https://github.com/margaretstorey/icse-MMR-TB-2025

The Methodological Implications of Using Generative AI in Software Engineering Research

WSESE 2025 Workshop, ICSE Ottawa

Margaret-Anne (Peggy) Storey

University of Victoria

Geir Hansen
SINTEF
Norway

Fabio Calefato
University of Bari
Italy

Kelly Blincoe
University of Auckland
New Zealand

Marcos Kalinowski
PUC-Rio
Brazil

Mauro Pezze
Università della
Svizzera Italiana, Italy

Paolo Tell
IT University of
Copenhagen, Denmark

Tom Zimmermann
Univ. of California
Irvine, USA

Bianca
Trinkenreich
Colorado State University,
USA

Who-What-How of SE Research

Generative AI in Science

GenAI use in SE Research

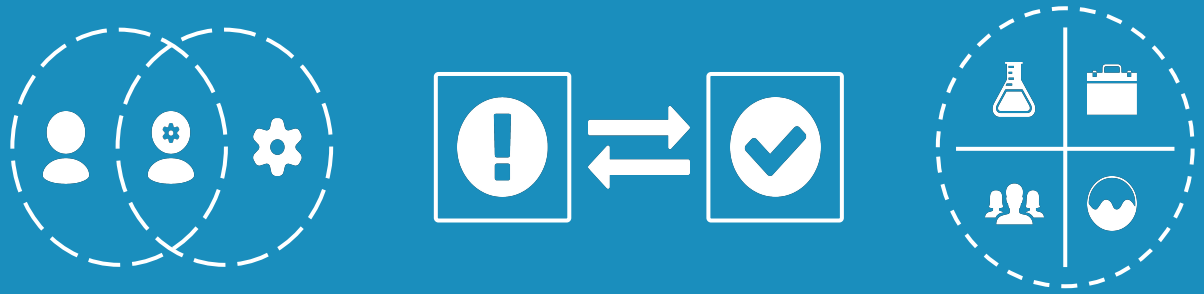
Implications?

Who-What-How of SE Research

Generative AI in Science

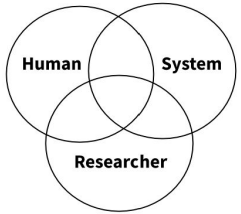
GenAI use in SE Research

Implications?



The Who, What, How of Software Engineering Research

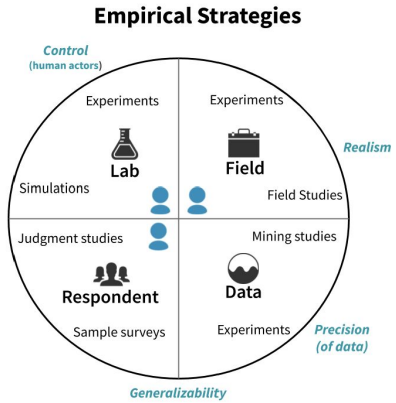
Who?
(is the main beneficiary)



What?
(type of research contribution)

Descriptive | Solution

How?
(which research strategies are used)

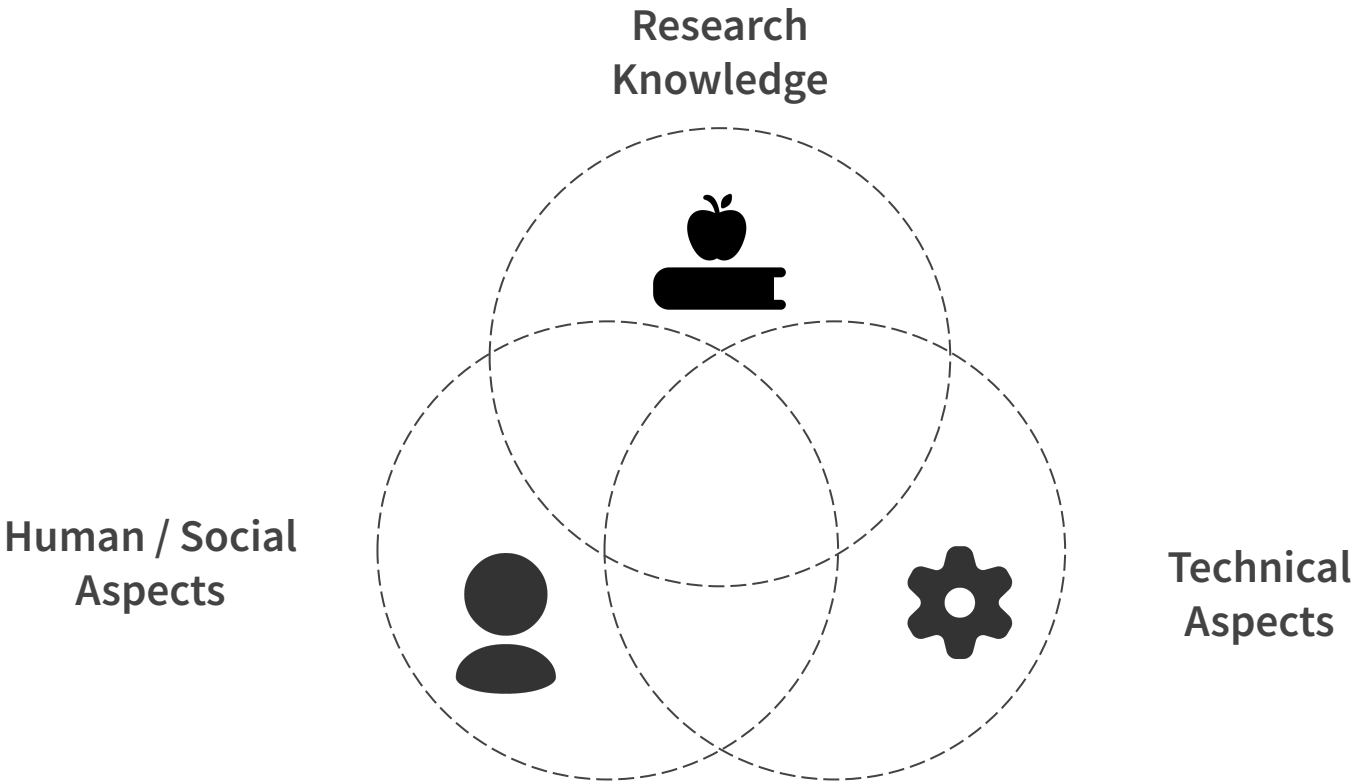


Non-Empirical Strategies

Formal Theory | Meta

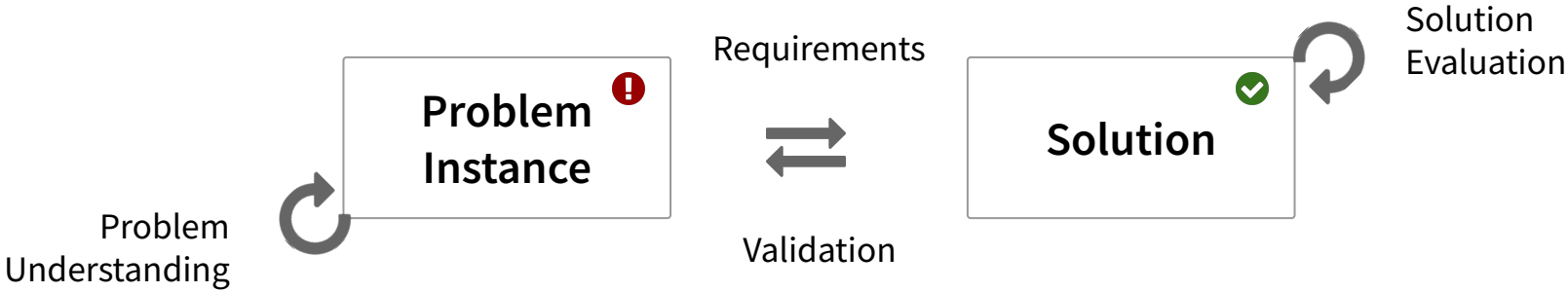
https://bit.ly/wwh-framework

Storey, M., Ernst, N.A., Williams, C. et al. The who, what, how of software engineering research: a socio-technical framework. Empir Software Eng 25, 4097–4129 (2020).



6 | **Who** is the claimed beneficiary of our research?

https://bit.ly/wwh-framework



 | **What** are we studying?

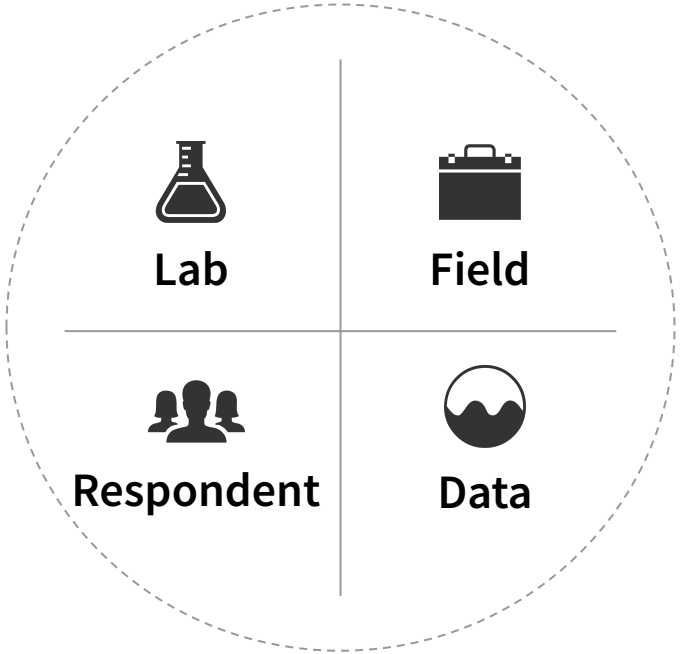
Engström, E., Storey, MA., Runeson, P., Host M., Baldasserra T., How software engineering research aligns with design science: a review. Empirical Software Eng 25, 2630–2660 (2020).

