mid	Min	18	18	22			
Mid	Max	100	93	89			
Final	Avg	60	54	49	1.62*	0.058; U=176.5	0.95; 95% CI [-2.15, 23.15]
Final	Std	16	20	22			
Final	Min	40	28	10			
Final	Max	88	93	86			

J. Yoo, M. M. Lau, C. Cheng, Y. C. A. Lam and Y. Takaku, "Transforming Student Assignments through GenAI Innovation in Higher Education," 2026 14th International Conference on Information and Education Technology (ICIET), Koriyama, Japan, 2026, pp. 37-40, doi: 10.1109/ICIET69664.2026.11561565.

Influence of AI on Grades in Higher Education

- Pros and Cons of the study and the assessment?

- How are the results assessed?
- Can you reproduce this study?

J. Yoo, M. M. Lau, C. Cheng, Y. C. A. Lam and Y. Takaku, "Transforming Student Assignments through GenAI Innovation in Higher Education," 2026 14th International Conference on Information and Education Technology (ICIET), Koriyama, Japan, 2026, pp. 37-40, doi: 10.1109/ICIET69664.2026.11561565.

Transnational Education (TNE) Sector: Perspectives from Academia and Industry

- Focus:
  - How GenAI tools are being utilised,
  - Skills required for effective human-AI collaboration
  - Necessary Adaptations within TNE frameworks to prepare future Graduates for an AI-driven world

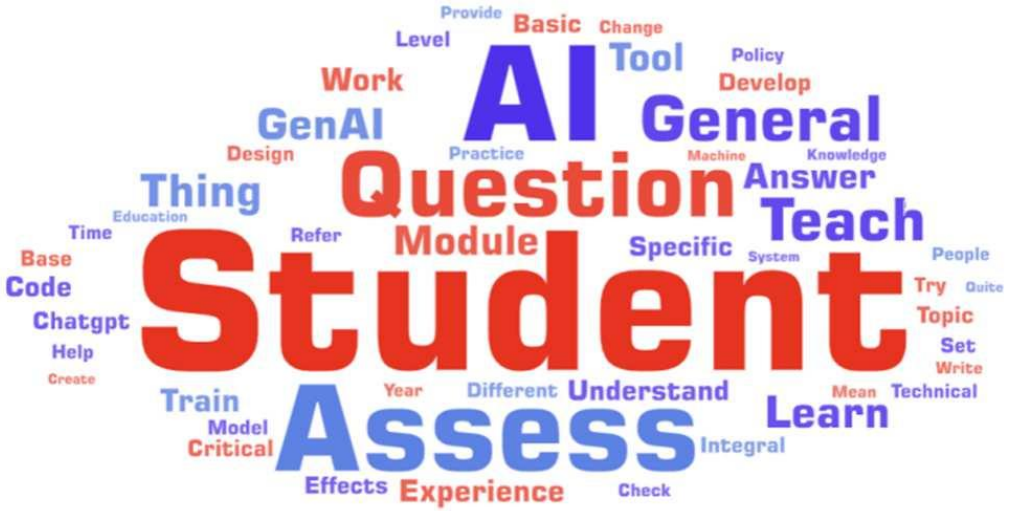
- Industry and academia contacts

- A qualitative research design, explicitly using focus group discussions (FGDs): a semi-structured format to guide the conversation while allowing participants the flexibility to elaborate on their experiences and perspectives
- 28 participants
- Before the sessions, participants were provided with an information sheet outlining the purpose of the study, the discussion topics, and the confidentiality measures in place
- FGDs were conducted in person or online at various locations
- Each session lasted 30 to 120 minutes and was moderated by an experienced facilitator
- All questions were reviewed and approved by the university research ethics committee - alignment with ethical standards

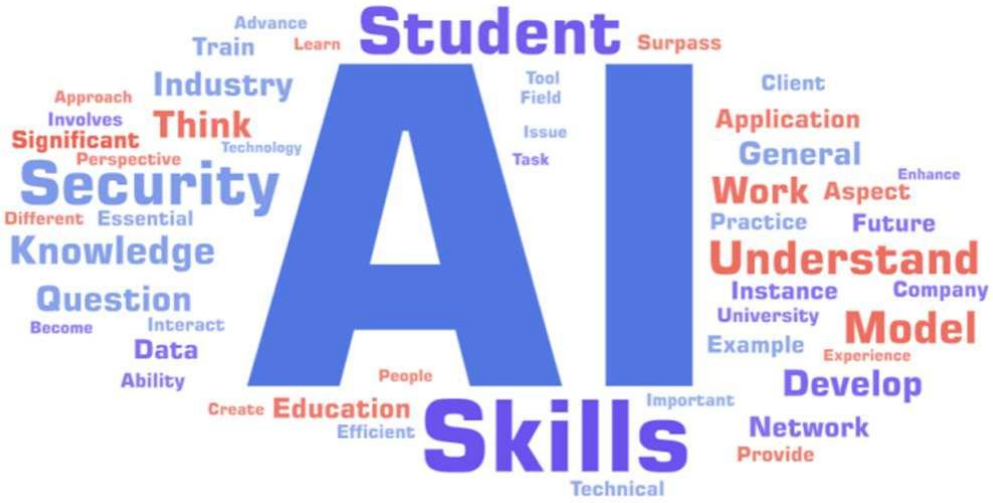
C. Liu, K. K. Chai and Y. Chen, "The Transformative Role of Generative AI in Higher Education: Perspectives from Academia and Industry," 2025 IEEE Global Engineering Education Conference (EDUCON), London, United Kingdom, 2025, pp. 1-6, doi: 10.1109/EDUCON62633.2025.11016643.

Transnational Education (TNE) Sector: Perspectives from Academia and Industry

- The word cloud analysis



Academic focus group



Industry focus group

C. Liu, K. K. Chai and Y. Chen, "The Transformative Role of Generative AI in Higher Education: Perspectives from Academia and Industry," 2025 IEEE Global Engineering Education Conference (EDUCON), London, United Kingdom, 2025, pp. 1-6, doi: 10.1109/EDUCON62633.2025.11016643.

Transnational Education (TNE) Sector: Perspectives from Academia and Industry

Academic focus group

- Usage of Generative AI Tools in Daily Work As A University Educator
  - Legislative restrictions and concerns about maintaining accuracy
  - Speaker 3, FGD2: "I used GenAI to set exam questions, but it (ChatGPT) doesn't answer them correctly. But that's fine"
  - Speaker 2, FGD2: "Being accurate in education is a very important thing. We don't want to deliver the wrong message to the students"
- Creating Engaging and Effective Learning Experiences with Gen AI
  - mixed views on GenAI's role in fostering critical thinking and technical skills, recognising both its potential benefits and challenges
  - GenAI can generate practical examples and scenarios for lab work, enhancing students' understanding of complex tools like MATLAB and improving problem-solving skills
  - Speaker 2, FGD2: "It will save me a lot of time and will give me some interesting examples (use cases) that I can show to the students"
- Novel Assessment Methods for Evaluating Collaboration with AI Tools
  - mixed views on using GenAI in student assessments
  - assessment recommendations: greater use of oral presentations and interactive assessments to develop essential soft skills, alongside regular updates to AI policies to maintain clarity and consistency.
  - Speaker 3, FGD1: "Allow students to use GenAI to assist their critical design."
- Support and Training for Educators and Students
  - need for university support and clear guidelines to effectively integrate GenAI in teaching and research
  - Speaker 2, FGD2: "The first thing is to set an agreement on AI policy. And then the next step will be to think about what training we need"

Industry focus group

- Usage of Generative AI Tools in Daily Industry Operations
  - GenAI is significantly enhancing daily office efficiency
  - Speaker 3: Almost every large company is developing its own AI large models…employees can use prompts to generate content efficiently…this significantly enhances efficiency"
- Critical GenAI Skills and Knowledge for Effective Human-AI Collaboration in the Industry
  - the importance of a balanced skill set, combining technical knowledge with the ability to apply AI in practical scenarios to enhance efficiency and decision-making processes
  - Speaker 5: "You don't necessarily need to optimise algorithms directly, but you should understand the foundational principles and common techniques used in AI"
- Essential Qualities and Skills for Graduates to Collaborate Effectively with AI Systems in Modern Industries
  - Graduates need a strong foundation in technical knowledge, including understanding basic concepts and techniques in their respective fields
  - Speaker 2: "Graduates need to develop unique skills and capabilities to stand out…A strong grasp of technical knowledge is essential"
- Evolving Skill Sets for the Workforce in the Age of AI
  - Critical thinking, problem-solving, and the ability to ask precise questions will become increasingly important as AI takes over routine tasks.
- Preparing Graduates to Leverage GenAI Tools in the Future Workforce
  - incorporate industry-specific elective courses, internships, and opportunities for competition participation
  - Speaker 2: "Graduates need to develop unique skills and capabilities to stand out… A strong grasp of technical knowledge is essential"

C. Liu, K. K. Chai and Y. Chen, "The Transformative Role of Generative AI in Higher Education: Perspectives from Academia and Industry," 2025 IEEE Global Engineering Education Conference (EDUCON), London, United Kingdom, 2025, pp. 1-6, doi: 10.1109/EDUCON62633.2025.11016643.

Transnational Education (TNE) Sector: Perspectives from Academia and Industry

- Pros and Cons of the study and the assessment?

- How are the results assessed?
  - Qualitative thematic/descriptive analysis?
  - They do not explicitly describe a recognised qualitative analysis procedure
- Can you reproduce this study?

C. Liu, K. K. Chai and Y. Chen, "The Transformative Role of Generative AI in Higher Education: Perspectives from Academia and Industry," 2025 IEEE Global Engineering Education Conference (EDUCON), London, United Kingdom, 2025, pp. 1-6, doi: 10.1109/EDUCON62633.2025.11016643.

AI for English Education

- Goal: Use AI to improve English skills
- Pre-defined questionnaire
- Two classes, 99 participants
- Higher vocational students upgrading to undergraduate colleges
- The teacher instructed students on the usage of prompts, providing examples and
- guiding them to describe their needs for GenAI to guide AI to generate targeted solutions for their requirements
- 16-week study
- Training sessions for lecturers (summer workshop)
- Pre (Figure 1 and 2) and post study questionnaire (Figure 3 and 4)

Yiman Ke. 2025. The GenAI Application of Personalized Learning and Human-AI Collaboration in English Education. In Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area Education Digitalization and Computer Science International Conference (EDCS '25). Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/3746469.3746496

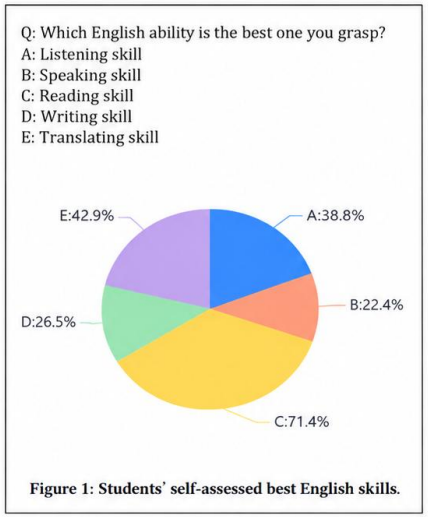


Figure 1: Students' self-assessed best English skills.