Non-Empirical

Empirical

Formal Theory

Meta



 | **How** we conduct our study (adapting McGrath)

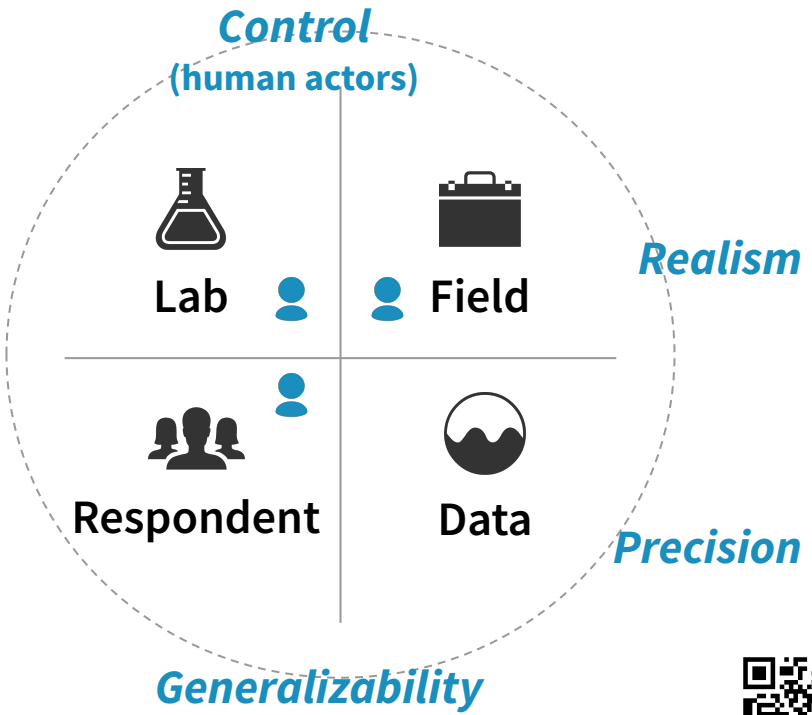
https://bit.ly/wwh-framework

Non-Empirical

Formal Theory

Meta

Empirical

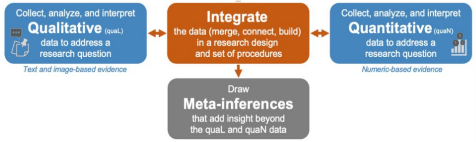


 | **How** we conduct our study (adapting McGrath)

https://bit.ly/wwh-framework

Mixed Methods Research Definition

Mixed methods research (**MMR**) is a research approach where multiple methods are used to collect, analyze, and **integrate both qualitative** and **quantitative** data to address a research problem and **produce novel insights**.



Landscape of MMR Design

Design Properties
- ❑ Research Questions
- ❑ Planned or Emergent
- ❑ Inductive or Deductive Dominance
- ❑ Timing (Sequential or Concurrent)

Integration Strategies
- ❑ Sequential integration
- ❑ Results-based integration
- ❑ Data-based integration
- ❑ Transformation-based integration

Research Designs
- ❑ Exploratory Sequential
- ❑ Explanatory Sequential
- ❑ Convergent Parallel
- ❑ Embedded

Principles to Guide MMR in SE

1 Methodological Rationale

Why did we select a particular MMR research design and set of procedures?
- ❑ Complementarity
- ❑ Expansion
- ❑ Development
- ❑ Triangulation
- ❑ Credibility
- ❑ Explanation
- ❑ Increased design flexibility

2 Novel Integrated Insights

What did we gain from using MMR?
- ❑ Improved problem understanding
- ❑ Greater depth and breadth
- ❑ Explaining unexpected results
- ❑ Complementary storytelling

3 Procedural Rigor

How well did we conduct the study?
- ❑ Is the use of mixed methods justified?
- ❑ Are the methods effectively integrated to answer the RQs?
- ❑ Are the findings from the mixed methods integrated?
- ❑ Are the different methods used rigorously conducted?

4 Ethical Research

How responsibly did we do it?
- ❑ Considering the why
- ❑ Privacy and confidentiality
- ❑ Respect and cultural sensitivity
- ❑ Safety and welfare

Antipatterns of MMR Designs

Presentation Antipatterns
- ❑ Uninvited guest or party crasher
- ❑ Smoke and mirrors
- ❑ Limitation shirker

Study Design Antipatterns
- ❑ Missing the mark
- ❑ Selling your soul
- ❑ Cargo cult research
- ❑ Sample contamination
- ❑ Lost opportunity
- ❑ Integration failure
- ❑ Questionable ethics

EMSE To Appear

Guiding Principles for Mixed Methods Research in Software Engineering

Margaret-Anne Storey · Rashina Hoda · Alessandra Maciel Paz Milani · Maria Teresa Baldassarre

March 24, 2025

Abstract *Mixed methods research is often used in software engineering, but researchers outside of the social or human sciences often lack experience when using these designs. This paper provides guiding principles and advice on how to design mixed method research, and to encourage the intentional, rigorous, and innovative application of mixed methods in software engineering. It also presents key properties of core mixed method research designs. Through a number of fictitious but recognizable software engineering research scenarios, we showcase how to choose suitable mixed method designs and consider the inevitable trade-offs any design choice leads to. We describe several antipatterns that illustrate what to avoid in mixed method research, and when mixed method research should be considered over other approaches.*

Keywords Mixed methods · Research methods · Methodology · Guiding Principles · Guidelines

Margaret-Anne Storey
University of Victoria
Victoria, BC, Canada
E-mail: mstorey@uvic.ca

Rashina Hoda
Monash University
Melbourne, VIC, Australia
E-mail: rashina.hoda@monash.edu

Alessandra Maciel Paz Milani
University of Victoria
Victoria, BC, Canada
E-mail: amilani@uvic.ca

Maria Teresa Baldassarre
University of Bari
Bari, Italy
E-mail: mariateresa.baldassarre@uniba.it

https://bit.ly/mmrse

Who-What-How of SE Research

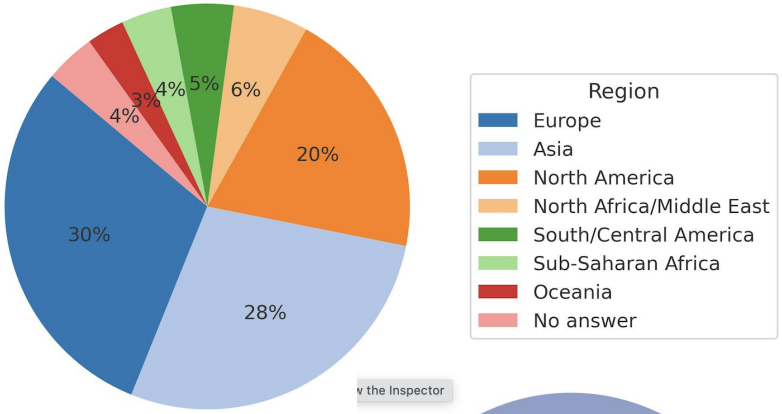
Generative AI in Science

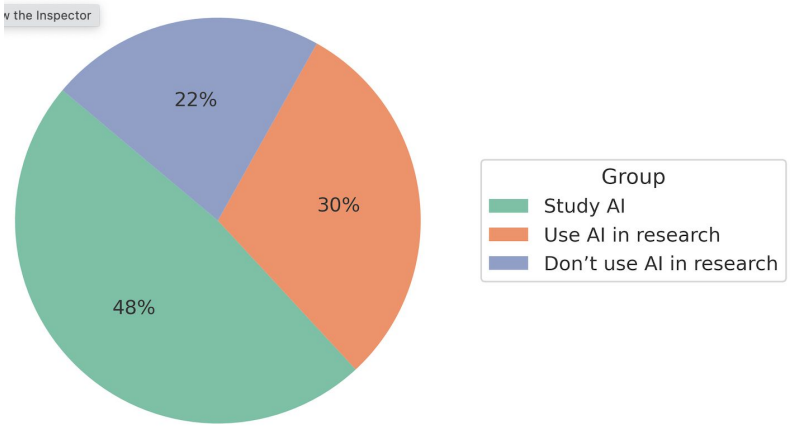
McLuhan Tetrad applied to GenAI use in SE Research

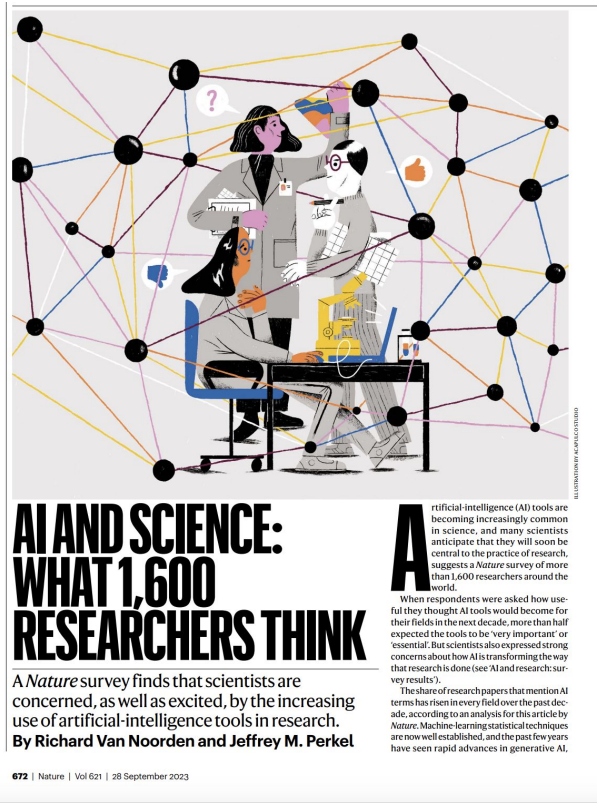
Implications?

Generative AI in Science

AI and Science







AI AND SCIENCE: WHAT 1,600 RESEARCHERS THINK

A *Nature* survey finds that scientists are concerned, as well as excited, by the increasing use of artificial-intelligence tools in research.