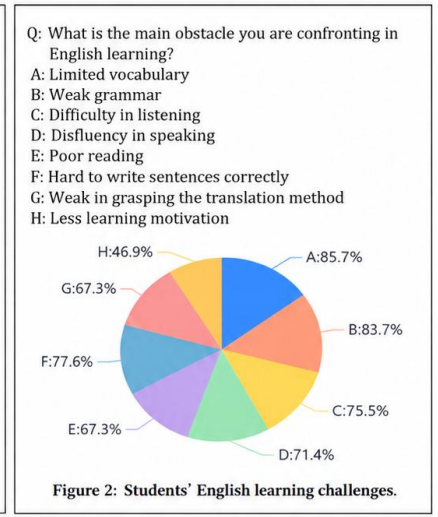


Figure 2: Students' English learning challenges.

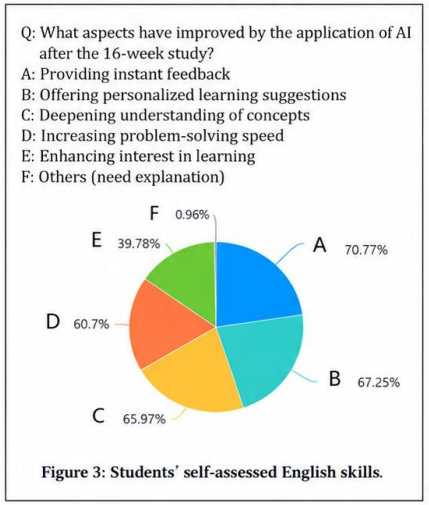


Figure 3: Students' self-assessed English skills.

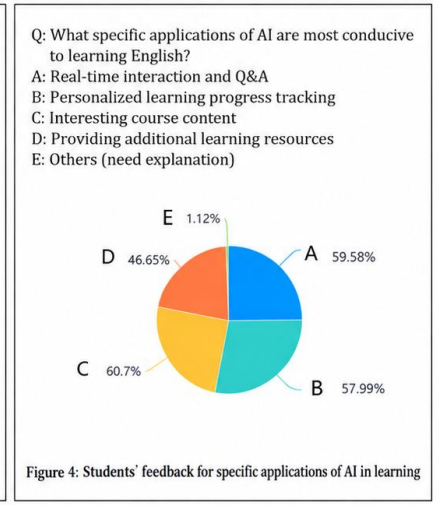


Figure 4: Students' feedback for specific applications of AI in learning

AI for English Education

- Pros and Cons of the study and the assessment?

- How are the results collected and assessed?
- Do results provide enough details to answer the questions?
- Can you reproduce this study?

Yiman Ke. 2025. The GenAI Application of Personalized Learning and Human-AI Collaboration in English Education. In Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area Education Digitalization and Computer Science International Conference (EDCS '25). Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/3746469.3746496

AI in Human-Computer Interaction (HCI) Education

- Goal: demonstrate to students both the potential benefits and inherent limitations of utilising GenAI when designing evaluation studies, while highlighting how GenAI might suggest ideas they previously overlooked or, conversely, fail to identify important aspects they themselves proposed
- Research questions:
  - How does GenAI support students' understanding of evaluation methods in Human-Computer Interaction (HCI) education?
  - What are the implications that emerge from using it in HCI education?
- Mixed-methods study:
- Measure: students' perceptions of how helpful GenAI was in enhancing their understanding of evaluation methods, the specific evaluation activities in which they used GenAI, and their views on the advantages and limitations of using AI during these tasks.
  - quantitative overview of trends and patterns in students' perceptions after the practical session
  - qualitative data: open-ended questions in the survey
  - quantitative data were analysed using descriptive statistics to reveal overall trends in perceived usefulness and patterns of GenAI utilisation. Inductive thematic analysis was conducted on qualitative responses using NVivo to identify common themes. The analysis followed Braun and Clark's phases - Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. Qualitative Research in Psychology 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa
- 74 registered undergraduate students - post-activity survey completed by 46 students
- In the lecture, the students were given three group activities, each about a different type of software
- Students discussed four questions:
  - what are the three most important things to evaluate
  - can the system be tested in a lab
  - would an experiment help and what would it measure
  - should it be tested by UX experts or real users
- After their group discussions, they were provided with prompts that performed the same activity as them (i.e., AI suggesting answers to the same activity) and asked to compare their answers with those of AI.
- In practical sessions, students performed various activities. E.g.,
  - students were asked to imagine themselves as expert evaluators, guided by GenAI (acting as a facilitator
  - having GenAI perform evaluations on their behalf. Students were encouraged to experiment with any evaluation method they believed GenAI could handle best

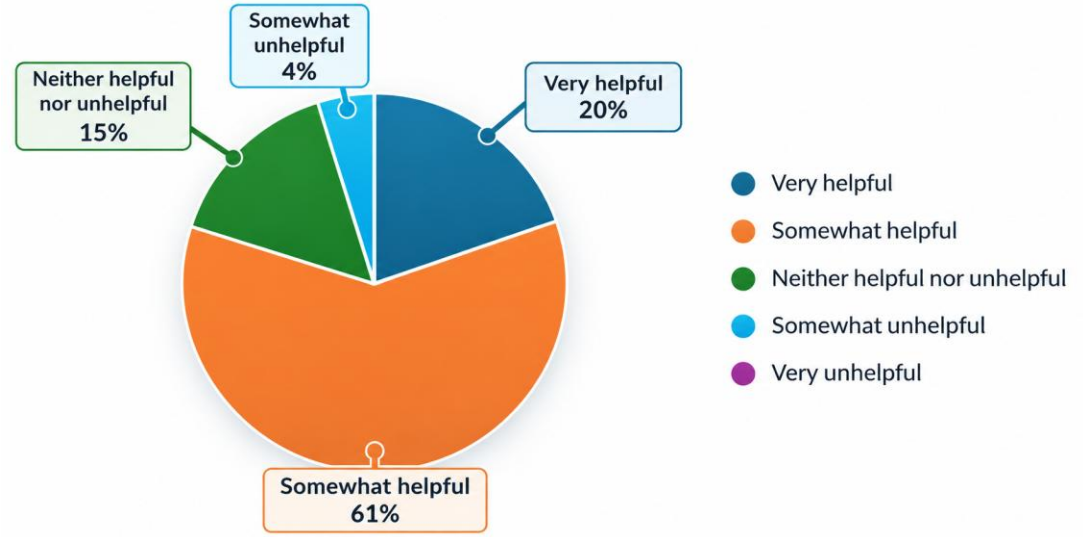
Abeer Aziz, Ahmed Kharrufa, and Ian Johnson. 2026. Integrating GenAI into HCI Pedagogy: Supporting Students' Learning of Evaluation Methods. In Proceedings of the 8th Annual Symposium on HCI Education (EduCHI '26). Association for Computing Machinery, New York, NY, USA, Article 15, 1–8. https://doi.org/10.1145/3803869.3803872

AI in Human-Computer Interaction (HCI) education

- GenAI as a reflective comparison partner
  - GenAI helped surface evaluation considerations they had not thought of
  - students identified important gaps in AI-generated suggestions
  - “the quality of the responses. . . depend on the quality of the prompts,”
  - GenAI is most effective when used after independent thinking, to “either seek confirmation or highlight what you missed.

- Emerging patterns in students’ use of GenAI in Evaluation methods
  - “Using GenAI in this practical made the evaluation process more structured and organised. It guided me step by step through the cognitive walkthrough and reminded me to think about key usability questions.”

GenAI Helpfulness in Learning Evaluation Methods



Abeer Aziz, Ahmed Kharrufa, and Ian Johnson. 2026. Integrating GenAI into HCI Pedagogy: Supporting Students’ Learning of Evaluation Methods. In Proceedings of the 8th Annual Symposium on HCI Education (EduCHI '26). Association for Computing Machinery, New York, NY, USA, Article 15, 1–8. https://doi.org/10.1145/3803869.3803872

AI in Human-Computer Interaction (HCI) education

- Pros and Cons of the study and the assessment?

- Can you reproduce this study?
- How are the results collected and assessed?
- How are the results presented?
- Do results provide enough details to answer the questions?

Abeer Aziz, Ahmed Kharrufa, and Ian Johnson. 2026. Integrating GenAI into HCI Pedagogy: Supporting Students' Learning of Evaluation Methods. In Proceedings of the 8th Annual Symposium on HCI Education (EduCHI '26). Association for Computing Machinery, New York, NY, USA, Article 15, 1–8. https://doi.org/10.1145/3803869.3803872

Generative AI in Higher Education: A Student Perspective

- Goal: A systematic meta-analysis and dimension analysis to investigate the adoption of Generative AI (GenAI) among higher education students
- RQ1. What are the students' perspectives and experiences regarding the use of Generative AI in higher education?
- RQ2. What factors affect students' acceptance and use of Generative AI?
- Systematic literature review - Keyword Cooccurrence Network
- Step 1 : systematic synthesis of research articles to identify and analyze student perspectives, experiences, and influential factors related to the use of GenAI in higher education.
- Step 2: a comprehensive framework of the five key dimensions that influence higher education students' acceptance and use of Generative AI technology, based on the findings synthesized in Step 1.

- Exposure and Actual Use - Students widely use GenAI tools (such as ChatGPT) to help them write, summarize, and think about ideas. Q&A and Class Participation. Usage continues despite the institution's policy being unclear or conflicting.
- Motivations for Use - GenAI is seen as a tool that promotes creativity and enhances learning. Supporting equity. User-Driven Learning and Administrative Task Management. Encourage Engagement, Content creation, and cross-disciplinary communication.
- Concerns and Uncertainties - There are ethical concerns such as over-reliance, bias, surveillance, academic integrity, and improper use of data. The usability and reliability of the data are questioned. Teachers and students are also concerned about plagiarism and misinformation.
- Learning Needs and Policy Clarity - Calling for an ethical policy, AI literacy, and training are clear. There must be a balanced approach to learning management that combines humans and AI. Academic integrity and responsible use of AI are emphasized.

W. Milinthapunya, P. Nilsook and P. Wannapiroon, "Generative AI in Higher Education: A Student Perspective," 2025 5th International Conference on Educational Communications and Technology (ICTAECT), Bangkok, Thailand, 2025, pp. 1-6, doi: 10.1109/ICTAECT67351.2025.11391308.

Generative AI in Higher Education: A Student Perspective

- Pros and Cons of the study and the assessment?

- Can you reproduce this study?
- What can we learn/communicate with this type of study?
- What are limitations of this type of study?

W. Milinthapunya, P. Nilsook and P. Wannapiroon, "Generative AI in Higher Education: A Student Perspective," 2025 5th International Conference on Educational Communications and Technology (ICTAECT), Bangkok, Thailand, 2025, pp. 1-6, doi: 10.1109/ICTAECT67351.2025.11391308.