By Richard Van Noorden and Jeffrey M. Perkel

Artificial-intelligence (AI) tools are becoming increasingly common in science, and many scientists anticipate that they will soon be central to the practice of research, suggests a *Nature* survey of more than 1,600 researchers around the world.

When respondents were asked how useful they thought AI tools would become for their fields in the next decade, more than half expected the tools to be 'very important' or 'essential'. But scientists also expressed strong concerns about how AI is transforming the way that research is done (see 'AI and research: survey results').

The share of research papers that mention AI terms has risen in every field over the past decade, according to an analysis for this article by *Nature*. Machine-learning statistical techniques are now well established, and the past few years have seen rapid advances in generative AI,

672 | Nature | Vol 621 | 28 September 2023

Van Noorden, R., & Perkel, J. M. (2023). AI and science: what 1,600 researchers think. *Nature*, *621*(7980), 672-675.

"The main problem is that AI is challenging our existing standards for proof and truth"

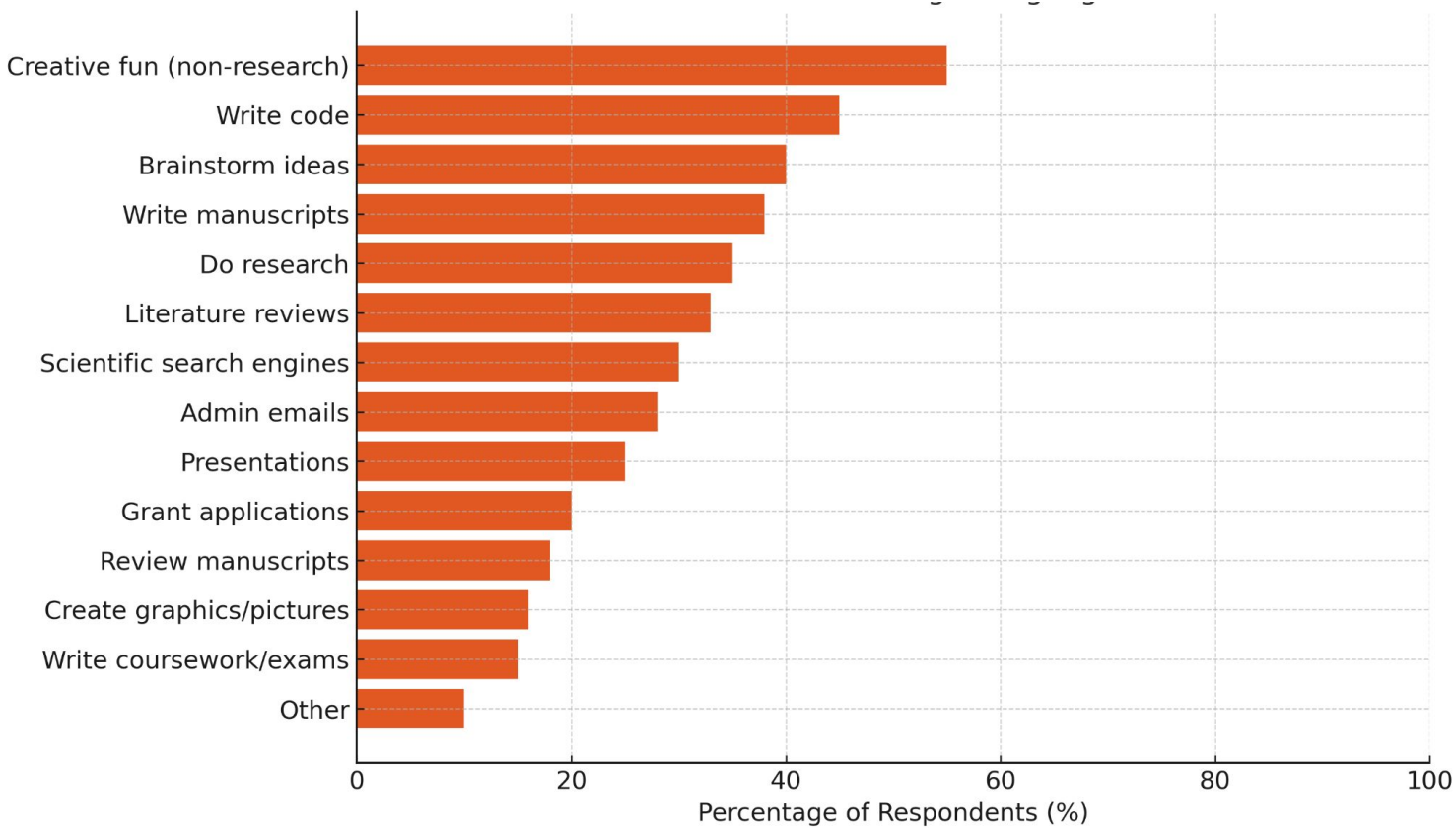
Cancer image analysis researcher at the Jackson Laboratory

"In my opinion, we don't understand well where the border between good use and misuse is."

PhD student in AI for medicine, University of Bristol

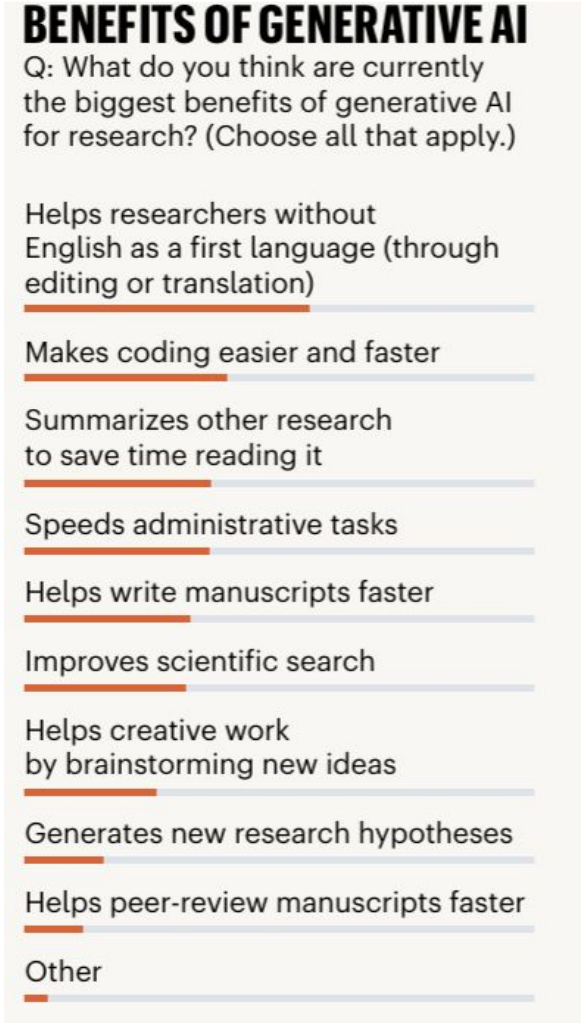
How researchers use **GenAI for science**

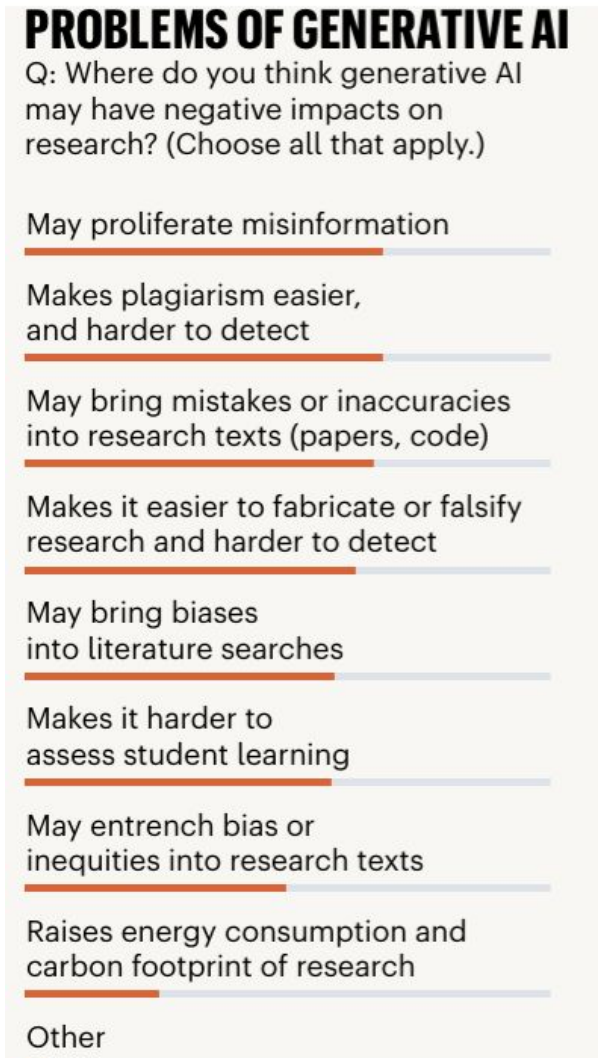
Van Noorden, R., & Perkel, J. M. (2023). AI and science: what 1,600 researchers think. *Nature*, *621*(7980), 672-675.