Generative AI in Programming Education

- Focus: AI in programming - student behaviour from sporadic tool
- assistance to infrastructure dependency
- The “watershed” moment: the widespread adoption of high-performance domestic Chinese Large Language Models (LLMs) such as DeepSeek (in early 2025)
- Study built upon a proprietary dataset gathered directly from the author’s instructional activities across three consecutive academic semesters
- Total students: 412
  - Before the changing event (pre-watershed): 89
  - After the changing event (post-watershed): 323
- Metrics:
  - LLM Usage and Categorization: Distribution of LLM types used (e.g., DeepSeek, ChatGPT, etc), derived from student questionnaires and observational statistics recorded during laboratory sessions
  - Code Submission Accuracy: Mean Score Rate (MSR), expressed as a normalized value between 0.0 and 1.0
  - Temporal Efficiency: Time-to-completion (TtC) for core coding tasks.
  - Examination Performance: Mean Score Rate (MSR) obtained exclusively from the programming specific questions in high-stakes

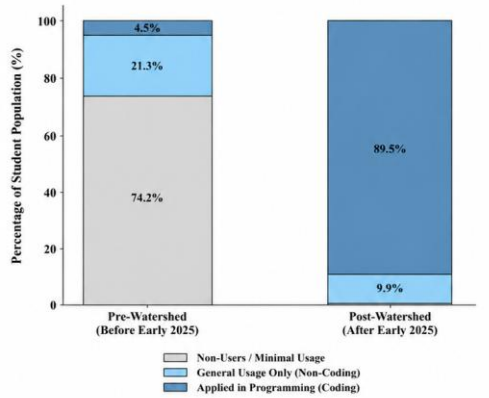


Figure 1: Shift in LLM Adoption Frequency.

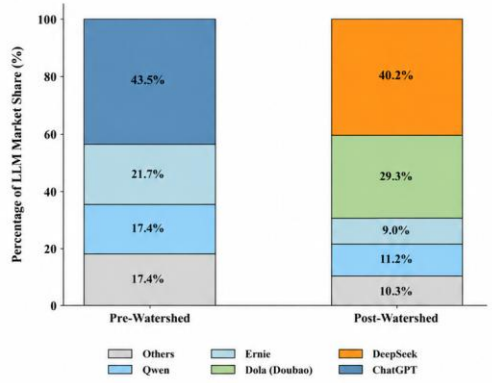


Figure 2: Migration of LLM Tool Preference.

Suyang Zhu and Zhengtao Liu. 2026. Generative AI in Programming Education: An Empirical Analysis of Student Performance and Assessment Challenges in the LLM Era. In Proceedings of the 2026 International Conference on Big Data and Informatization Education (ICBDIE '26). Association for Computing Machinery, New York, NY, USA, 26–32. https://doi.org/10.1145/3806980.3806985

Generative AI in Programming Education

- Pros and Cons of the study and the assessment?
- Can you reproduce this study?
- Course-specific study, course-specific metrics

Suyang Zhu and Zhengtao Liu. 2026. Generative AI in Programming Education: An Empirical Analysis of Student Performance and Assessment Challenges in the LLM Era. In Proceedings of the 2026 International Conference on Big Data and Informatization Education (ICBDIE '26). Association for Computing Machinery, New York, NY, USA, 26–32. https://doi.org/10.1145/3806980.3806985

GenAI Analytics and Pedagogical Configurations

- Focus: which GenAI Analytics and Pedagogical Configurations are more useful for university teachers through a GenAI system developed by the authors.
- 16 university teachers; 5 different institutions
- One-hour face-to-face workshop: participants completed the consent form together with a pre-questionnaire (demographics, use of AI tools in teaching)
  - Net Promoter Score (NPS) - Frederick F Reichheld. 2003. The one number you need to grow. Harvard business review 81, 12 (2003), 46–55
  - Participants received a 15-min introduction to the concept of teacher autonomy with GenAI-powered chatbots
- Each participant was provided with teacher and student credentials to access a university programming course deployed on Canvas LMS where the GenAI chatbot was integrated.
- Participants were asked to pedagogically configure the GenAI tool, test it, and understand the different behaviours of the course students through the implemented GenAI Analytics.
- Post-questionnaire:
  - The evaluation of every GenAI Analytic indicator implemented in the GenAI system, rating its usefulness on a 5-point Likert-type item plus an open-ended question to justify the scores given.
  - The evaluation of every Pedagogical Configuration implemented in the GenAI system, rating its usefulness on a 5-point Likert-type item plus an open-ended question to justify the scores given.
  - The potential adoption of the GenAI system within their teaching practice through the NPS, thus allowing to compare the results with the ones given for ChatGPT (or other GenAI chatbots) in the pre-questionnaire.
  - The usability of the GenAI system through the *UMUX-Lite questionnaire*, a seven-scale two-item questionnaire with a high internal reliability, able to predict and compare results obtained with the System Usability Scale (SUS) questionnaire.
    - James R Lewis, Brian S Utesch, and Deborah E Maher. 2013. UMUX-LITE: when there's no time for the SUS. In Proceedings of the SIGCHI conference on human factors in computing systems. 2099–2102. doi:10.1145/2470654.2481287

Alejandro Ortega-Arranz, Paraskevi Topali, Pablo García-Zarza, Miguel L. Bote-Lorenzo, Juan I. Asensio-Pérez, and Inge Molenaar. 2026. GenAI Analytics and Pedagogical Configurations: Featuring New Generative AI Systems in Education. In Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference (LAK '26). Association for Computing Machinery, New York, NY, USA, 779–785. https://doi.org/10.1145/3785022.3785054

GenAI Analytics and Pedagogical Configurations

- All evaluated analytics were positively rated (score 4 or 5) by more than 75% of the participants – Figure 1
- All configuration parameters were positively rated by 50%+ participants – Figure 2

- The implemented GenAI Analytics are low-level analytics (i.e., raw data)
  - as suggested by one participant, it would be interesting to explore higher-level analytics that combine multiple indicators to derive participants' profiles or engagement levels with AI
- The study involved a limited sample of practitioners, all of them within the field of educational technology.
  - it is planned to involve teachers from other disciplines and educational levels to understand whether there are differences in GenAI analytics or pedagogical configurations according to such variables.

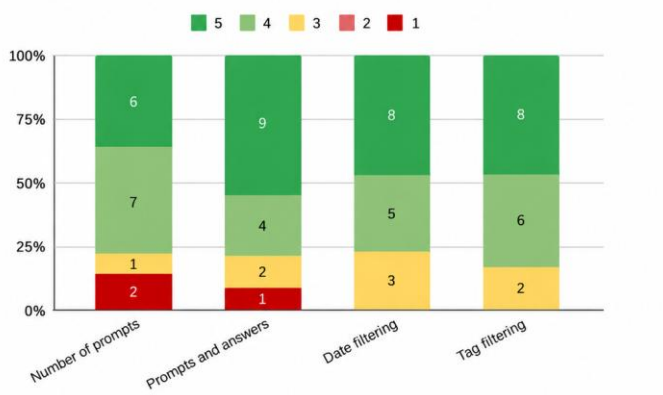


Figure 1: GenAI Analytics

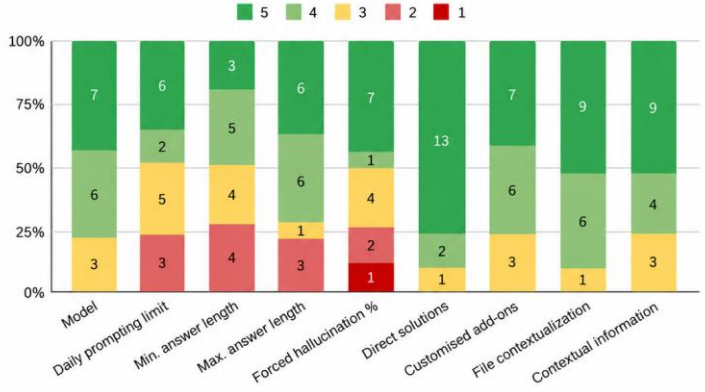


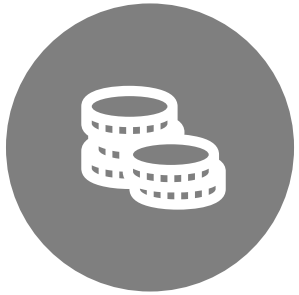
Figure 2: Pedagogical Configurations

Alejandro Ortega-Arranz, Paraskevi Topali, Pablo García-Zarza, Miguel L. Bote-Lorenzo, Juan I. Asensio-Pérez, and Inge Molenaar. 2026. GenAI Analytics and Pedagogical Configurations: Featuring New Generative AI Systems in Education. In Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference (LAK '26). Association for Computing Machinery, New York, NY, USA, 779–785. https://doi.org/10.1145/3785022.3785054

Complement your Work with the Notes

- Goals
- Challenges
- Metrics
- Data collection techniques
- Assessment approaches
- Results presentation
- Results abstraction
- Conclusions
- Things to avoid