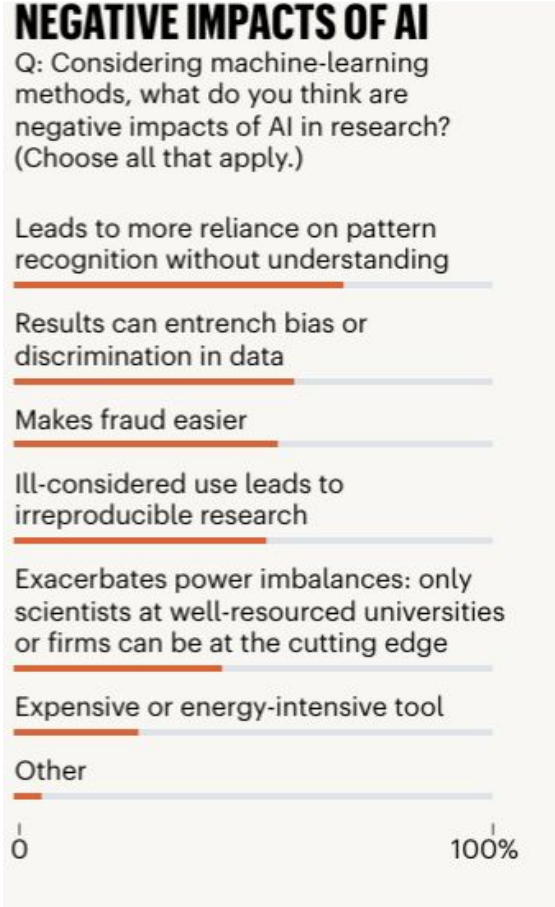
15 | How researchers use **GenAI for science**

Van Noorden, R., & Perkel, J. M. (2023). AI and science: what 1,600 researchers think. *Nature*, *621*(7980), 672-675.

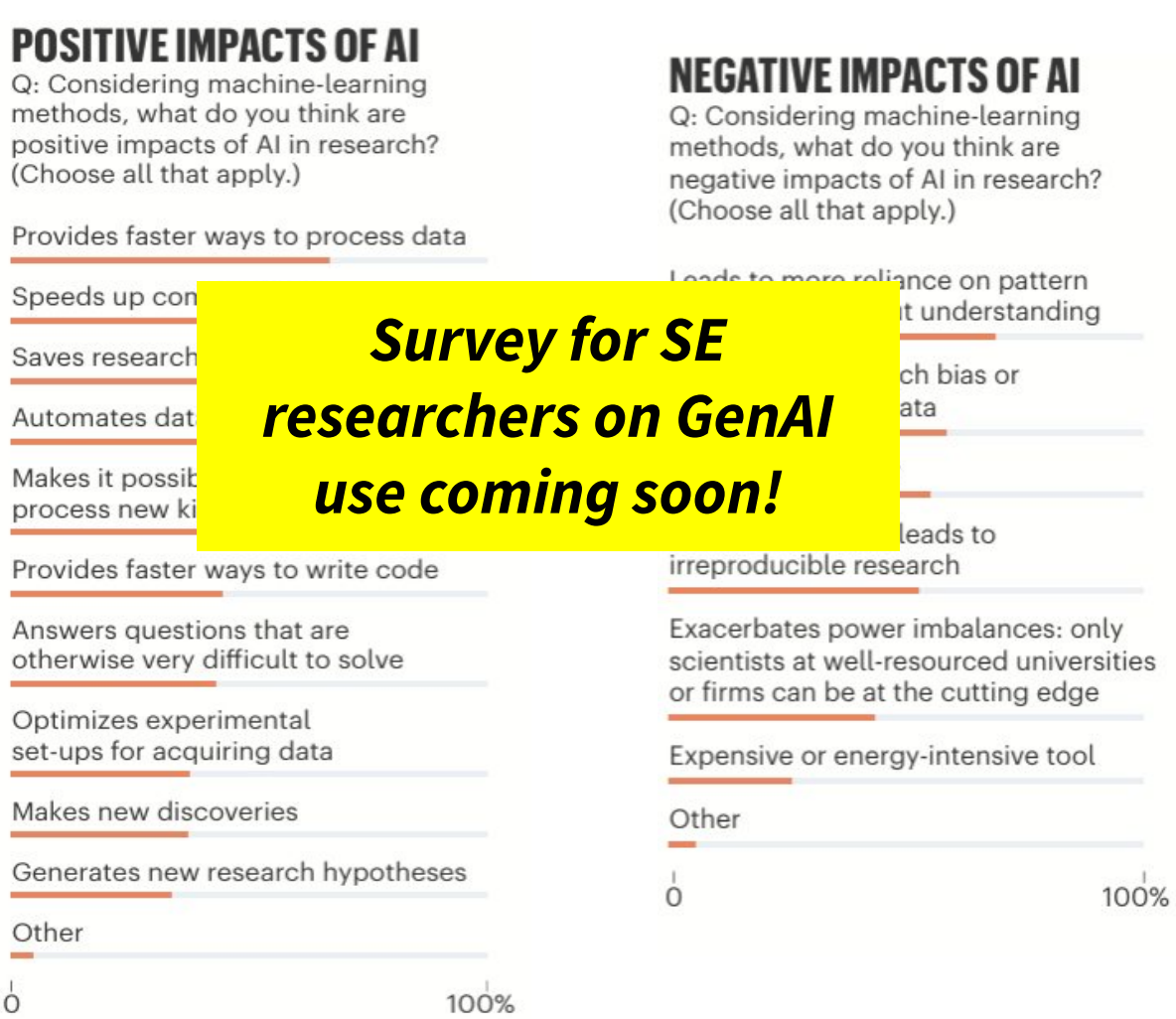




Van Noorden, R., & Perkel, J. M. (2023). AI and science: what 1,600 researchers think. *Nature*, *621*(7980), 672-675.



Van Noorden, R., & Perkel, J. M. (2023). AI and science: what 1,600 researchers think. *Nature*, *621*(7980), 672-675.



Survey for SE researchers on GenAI use coming soon!

Van Noorden, R., & Perkel, J. M. (2023). AI and science: what 1,600 researchers think. *Nature*, *621*(7980), 672-675.

Who-What-How of SE Research

Generative AI in Science

GenAI use in SE Research

Implications?

GenAI use in SE Research


ALFRED P. SLOAN
FOUNDATION
CARLSBERG
FOUNDATION
The Copenhagen
Symposium on
Human
Centered
Software
Engineering AI
November 17-20, 2024
at Aalborg University Copenhagen

GenAI in software engineering must be human centered

Russo, D., Baltes, S., van Berkel, N., Avgeriou, P., Calefato, F., Cabrero-Daniel, B., ... & Vasilescu, B. (2024). Generative AI in Software Engineering Must Be Human-Centered.

research

GenAI in software engineering must be human centered

^

Russo, D., Baltes, S., van Berkel, N., Avgeriou, P., Calefato, F., Cabrero-Daniel, B., ... & Vasilescu, B. (2024). Generative AI in Software Engineering Must Be Human-Centered.

Get on the Train or be Left on the Station: Using LLMs for Software Engineering Research

Bianca Trinkenreich
bianca.trinkenreich@colostate.edu
Colorado State University
Fort Collins, USA

Fabio Calefato
fabio.calefato@uniba.it
University of Bari
Bari, Italy

Geir Hanssen
ghanssen@sintef.no
SINTEF
Trondheim, Norway

Kelly Blincoe
k.blincoe@auckland.ac.nz
University of Auckland
Auckland, New Zealand

Marcos Kalinowski
kalinowski@inf.puc-rio.br
PUC-Rio
Rio de Janeiro, Brazil

Mauro Pezzè
mauro.pezze@usi.ch
USI Università della Svizzera Italiana
Lugano, Italy

Paolo Tell
pate@itu.dk
IT University of Copenhagen
Copenhagen, Denmark

Margaret-Anne Storey
mstorey@uvic.ca
University of Victoria
Victoria, Canada

Abstract

The rapid adoption of Large Language Models (LLMs) is not only transforming software engineering (SE) practice but is also poised to fundamentally disrupt how research is conducted in the field. While perspectives on this transformation range from viewing LLMs as mere productivity tools to considering them revolutionary forces, we argue that the SE research community must proactively engage with and shape the integration of LLMs into research practices, emphasizing human agency in this transformation. As LLMs rapidly become integral to SE research—both as tools that support investigations and as subjects of study—a human-centric perspective is essential. Ensuring human oversight and interpretability is necessary for upholding scientific rigor, fostering ethical responsibility, and driving meaningful advancements in the field. Drawing from discussions at the 2nd Copenhagen Symposium on Human-Centered AI in SE, this position paper employs Marshall McLuhan's Tetrad of Media Laws to analyze the impact of LLMs on SE research. Through this theoretical lens, we examine how LLMs enhance research capabilities through accelerated ideation and automated processes, make some traditional research practices obsolete, retrieve valuable aspects of historical research approaches, and risk reversal effects when taken to extremes. Our analysis reveals opportunities for innovation and potential pitfalls that require careful consideration. We conclude with a call to action for the SE research community to proactively harness the benefits of LLMs while developing frameworks and guidelines to mitigate their risks, to ensure continued rigor and impact of research in an AI-augmented future.

Keywords

Generative AI, LLM, AI4SE, McLuhan's Tetrad

FSE Companion '25, Trondheim, Norway

ACM ISBN 979-8-4007-1276-0/2025/06
https://doi.org/10.1145/3696630.3731666

ACM Reference Format:
Bianca Trinkenreich, Fabio Calefato, Geir Hanssen, Kelly Blincoe, Marcos Kalinowski, Mauro Pezzè, Paolo Tell, and Margaret Anne Storey. 2025. Get on the Train or be Left on the Station: Using LLMs for Software Engineering Research. In *33rd ACM International Conference on the Foundations of Software Engineering (FSE Companion '25), June 23–28, 2025, Trondheim, Norway.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3696630.3731666

1 Introduction

Integrating Large Language Models (LLMs) into Software Engineering (SE) research reflects a broader transformation across scientific disciplines. Generative AI technologies are fundamentally changing how research is conducted, from accelerating hypothesis generation to enhancing data analysis and interpretation [26]. This transformation is particularly relevant for SE research, where LLMs are becoming integral both as subjects of our investigations and as tools we use to conduct research. These models have demonstrated their potential to revolutionize research in our field by supporting various tasks, such as enhancing brainstorming processes [23], generating representative data [7, 20], aiding in data analysis and qualitative research [3], and automating repetitive or tedious tasks.

As SE researchers increasingly incorporate LLMs into their workflows, it becomes crucial to maintain a human-centric perspective, particularly when studying human aspects of SE [18]. The transformative potential of LLMs extends beyond mere automation as these tools can augment our ability to understand developer experiences, team dynamics, and socio-technical interactions in software development. However, this potential must be balanced against the need to preserve human agency and ensure that our research methods remain rigorous, transparent, and ethically sound. This is particularly important as we study how software developers adapt to and integrate LLMs into their work practices, requiring us to critically examine our own use of these tools in researching such phenomena. Therefore, understanding the broad impact of LLMs requires a comprehensive framework that evaluates their benefits and potential unintended consequences.

Marshall McLuhan's Tetrad of Media Effects [14] provides a compelling lens through which to examine the different ways a new

Paper:
https://biancatrink.github.io/files/papers/HumanAISE2025.pdf

WSESE 2025

Towards Evaluation Guidelines for Empirical Studies involving LLMs
Stefan Wagner, Marvin Muñoz Barón, Falessi Davide, Sebastian Baltes (most similar to what we are doing …)

A Framework for Using LLMs for Repository Mining Studies in Empirical Software Engineering
Vincenzo De Martino, Joel Castaño Fernández, Fabio Palomba, Xavier Franch, Silverio Martínez-Fernández

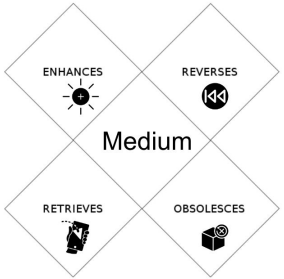
Applications and Implications of Large Language Models in Qualitative Analysis: A New Frontier for Empirical Software Engineering
Matheus de Morais Leça, Lucas Valença, Reydne Bruno dos Santos, Ronnie de Souza Santos

Large Language Model for Qualitative Research - A Systematic Mapping Study
Cauã Ferreira Barros, Bruna Borges Azevedo, Valdemar Graciano Neto, Mohamad Kassab, Marcos Kalinowski, Hugo Alexandre D. do Nascimento, Michelle C.G.S.P. Bandeira

Can Machine Learning Support the Selection of Studies for Systematic Literature Review Updates?
Marcelo Costalonga, Bianca Minetto Napoleão, Maria Teresa Baldassarre, Katia Felizardo, Igor Steinmacher, Marcos Kalinowski

On the difficulties of conducting and replicating systematic literature reviews studies using LLMs in software engineering
Katia Romero Felizardo, Anderson Deizepe, Daniel Coutinho, Genildo Gomes da Silva Junior, Maria Alcimar Costa Meireles, Marco Gerosa, Igor Steinmacher

A Research Playbook for Studying the Impacts of a Disruptive Technology

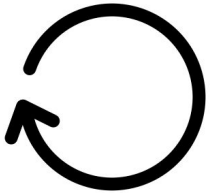


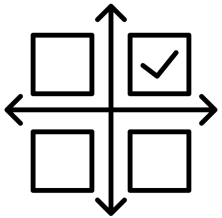
1

Use McLuhan's tetrad to **map out different impacts** of a specific application of the disruptive technology

2

Consider which **phenomenon and ideas** about these phenomenon are relevant to study

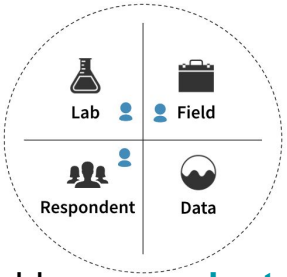




3

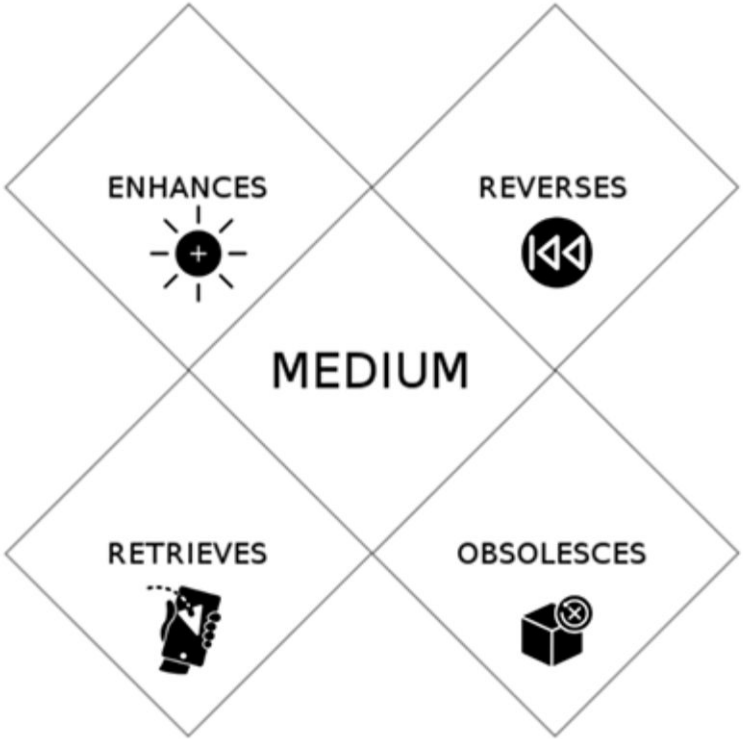
Develop specific **research questions** selecting units of analysis and determine the desired theoretical contributions

4



Select suitable **research strategies** that align with the research questions and phenomena to be studied

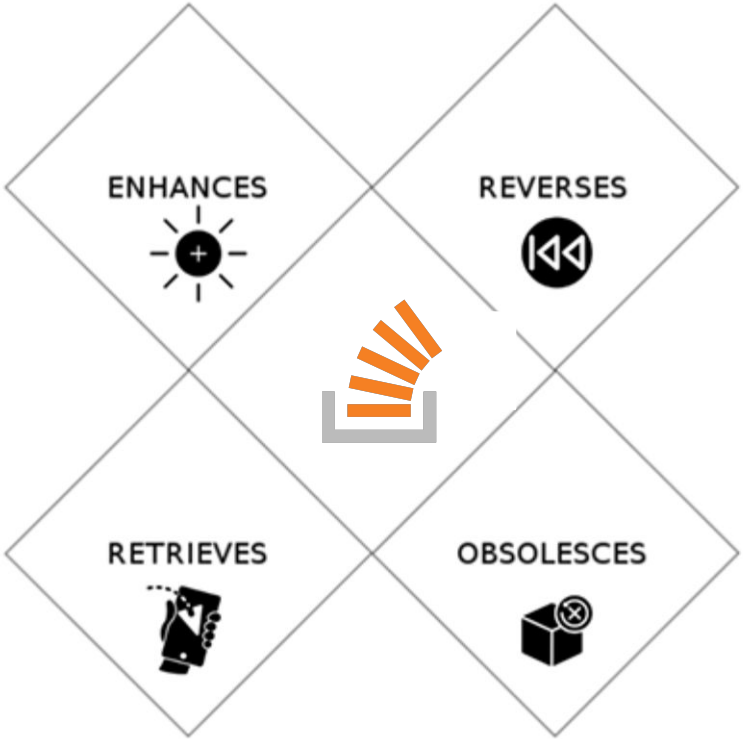
Storey, M. A., Russo, D., Novielli, N., Kobayashi, T., & Wang, D. (2024). A disruptive research playbook for studying disruptive innovations. *ACM Transactions on Software Engineering and Methodology*, *33*(8), 1-29.



27 | McLuhan’s **tetrad**: Four laws of media

McLuhan, M. (1975). McLuhan's Laws of the Media. *Technology and culture*, *16*(1), 74-78.

Faster answers to questions, Debugging, Community



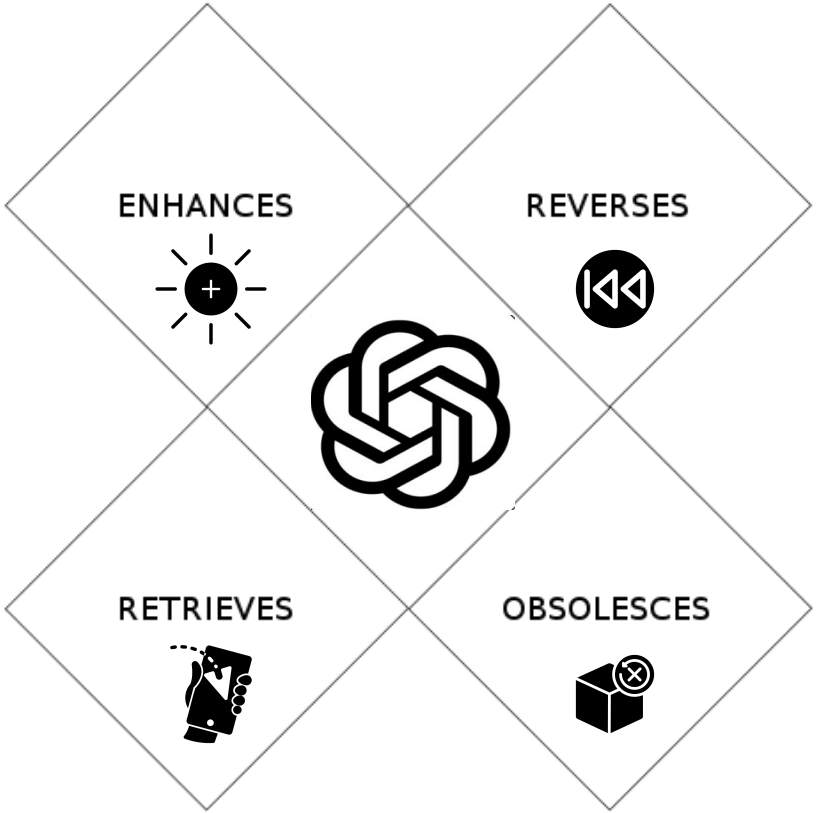
Trust, Blind Hacking, Flips into: LLMs (drive for more customization)

Gurus, Portfolios

Email, Documentation, Textbooks, Onboarding

https://bit.ly/research-playbook

Automation possibilities
Productivity enhancements
Quality improvements
Personalized solutions
Documentation, code reviews
Better time estimates
Communication, Teamwork
Creativity of dev and teamwork
Who can be a developer