GenAI Adoption Approaches



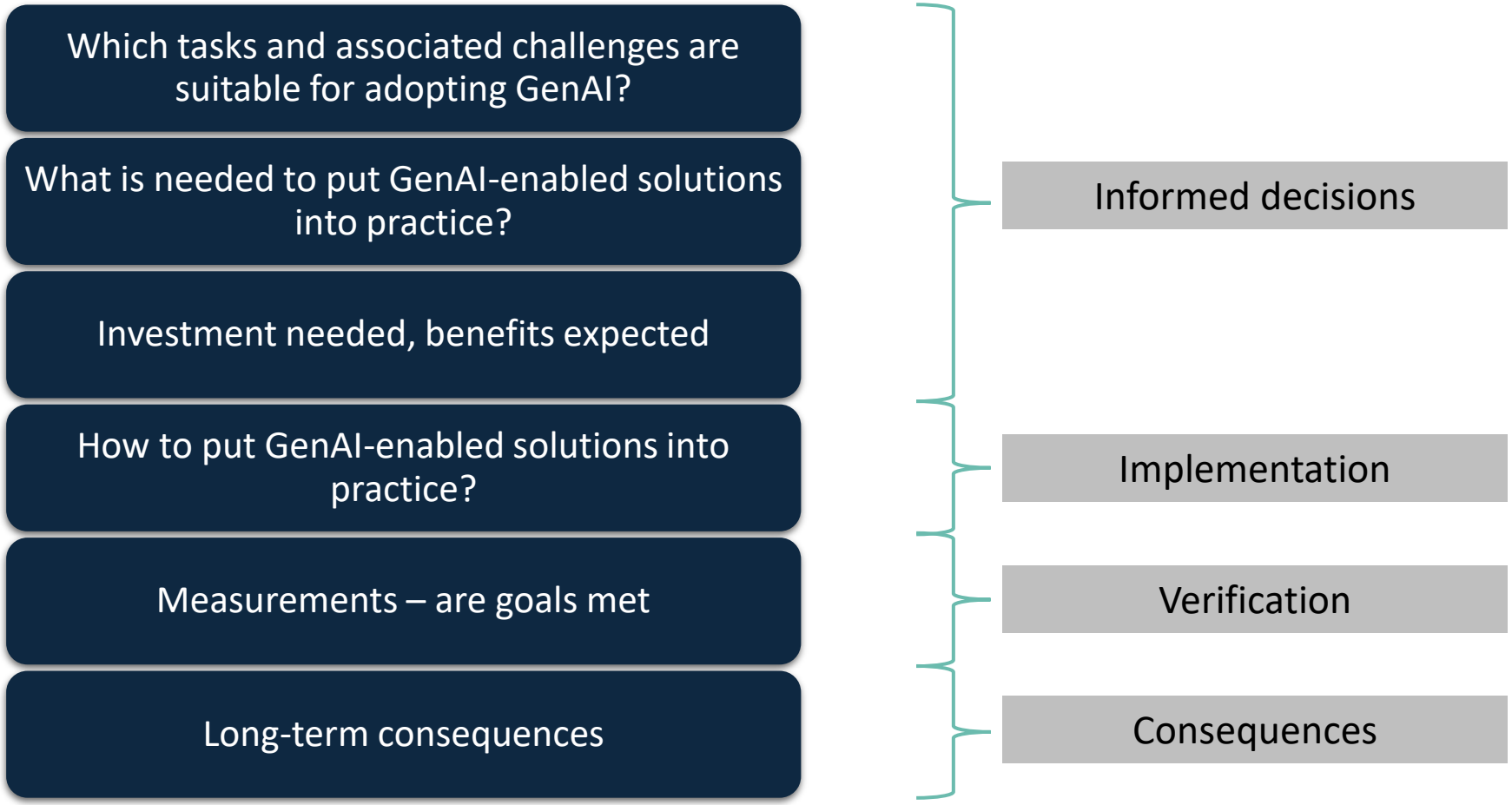
Bottom up (with a motivation from the top): here are the tools, here are the tokens (with or without a limit) – experiment and figure out if we can get some benefit

Top down: we have identified challenges which GenAI can solve, we will translate the solution into strategy and methodology, and make it reproducible. E.g., part of the process.

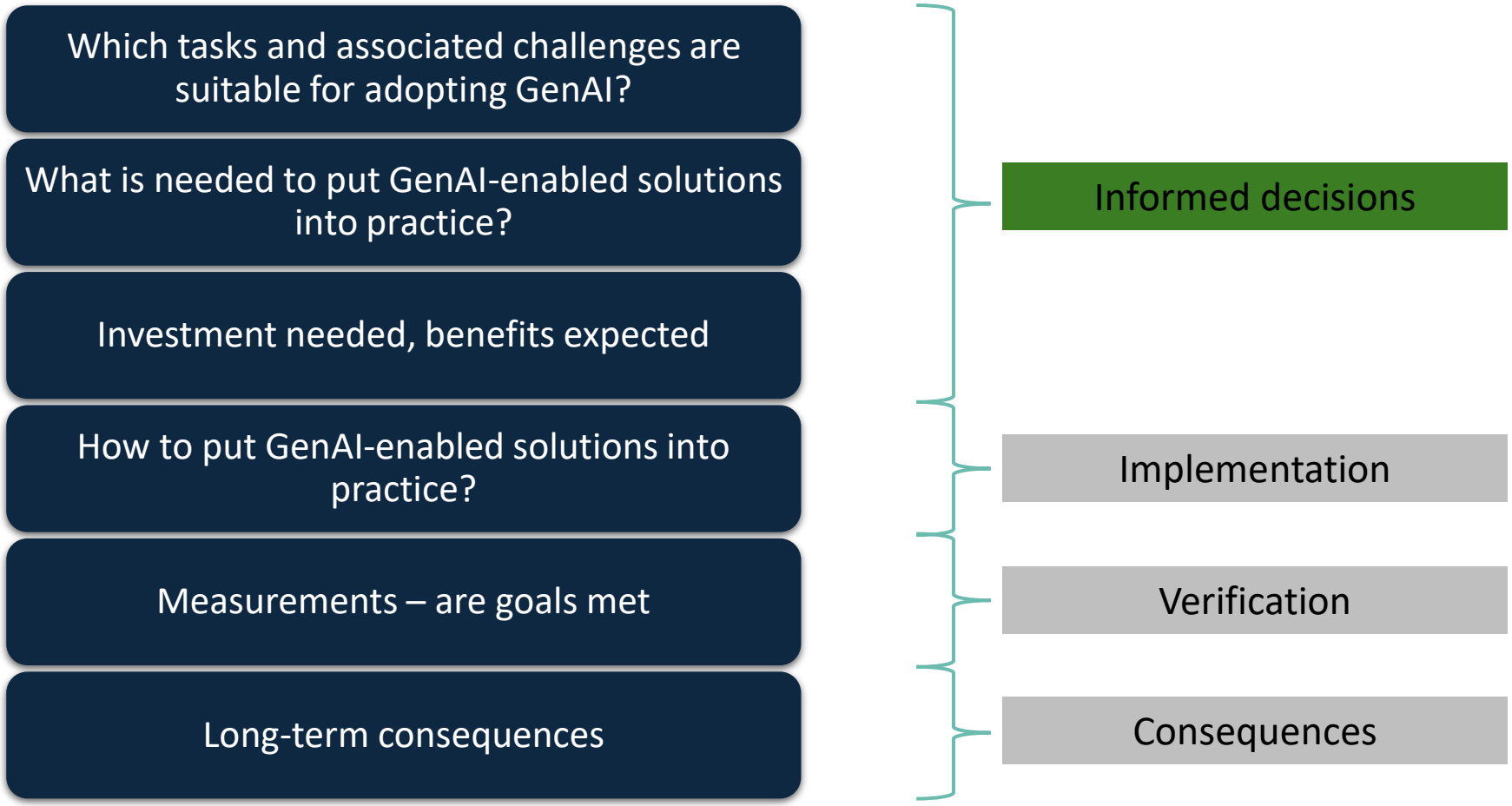
Bottom up: is how we usually understands what works

Top down: is how we usually enforce good practices

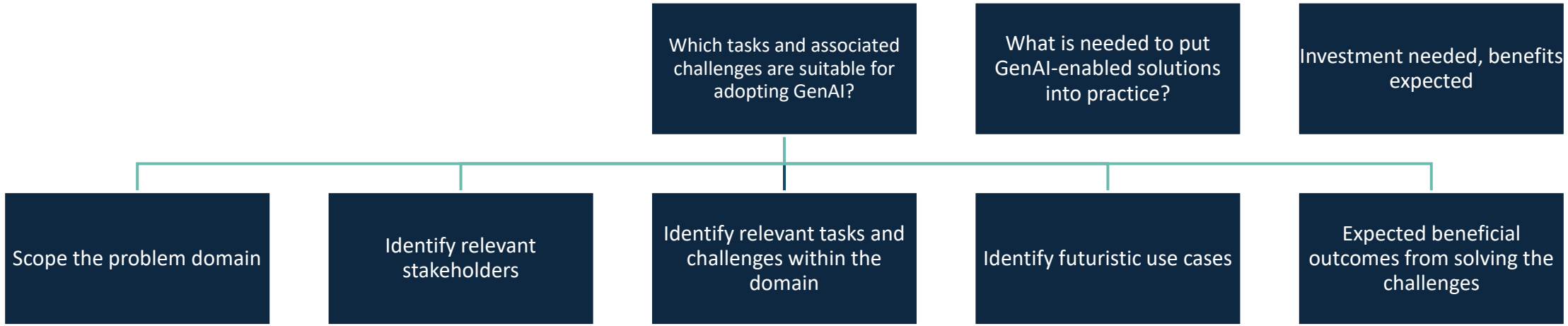
Top Down Approach



Top Down Approach



Informed Decisions



- Task and challenges:
- Background: ...
- Consequences: ...
- Common challenges: ...
- Sources of complexity: ...

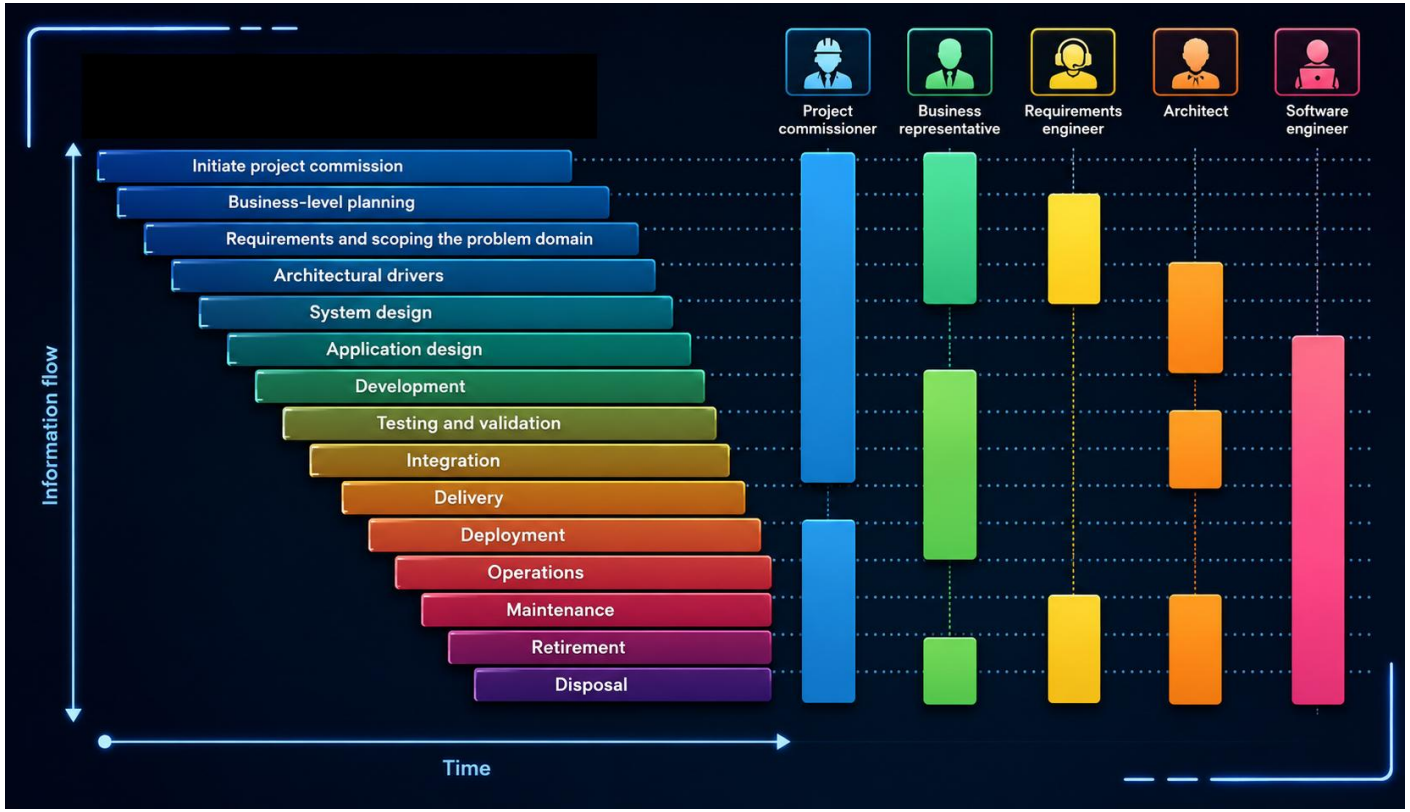
- Input/Output sets
  - Input
  - Action
  - Output (beneficial outcome)

Informed Decisions – SDLC Problem Scoping Example

Which tasks and associated challenges are suitable for adopting GenAI?