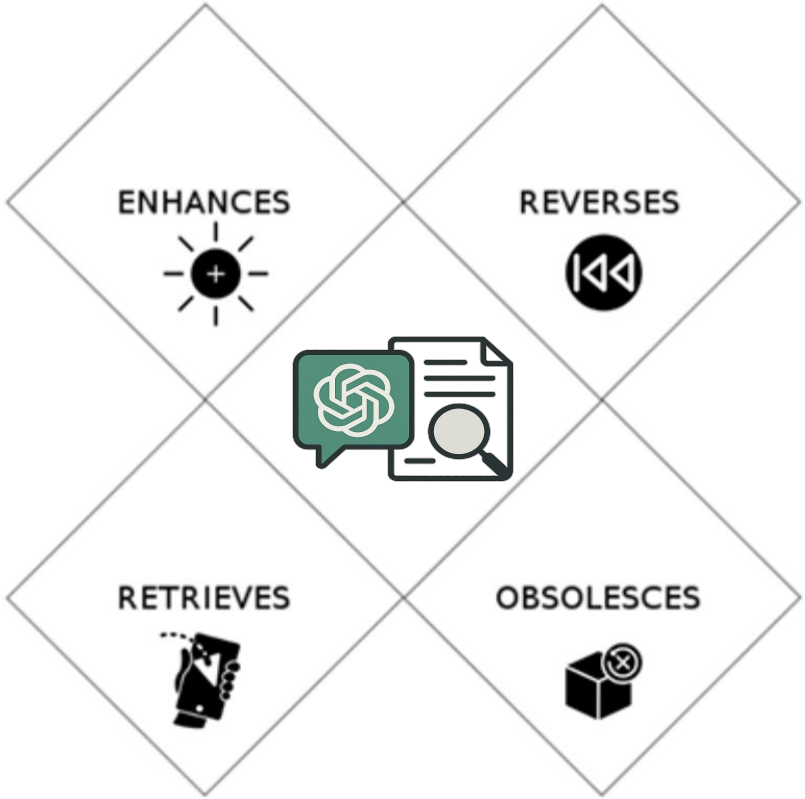


Over-reliance
Homogeneous solutions
Devaluation of dev craft
Loss of control, trust
Low understanding
Lost provenance
Model collapse
Lack of empathy for humans

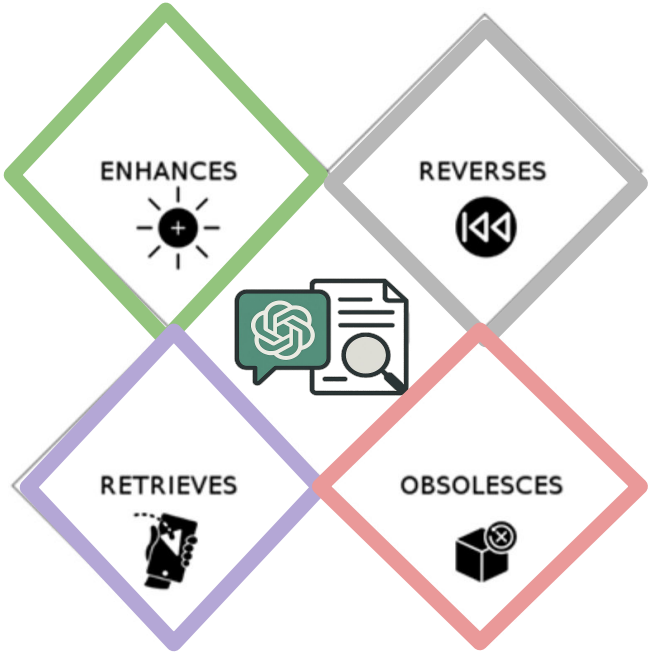
Natural language (pseudocode)
Chatbots, AI agents
End user programmers
Ethical/privacy technology concerns
Improved diversity

Barriers to entry
Writing code
Search, Stack Overflow
Manual documentation,
Manual tests, code reviews
Human troubleshooting
Nerd culture, narrow skills
Traditional education
Some areas of SWEBOK

Software Engineering by and for Humans in an AI Era
Authors: Silvia Abrahão, John Grundy, Mauro Pezzè, Margaret-Anne Storey, Damian Andrew Tamburri, TOSEM, Feb 2025.



 | GenAI for **SE Research** through the lens of McLuhan's Tetrad



1. **Enhance** research capabilities
2. **Obsolesce** traditional research practices or activities
3. **Retrieve** valuable aspects of historical research approaches
4. **Reverse** what happens when taken to extremes (flip ito?)

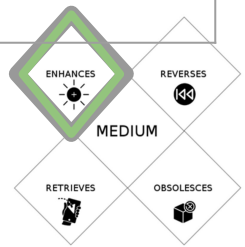
Research Pipeline Stage	Enhance What does it amplify?	Obsolesce What does it push aside?	Retrieve What does it bring back?	Reverse What happens when pushed to extremes?
Research Goals and Questions Formulation (*)	Rapid idea generation (*), auto-suggested hypotheses, literature summarization automation	Manual literature review (*), brainstorming without AI assistance	"Coffee house research" (*), switch trending research topics (*), sketchbook of ideas	Creativity echo chamber (*), Homogenized research questions, potential loss of novelty due to AI-generated ideas
Experimental Design & Methodology	Automated experiment setup, code synthesis for study prototypes, reproducibility improvements	Tedious manual setup, reliance on domain experts for experiment structuring	Human "intractable" models of research field, Modular and reusable experimental designs	Over-reliance on AI-generated methodologies may lead to reduced critical evaluation
Data Collection	Faster extraction from repositories (GitHub, Stack Overflow), automated data cleaning	Human-driven data curation, traditional data wrangling techniques	Historical datasets revisited for new insights	Bias amplification in datasets, lack of transparency in synthetic data creation
Processing	Improved statistical modeling via AI, anomaly identification	Pushing aside the risk of human errors	Large-scale or longitudinal ethnographic studies	Errors and biases that humans cannot easily detect
Analysis & Interpretation (*)	Quantitative, qualitative, and mixed-methods analysis (*), Diverse viewpoints (*)	Manual coding (qualitative and quantitative), Manual selection and execution of statistical techniques (*)	Finding related theories in other domains (*), Holistic and interdisciplinary analysis (*)	AI hallucinations (*) and misleading interpretations if results are blindly trusted
Writing & Dissemination	Automated paper drafting, AI-assisted summaries, multilingual dissemination	Manual academic writing, sole reliance on human synthesis	Collaborative, rapid prototyping of research papers	Proliferation of low-quality or AI-generated papers, diminishing originality and rigor
Cross-cutting impacts (*)	Research speed and creativity	Manual/tedious research tasks	Impactful research	Lower skills of researchers

Enhance research capabilities

What aspects of a researcher's experience or capability does it **amplify or make more efficient**?

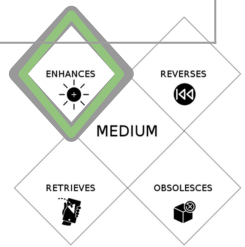
Enhance research capabilities *per research stage*

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Idea generation Auto-suggested hypotheses	Experiment design Code synthesis for study prototypes	Extraction from repositories Data cleaning	Statistical modeling Anomaly identification	Qualitative, quantitative, and mixed-methods analysis	Paper drafting AI-assisted summaries Multilingual dissemination
Cross-Cutting Impact: Research speed and creativity					



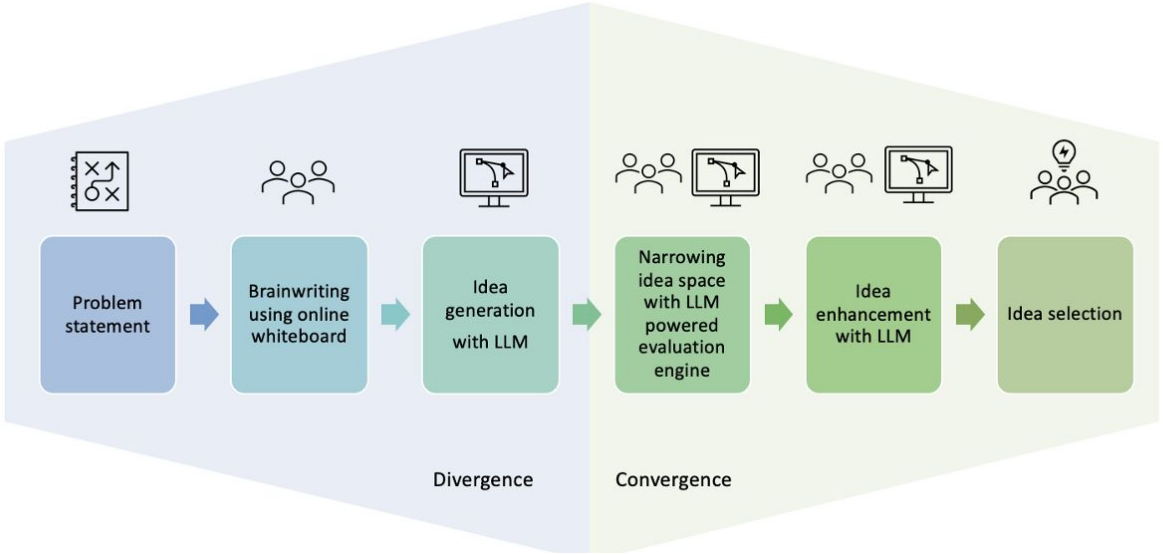
Enhance research capabilities *per research stage*

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Idea generation Auto-suggested hypotheses	Experiment design Code synthesis for study prototypes	Extraction from repositories Data cleaning	Statistical modeling Anomaly identification	Qualitative, quantitative, and mixed-methods analysis	Paper drafting AI-assisted summaries Multilingual dissemination
Cross-Cutting Impact: Research speed and creativity					



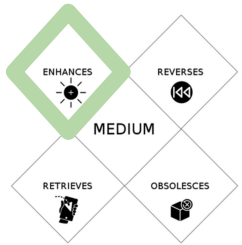
Enhance research capabilities

An Example: Idea Generation - Brainwriting



16 undergraduate students enrolled in an advanced undergraduate course on **tangible interaction design**
5 project teams with 3–4 members. The study was part of a 70-minute in-class session in February 2023

Shaer, O., Cooper, A., Mokryn, O., Kun, A. L., & Ben Shoshan, H. (2024, May). AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-17).