What is needed to put GenAI-enabled solutions into practice?

Investment needed, benefits expected



Informed Decisions – SDLC Problem Scoping Example

Which tasks and associated challenges are suitable for adopting GenAI?

What is needed to put GenAI-enabled solutions into practice?

Investment needed, benefits expected

Too many meetings and meeting notes: misunderstandings lead to expensive re-engineering and lower quality of software reducing it value, data analysis is expensive.

- **Background**: Social media and marketing/business expert has many meetings and takes many notes. It is hard to find answers important to engineering in those notes or analyze the data manually.
- **Consequences**: Misunderstandings are frequent, leading to the need to re-engineer solutions (expensive) or to software with unsatisfactory quality (e.g., performance).
- **Common challenges**: Overwhelming complexity, scale and size of the problem, lack of technical support in terms of tools that can approximate and summarize the notes. Manual analysis of data.
- **Sources of complexity**: Software engineering is a collaborative effort, but humans (nature of our perception) tend to forget things or reinterpret them when trying to communicate, leading to misunderstandings or misinterpretation. In general, humans are limited by the speed we can understand complexity. Too many skills needed - lack of technical skills to create scripts for data analysis.

IO set #001

- Inputs: Meeting notes
- Action: Create abstracted meeting notes
- Outputs: Summarised meeting notes suitable for understanding by different stakeholders in the SDLC

IO set #002

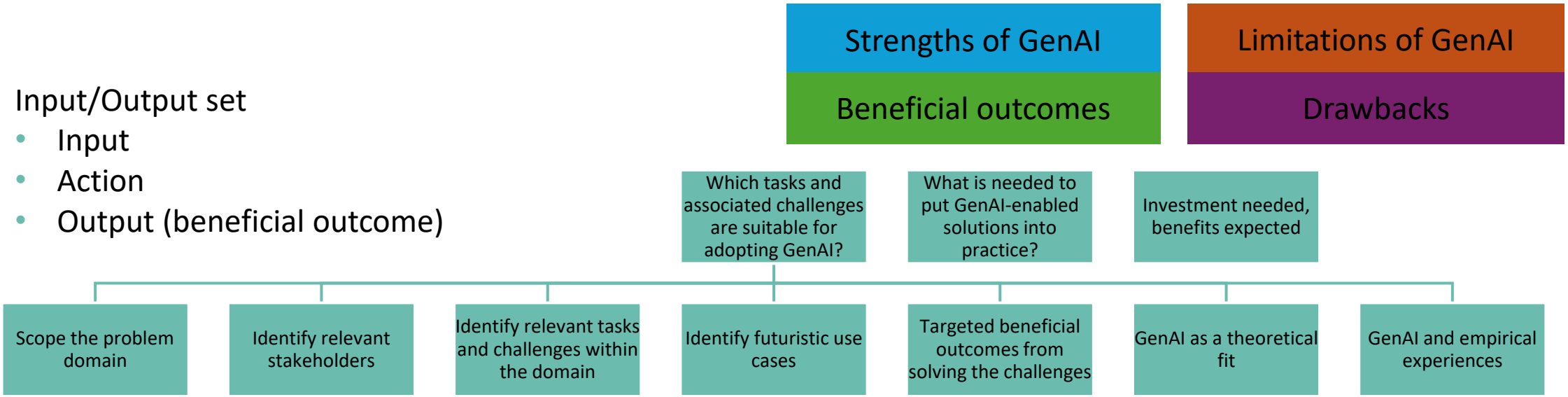
- Inputs: Data, description of algorithmic transformation
- Action: Create scripts that take input data, make analysis, and generate visualized representation
- Outputs: Python scripts

Informed Decisions

Celal Ziftci, Stoyan Nikolov, Anna Sjövall, Bo Kim, Daniele Codecasa, and Max Kim. 2025. Migrating Code At Scale With LLMs At Google. In Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering (FSE Companion '25). Association for Computing Machinery, New York, NY, USA, 162–173. https://doi.org/10.1145/3696630.3728542

Input/Output set

- Input
- Action
- Output (beneficial outcome)



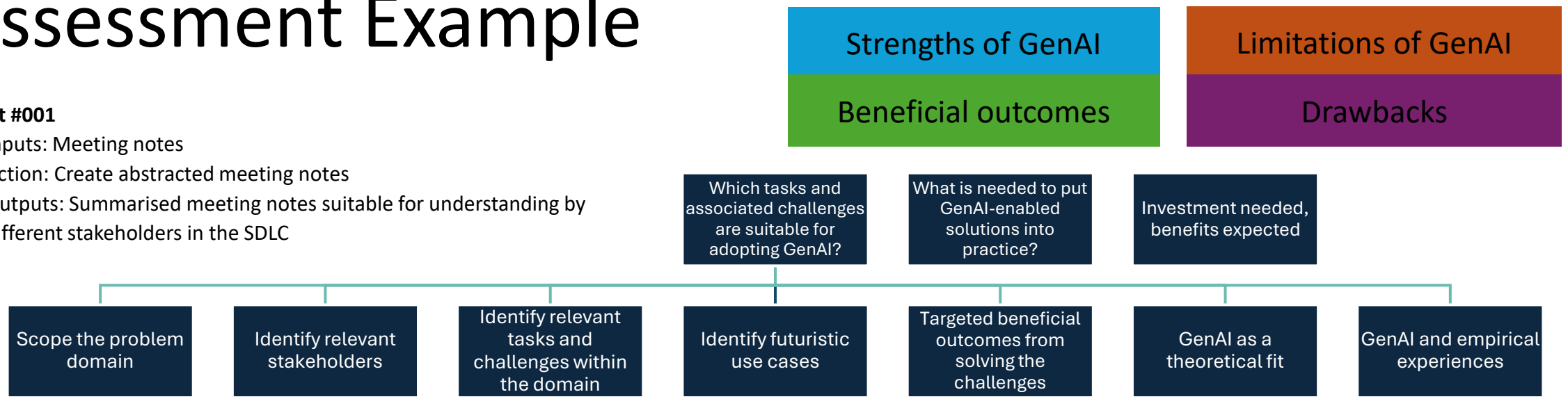
Points in favour of adopting GenAI as an adequate solution in this context, experienced GenAI strengths and beneficial outcomes	Assumptions made for selecting GenAI, risks of the adoption, experienced drawbacks and limitations
What are the factors that will influence the decision? What can we expect to significantly increase?	As GenAI solves one problem, does it create others?

Decision template adapted from Jens Knodel and Matthias Naab. 2016. Pragmatic Evaluation of Software Architectures (1st. ed.). Springer Publishing Company, Incorporated.

Informed Decisions – GenAI Assessment Example

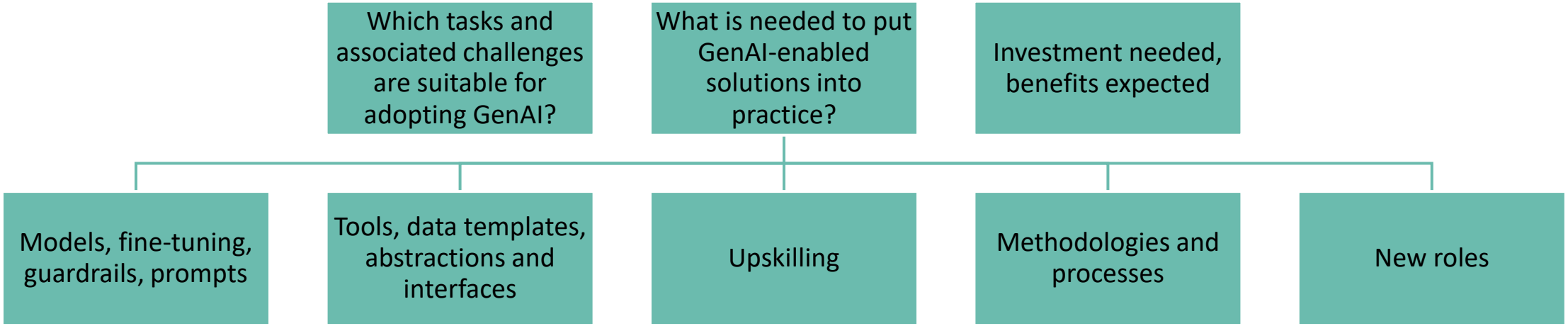
IO set #001

- Inputs: Meeting notes
- Action: Create abstracted meeting notes
- Outputs: Summarised meeting notes suitable for understanding by different stakeholders in the SDLC



GenAI tools show high correctness, completeness, and readability when generating UML diagrams. Hence, they can automate visualisation of notes. GenAI tools demonstrated the ability to use well-structured visual language (UML) to generate diagrams contributing to better understandability between stakeholders and increasing the throughput of the whole organisation. To reduce the effort, GenAI tools can keep updating the diagrams automatically.	There are challenges with hallucinations, misinterpretation, and theoretical limitations of GenAI due to the lack of contextual knowledge. Generated diagrams need to be evaluated manually. Wrong information in diagrams can cause confusion and introduce more overhead. The beneficial outcomes in terms of reducing the time are not quantified in the study, so we can only assume them.
Size and complexity of notes, number of UML diagrams that a human needs to manually evaluate, the level of involvement in each of the diagrams, in case there is a need to change something.	Improve understanding between stakeholders **vs** the quality of the generated UML diagrams and drawbacks (e.g., confusion) that the low quality of the diagrams can introduce. Time saved by automatically creating and updating the diagrams **vs** time needed to manually evaluate and fix the diagrams

Informed Decisions



GenAI-related tasks - Finding, configuring and tuning models, and using chatbots and agents.

GenAI assistance - Using GenAI-enabled tools, toolchains and methodologies.

GenAI methodologies and toolchains - Integrating tools with GenAI and GenAI methodology with software development methodologies.