Enhance research capabilities

An Example: Idea Generation - Brainwriting

- **Expanded Idea Diversity**

  GPT-3 introduced novel perspectives and ideas outside human clusters.

- **Increased Idea Volume**

  LLM contributions led to more total ideas; GPT-3 acted as a 'fourth teammate'.

- **Jumpstarting Creativity**

  Helped teams overcome creative blocks and inspired further human ideation.

- **Enhanced Specificity & Novelty**

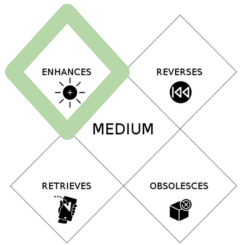
  LLM ideas were often more detailed, technical, and concrete.

Support for the divergence stage of idea generation, and the convergence stage of evaluation and selection of ideas

“It gave us a bunch of ideas we had not thought to offer on our own.”

“A fourth teammate.”

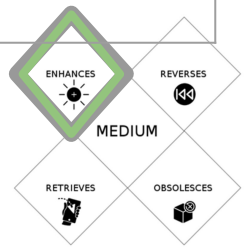
“Inspiration to build upon.”



Shaer, O., Cooper, A., Mokryn, O., Kun, A. L., & Ben Shoshan, H. (2024, May). AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-17).

Enhance research capabilities *per research stage*

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Idea generation Auto-suggested hypotheses	Experiment design Code synthesis for study prototypes	Extraction from repositories Data cleaning	Statistical modeling Anomaly identification	Qualitative, quantitative, and mixed-methods analysis	Paper drafting AI-assisted summaries Multilingual dissemination
Cross-Cutting Impact: Research speed and creativity					

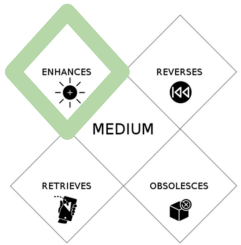


Enhance research capabilities

Experiment Design - Mixed-Methods Research

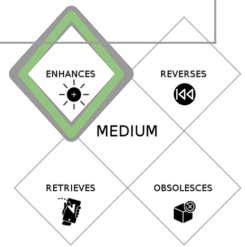
Advise the researcher to avoid known **anti-patterns:**

- ‘Sample contamination’
  - When the same participants are used in multiple phases of a mixed methods study (e.g., in both interviews and follow-up surveys) without accounting for potential bias or influence
- ‘Integration failure’
  - When findings from qualitative and quantitative strands are not meaningfully connected or synthesized



Enhance research capabilities *per research stage*

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Idea generation Auto-suggested hypotheses	Experiment design Code synthesis for study prototypes	Extraction from repositories Data cleaning	Statistical modeling Anomaly identification	Qualitative, quantitative, and mixed-methods analysis	Paper drafting AI-assisted summaries Multilingual dissemination
Cross-Cutting Impact: Research speed and creativity					



Enhance research capabilities

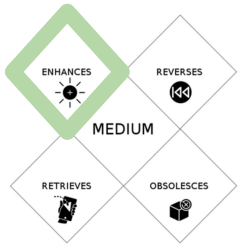
Qualitative, Quantitative and Mixed-Methods Analysis

Quantitative data (e.g., controlled study measurements, test results, defect reports, commit logs)

Analyze structured datasets to extract key patterns and relationships

Qualitative data (e.g., interviews, meeting transcripts)

Identify sentiment and themes (inductively or deductively)

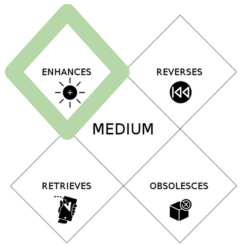


Enhance research capabilities

An example: Sentiment Analysis

Sentiment Analysis: comparing small (e.g., BERT, RoBERTa) and big LLMs (e.g., Llama 2-Chat, Vicuna, WizardLM)

Feature	bLLMs (e.g., Llama 2-Chat)	Traditional Techniques / sLLMs
Data Requirement	Work well with little or **no labeled data** (zero-shot/few-shot)	Require **large labeled datasets** for fine-tuning
Generalization	Handle **imbalanced** or **cross-platform** data better	Struggle with generalizing across platforms
Ease of Use	Require only **natural language prompts**	Require **model training and hyperparameter tuning**
Time and Resource Efficiency	Skip costly fine-tuning; adaptable via **in-context learning**	Costly to train and fine-tune, especially at scale
Versatility	Strong performance across multiple SE domains	Often **domain-specific**, limited generalizability



Zhang, T., Irsan, I. C., Thung, F., & Lo, D. (2025). Revisiting sentiment analysis for software engineering in the era of large language models. *ACM Transactions on Software Engineering and Methodology*, *34*(3), 1-30.

Enhance research capabilities

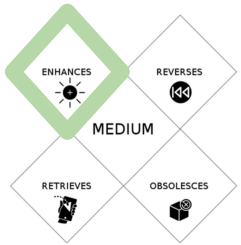
An Example: Thematic Analysis

1. Prompting
   - Output highly sensitive to prompt wording

2. Research Temperature
   - Temperature = 0 for consistency/reproducibility.
   - Temperature = 0.5 or higher to test for theme variation and creativity

Repeated theme extraction at different temperatures to identify the most robust themes
Used LLMs in the phase of defining themes from codes

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, *42*(4), 997-1019.



A Framework for Using LLMs for Repository Mining Studies in Empirical Software Engineering

Vincenzo De Martino, Joel Castaño Fernández, Fabio Palomba, Xavier Franch, Silverio Martínez-Fernández

Applications and Implications of Large Language Models in Qualitative Analysis: A New Frontier for Empirical Software Engineering

Matheus de Morais Leça, Lucas Valença, Reydne Bruno dos Santos, Ronnie de Souza Santos

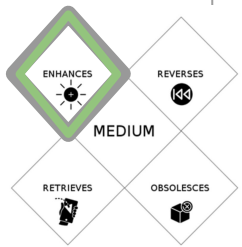
Large Language Model for Qualitative Research - A Systematic Mapping Study

Cauã Ferreira Barros, Bruna Borges Azevedo, Valdemar Graciano Neto, Mohamad Kassab, Marcos Kalinowski, Hugo Alexandre D. do Nascimento, Michelle C.G.S.P. Bandeira

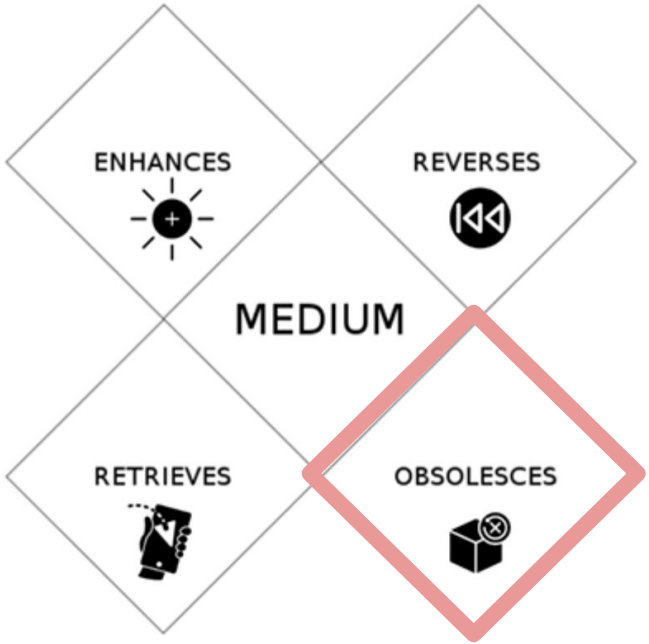
Enhance research capabilities *per research stage*

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Idea generation Auto-suggested hypotheses	Experiment design Code synthesis for study prototypes	Extraction from repositories Data cleaning	Statistical modeling Anomaly identification	Qualitative, quantitative, and mixed-methods analysis	Paper drafting AI-assisted summaries Multilingual dissemination

Cross-Cutting Impact: Research speed and creativity



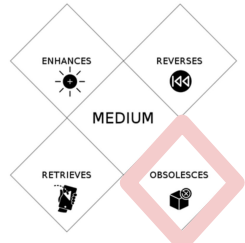
Obsolesce traditional research practices



What does it **displace, reduce, or replace**?

Obsolesce research practices *per research stage*

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Manual literature review	Tedious protocol design	Human-driven data collection	Manual data wrangling Human errors	Manual coding/scripting/ interpretation Decision trees for designs API docs	Background section writing Writing from scratch Hiring editors
Crosscutting: Tedious, repetitive tasks that require low level knowledge					



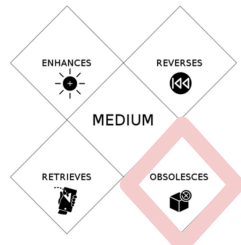
Obsolesce traditional research practices

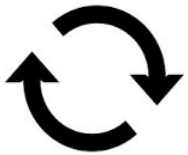
An Example: Manual literature review

- Pair with researcher to apply inclusion/exclusion criteria
- Support refining the search string (from control papers)
- Record the study and update the review years later

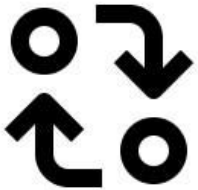
"It is not advisable to completely outsource the selection process to ChatGPT. However, it could be valuable as a support tool, aiding novice researchers or even experienced ones when they are in doubt."