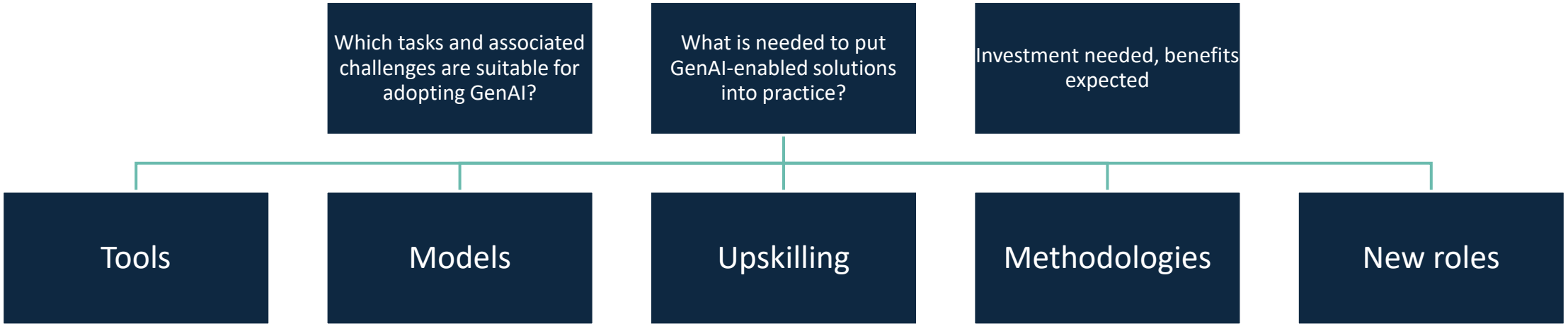
Stakeholders need to perform new **GenAI-related tasks**.

Due to GenAI assistance, the **effort** is **redistributed** (for example, when effort is moved from one stakeholder to another).

There is a **methodology** in place that helps stakeholders, but someone has to create and maintain the methodology.

Do you have relevant skills in your organisation?

Informed Decisions

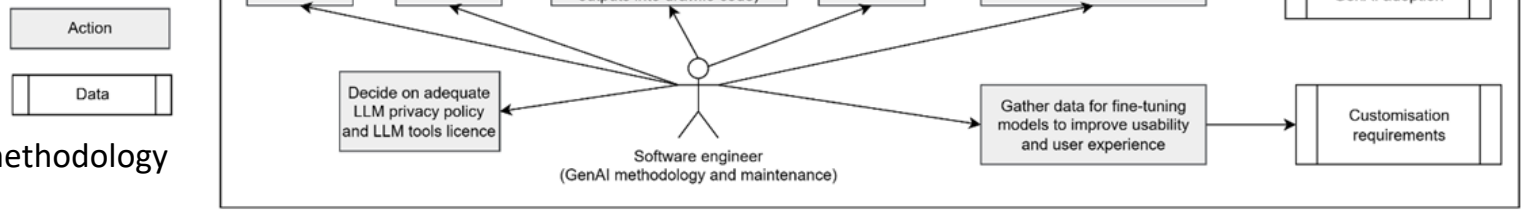


	Assumptions and risks (constraints)
Points in favour of adopting GenAI as an adequate solution in this context, experienced GenAI strengths and beneficial outcomes	Assumptions made for selecting GenAI, risks of the adoption, experienced drawbacks and limitations
What are the factors that will influence the decision? What can we expect to significantly increase?	As GenAI solves one problem, does it create others?

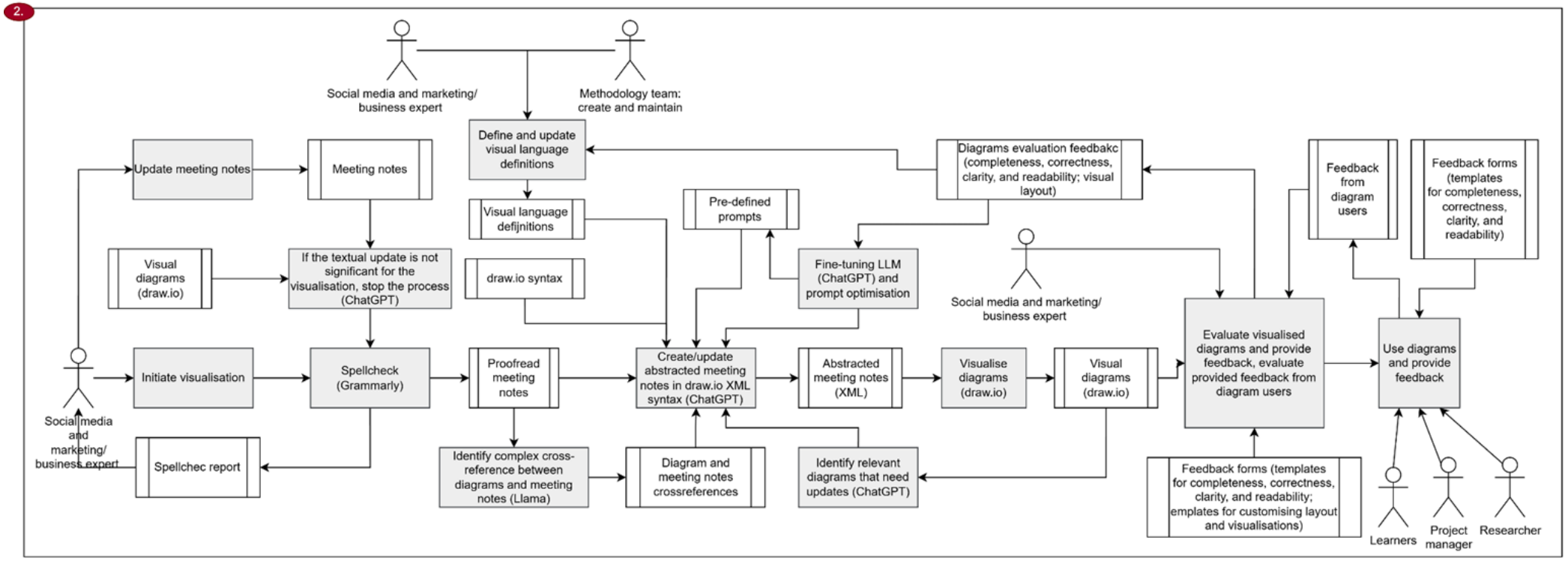
Decision template adapted from Jens Knodel and Matthias Naab. 2016. Pragmatic Evaluation of Software Architectures (1st. ed.). Springer Publishing Company, Incorporated.

Informed Decisions – Methodology Example

1. Create and maintain LLM infrastructure.



2. GenAI-enabled methodology



Informed Decisions – Methodology Example

Which tasks and associated challenges are suitable for adopting GenAI?

What is needed to put GenAI-enabled solutions into practice?

Investment needed, benefits expected

Assumptions and risks (constraints)

…	… All communication between social media and marketing/business expert and other stakeholders will be supported by visualised, summarised, and abstracted previous notes. Social media and marketing/business do not use raw LLMs. Instead, there will be a toolchain in place. One software engineer - methodology team - will create and maintain LLM infrastructure (choose and configure). Methodology team will define a modelling profile in draw.io based on the defined visual language. They will use the defined modelling language to guide creation of prompts (define prompt templates) and use draw.io to visualise generated diagrams (these are generated in XML using draw.io syntax). Syntax and grammar check – use Grammarly. There is a similar tool powered by GenAI: https://notebooklm.google/, so this approach is feasible. Training in prompting, knowledge of: structured prompt engineering (Marvin et al, 2023), prompt engineering best practices (Arvidsson and Axell, 2023; OpenAI Developers), and, prompt patterns (White et al, 2023) are also essential. Methodology team will need to acquire skills necessary for the LLM-related tasks (crafting prompts, fine-tuning LLMs), creation of workflows, and creation of templates. Social media and marketing/business expert needs to evaluate visualised notes, hence they need abstraction and visualisation skills. A course on abstractions in software engineering (e.g., software architecture) would help here.
…	…

Business Aspects

Business aspects :

- cost reduction
- productivity gains
- revenue uplift
- time-to-market reduction
- risk avoidance (expected loss reduction)
- develop technology acumen

Costs

- initial investments (one-off)
- operational investments (recurring)

Investment

Initial investments (one-off or upfront)

- Technology: software licences, infrastructure such as cloud setup, integration work and data preparation.
- People: training and upskilling, external consultants and initial hiring.
- Process and governance: security and compliance setup, privacy policy and architecture design.
- Change enablement: pilots and proofs of concept.

Operational investments (ongoing costs)

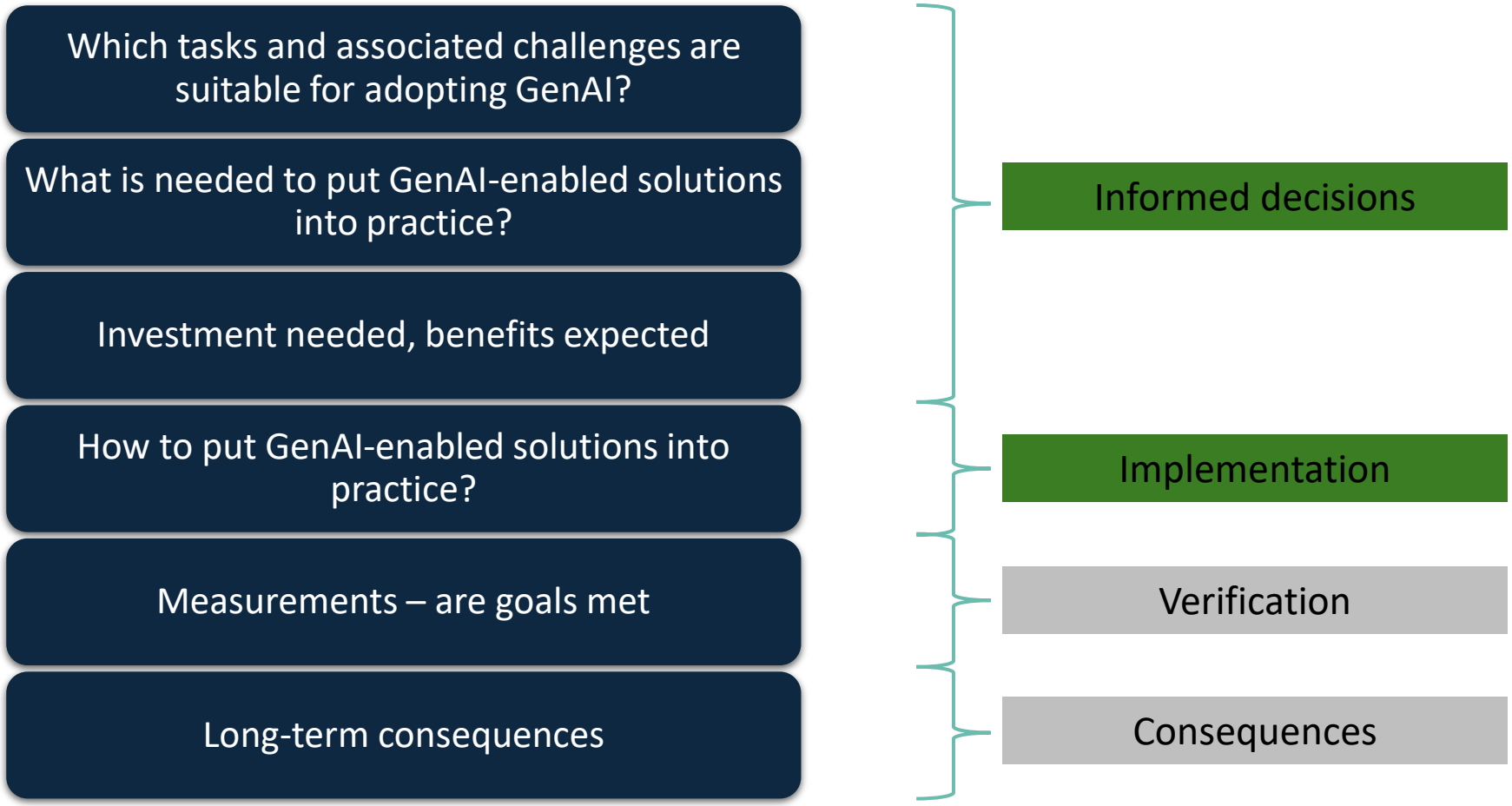
- Run costs: infrastructure, tokens and licences.
- People: support staff, model or system maintenance and governance roles.
- Risk: legal and regulatory overhead hindering beneficial outcomes.
- Evolution: upgrades, retraining or re-prompting, vendor changes, technical debt, organisation changes and software architecture changes.

Include costs into decision-making process

Decision-making Template

Points in favour of adopting GenAI as an adequate solution in this context, experienced GenAI strengths and beneficial outcomes	Assumptions made for selecting GenAI, risks of the adoption, experienced drawbacks and limitations
What are the factors that will influence the decision? What can we expect to significantly increase?	As GenAI solves one problem, does it create others?

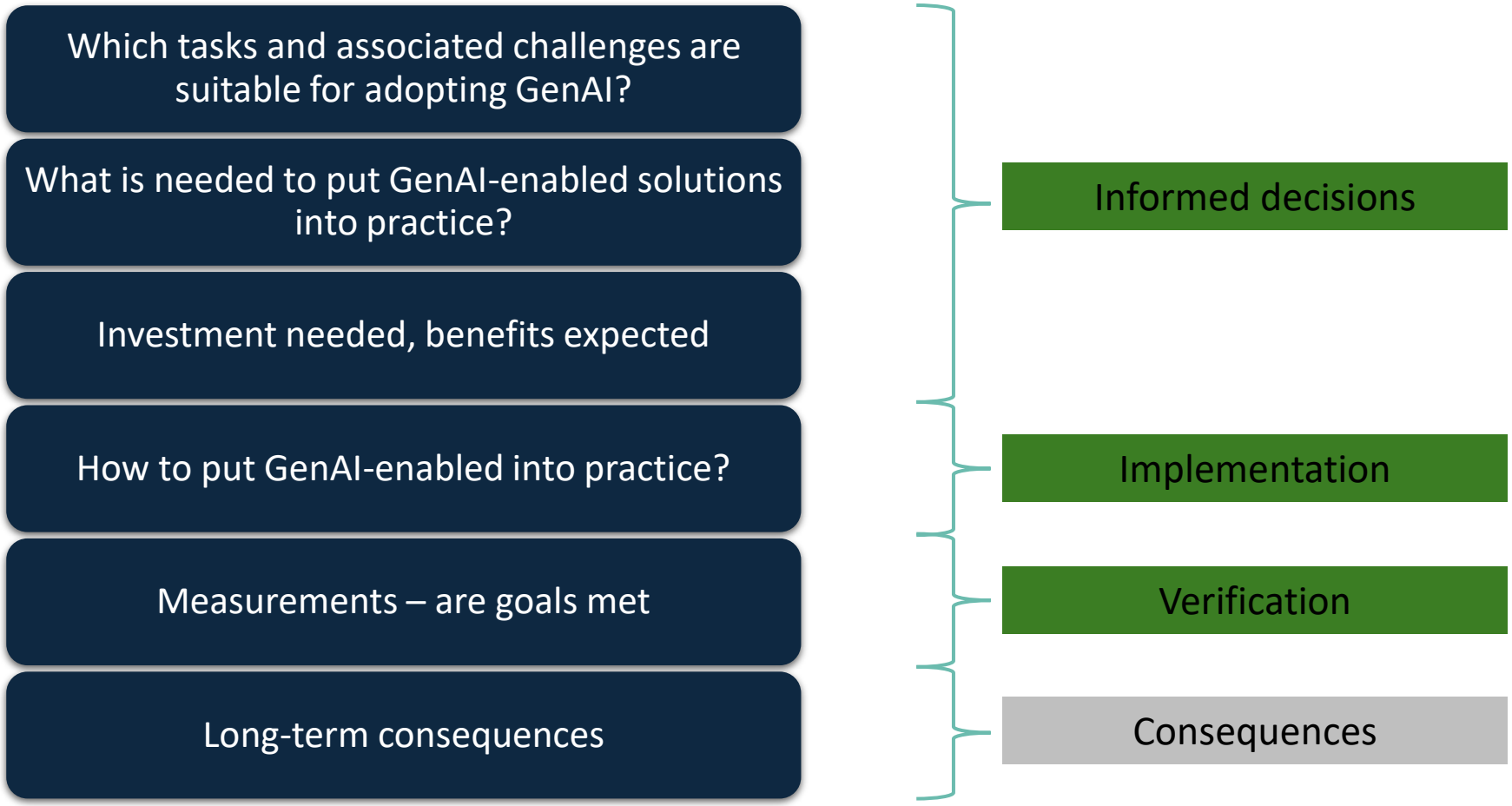
Top Down Approach



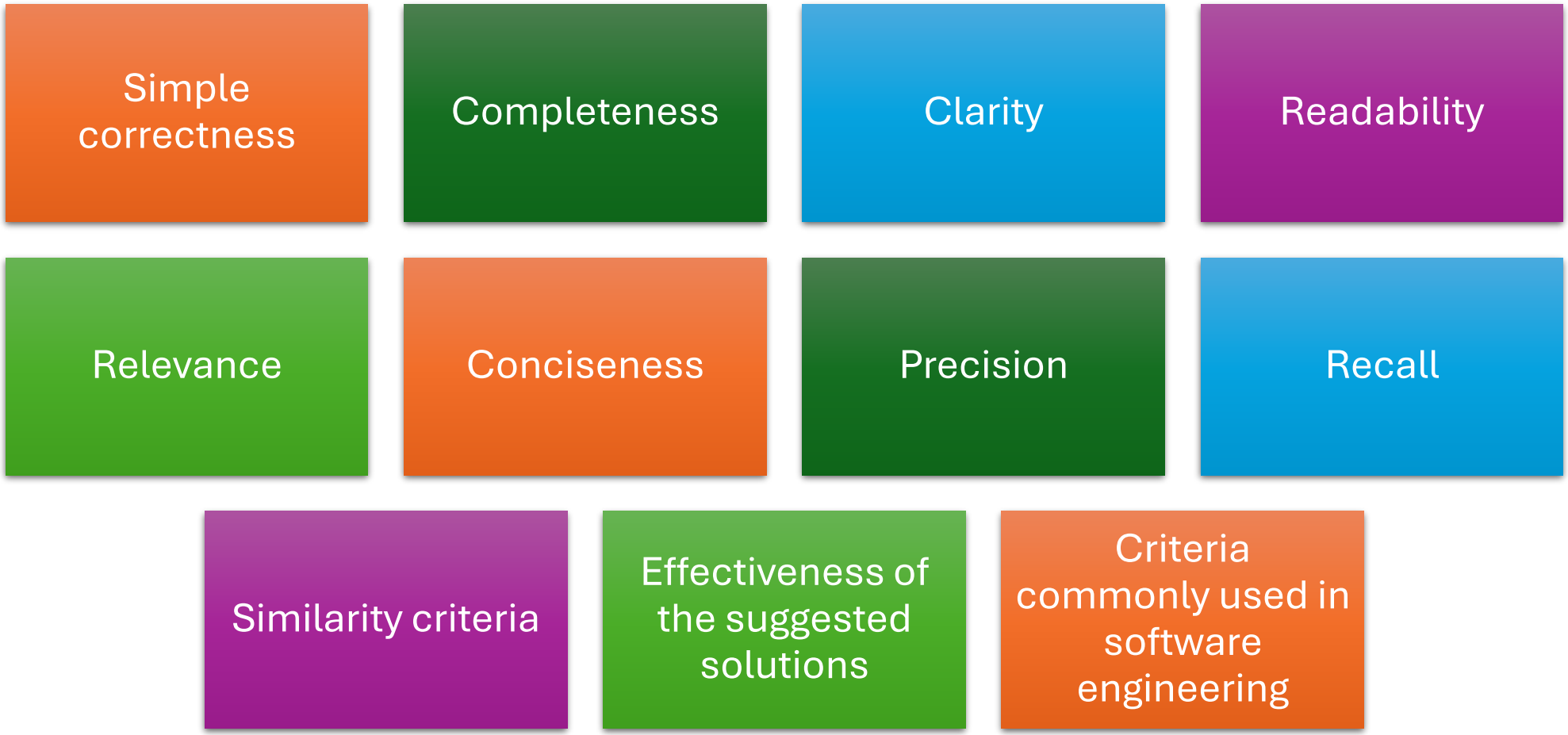
Implementation - How to put GenAI-enabled into Practice?

- Unlimited access
- Standardised access (prompt templates, data templates)
- Hidden behind an interface

Top Down Approach

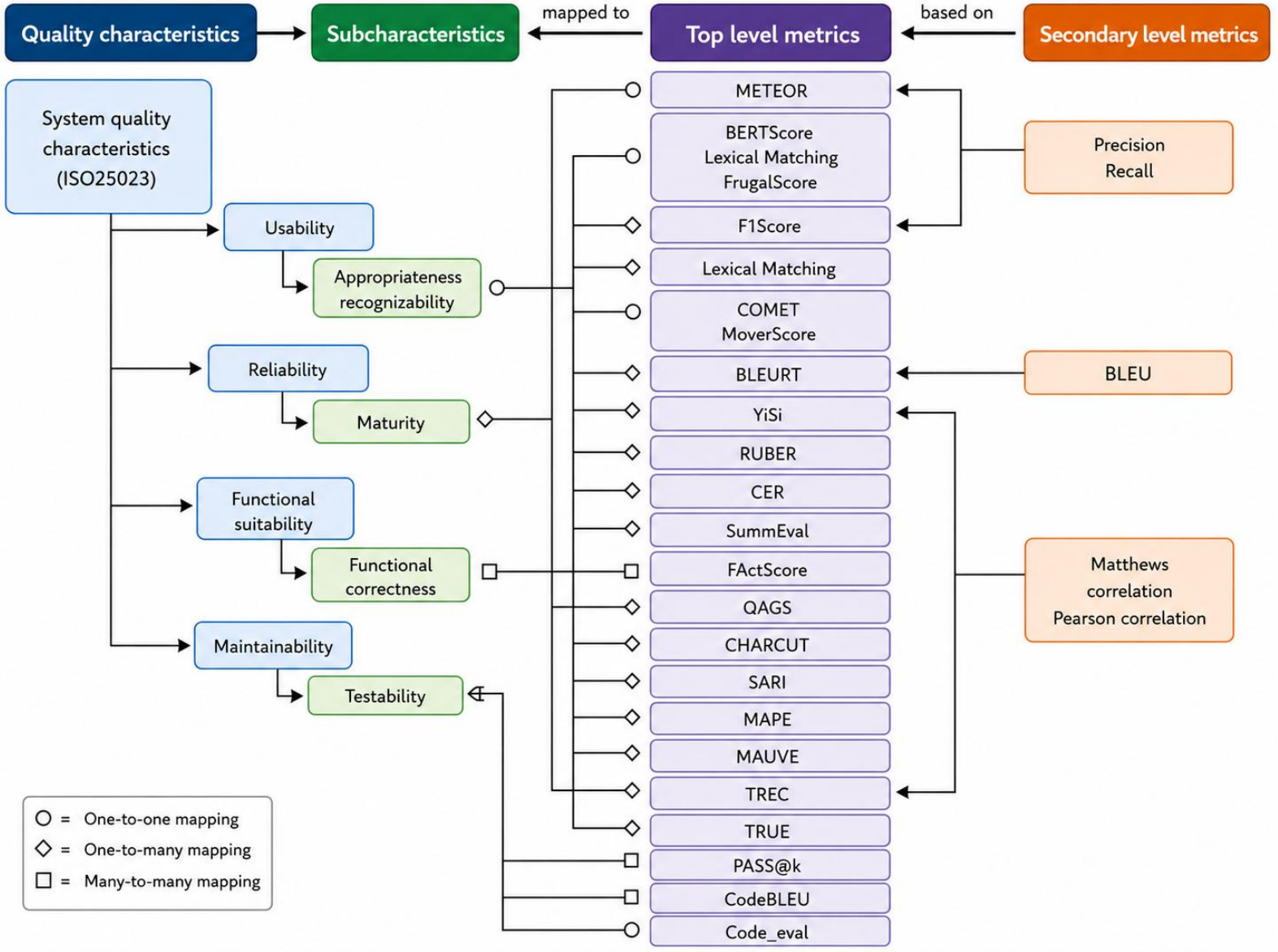


Measuring the Impact of AI Adoption

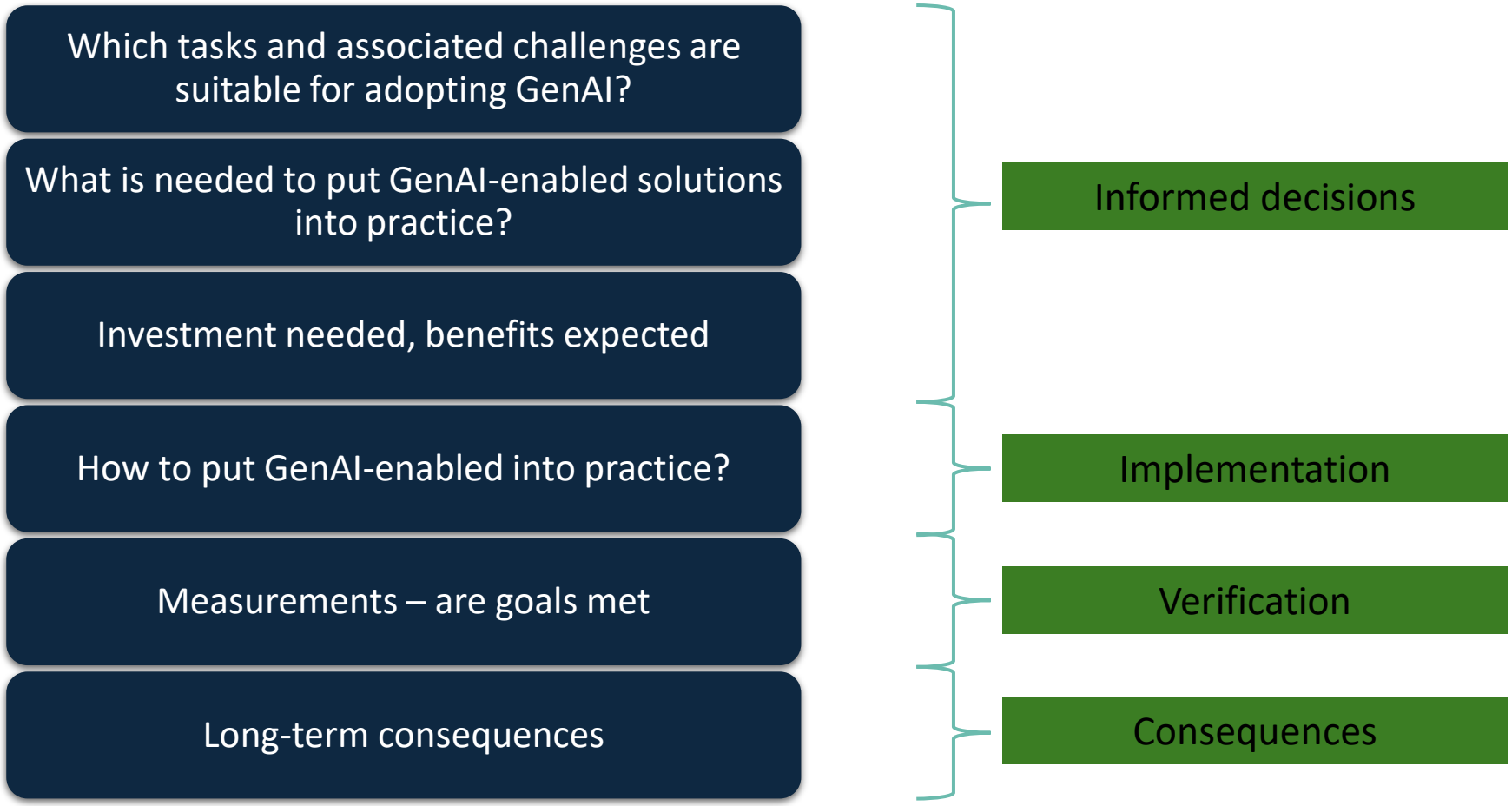


Mapping the identified Metrics to ISO/IEC 25023 Standard for Software Systems

Liang Yu, Emil Alégroth, Panagiota Chatzipetrou, Tony Gorschek, Measuring the quality of generative AI systems: Mapping metrics to quality characteristics — Snowballing literature review, Information and Software Technology, Volume 186, 2025 - https://www.sciencedirect.com/science/article/pii/S0950584925001417

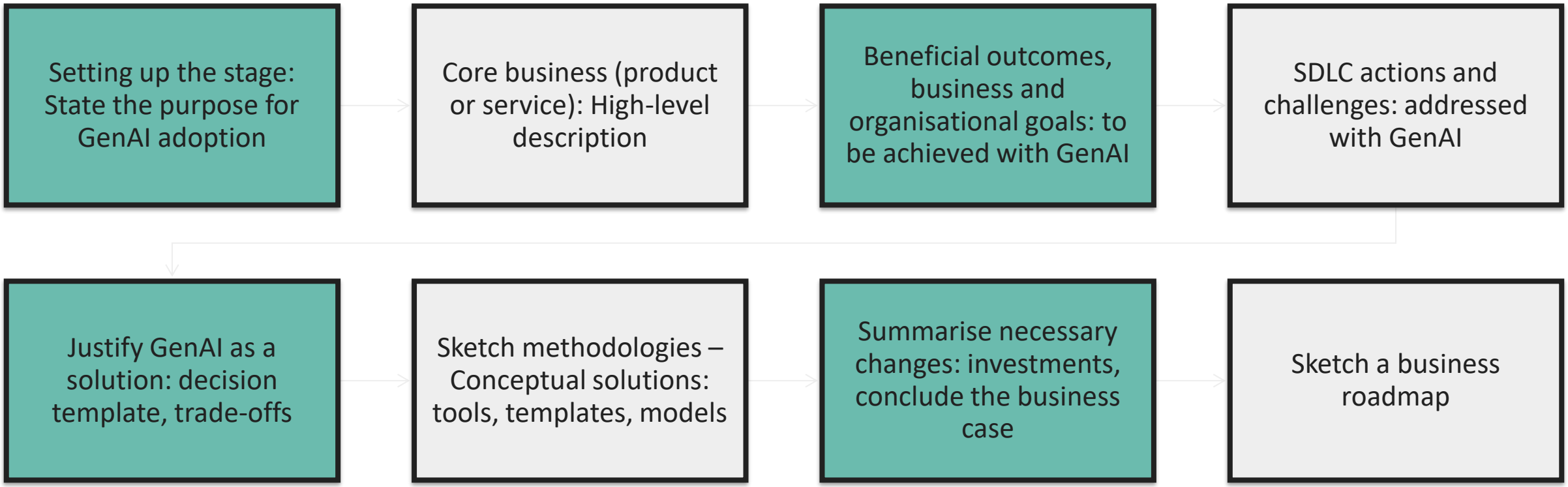


Top Down Approach



Long-term Effects

Summary


Setting up the stage: State the purpose for GenAI adoption
Core business (product or service): High-level description
Beneficial outcomes, business and organisational goals: to be achieved with GenAI
SDLC actions and challenges: addressed with GenAI
Justify GenAI as a solution: decision template, trade-offs
Sketch methodologies – Conceptual solutions: tools, templates, models
Summarise necessary changes: investments, conclude the business case
Sketch a business roadmap

Summary

Problem domain: GenAI cannot tell us what we want, can help to shape the problem domain

- E.g., requirements engineering

Solution: GenAI cannot tell us what we want to achieve, can help with suggestions

- E.g., software architecture design

Too many assumptions make solution inadequate

Relevant Conferences and Journals

- International Conference on Artificial Intelligence in Education - https://iaied.org/
- International Conference on Educational Data Mining - https://educationaldatamining.org/conferences/
- ACM Learning @ Scale - https://learningatscale.hosting.acm.org/
- International Journal of Artificial Intelligence in Education - https://www.sciencedirect.com/journal/international-journal-of-artificial-intelligence-in-education
- IEEE Transactions on Learning Technologies (TLT) - https://ieee-edusociety.org/publication/about-publications/tlt

Staying in Contact

- jj542@cam.ac.uk
  - Mailing list: send me an email

- LinkedIn: https://www.linkedin.com/in/jasmin-jahi%C4%87-a3aaa833/

- Generative AI in the Software Development Lifecycle - https://advanceonline.cam.ac.uk/courses/generative-ai-in-the-software-development-lifecycle
- AI-Augmented Software Engineering - https://www.cl.cam.ac.uk/teaching/2627/L174/

Questions