Felizardo, K. R., Lima, M. S., Deizepe, A., Conte, T. U., & Steinmacher, I. (2024, October). ChatGPT application in Systematic Literature Reviews in Software Engineering: an evaluation of its accuracy to support the selection activity. In Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (pp. 25-36).





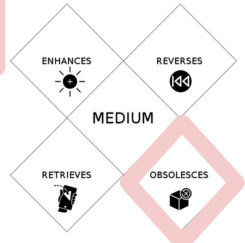
Can Machine Learning Support the Selection of Studies for Systematic Literature Review Updates?
Marcelo Costalonga, Bianca Minetto Napoleão, Maria Teresa Baldassarre, Katia Felizardo, Igor Steinmacher, Marcos Kalinowski



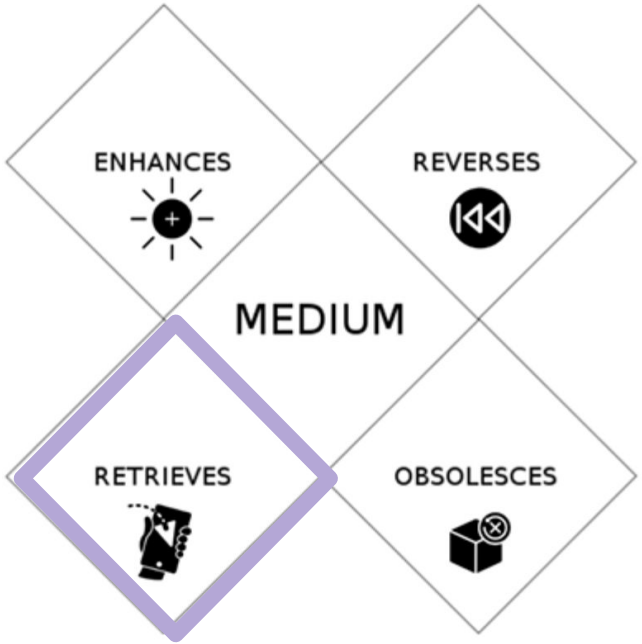
On the difficulties of conducting and replicating systematic literature reviews studies using LLMs in software engineering
Katia Romero Felizardo, Anderson Deizepe, Daniel Coutinho, Genildo Gomes da Silva Junior, Maria Alcimar Costa Meireles, Marco Gerosa, Igor Steinmac

Obsolesce research practices *per research stage*

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Manual literature review	Tedious protocol design	Human-driven data collection	Manual data wrangling Human errors	Manual coding/scripting/ interpretation Decision trees for designs API docs	Background section writing Writing from scratch Hiring editors
Crosscutting: Tedious, repetitive tasks that require low level knowledge					



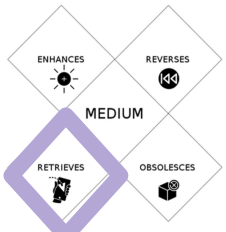
Retrieve aspects of historical research approaches



What **older media, tool, or process** does it bring back in a new way?

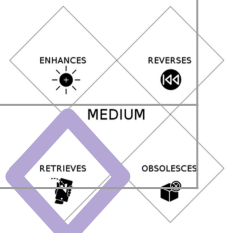
Retrieve historical research approaches by research stage

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Research ideas shaped by dialogue Notebook, index cards, marginalia Former trendy topics	Borrowing of methodologies from other disciplines “Participatory” design	Use of question banks and standard protocols	Hands on style of pipeline to clean, merge, label, re code etc.	Hermeneutics by supporting narrative synthesis Exploratory statistics, where interpretation may guide discovery over hypothesis testing	ChatGPT serves as a modern-day amanuensis—drafting and rephrasing as human scholars dictate and converse about ideas.
Crosscutting: Impactful research (less incremental contributions?)					



Retrieve historical research approaches by research stage

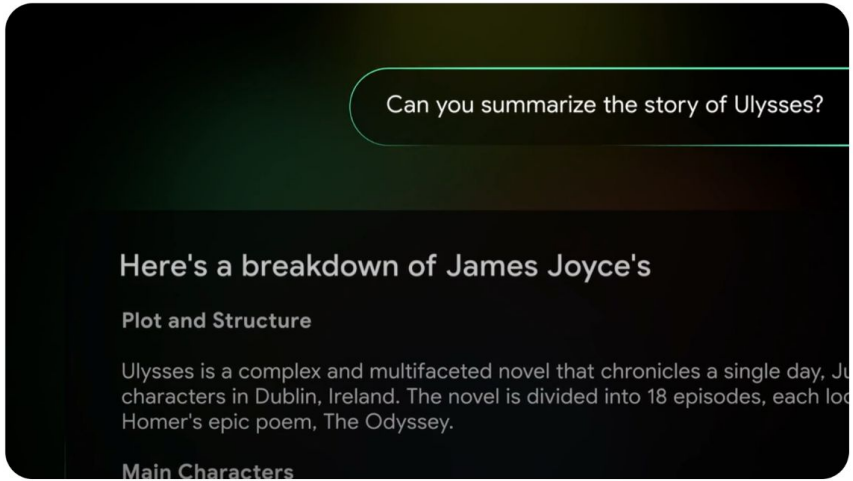
Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Research ideas shaped by dialogue Notebook, index cards, marginalia Former trendy topics	Borrowing of methodologies from other disciplines “Participatory” design	Use of question banks and standard protocols	Hands on style of pipeline to clean, merge, label, re code etc.	Hermeneutics by supporting narrative synthesis Exploratory statistics, where interpretation may guide discovery over hypothesis testing	ChatGPT serves as a modern-day amanuensis—drafting and rephrasing as human scholars dictate and converse about ideas.
Crosscutting: Impactful research (less incremental contributions?)					

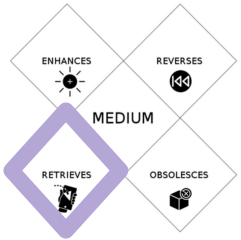


Your Personalized AI Research Assistant

Upload your sources

Upload PDFs, websites, YouTube videos, audio files, Google Docs, or Google Slides, and NotebookLM will summarize them and make interesting connections between topics, all powered by Gemini 2.0's multimodal understanding capabilities.





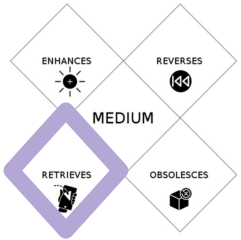
Retrieve aspects of historical research approaches

An Example: Study goals – Bring back former topics of interest

Requirements engineering is making a comeback as a “hot topic”

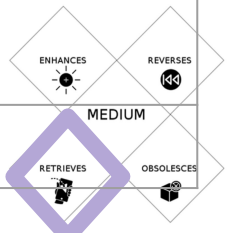
- GenAI can transform unstructured data into structured requirements, supporting automated elicitation, inconsistency detection, and traceability

Krishna, M., Gaur, B., Verma, A., & Jalote, P. (2024, June). Using LLMs in software requirements specifications: an empirical evaluation. In *2024 IEEE 32nd International Requirements Engineering Conference (RE)* (pp. 475-483). IEEE.

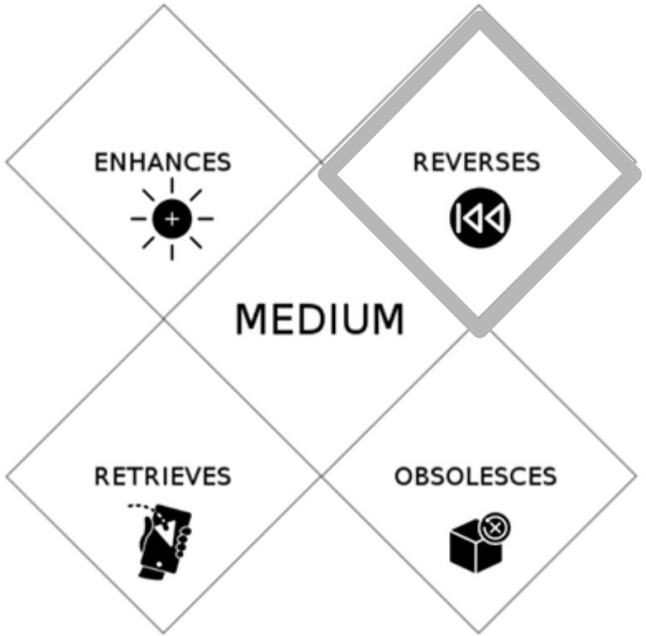


Retrieve historical research approaches by research stage

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Research ideas shaped by dialogue Notebook, index cards, marginalia Former trendy topics	Borrowing of methodologies from other disciplines “Participatory” design	Use of question banks and standard protocols	Hands on style of pipeline to clean, merge, label, re code etc.	Hermeneutics by supporting narrative synthesis Exploratory statistics, where interpretation may guide discovery over hypothesis testing	ChatGPT serves as a modern-day amanuensis—drafting and rephrasing as human scholars dictate and converse about ideas.
Crosscutting: Impactful research (less incremental contributions?)					



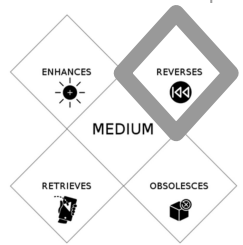
Reverses what was intended when taken to extreme



When used to the extreme, what are the **negative or opposite effects (of those intended)** that emerge? What does the "technology" flip into?

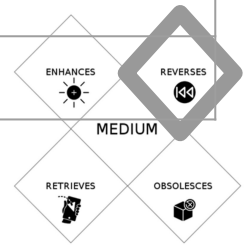
Reverses what was intended when taken to extreme

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Loss of deep situated context leading to shallow questions Studying the obvious rather than what is of societal importance	Less critical evaluation of fewer methodological choices Cargo cult methodological choices	Over standardization Loss of context of humans studied Biases in synthetic data	Errors due to black boxes in the pipeline	Too much trust without domain understanding	Diminishing rigor and originality and quality of papers
Crosscutting: Lossed skills by researchers					



Reverses what was intended when taken to extreme

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Loss of deep situated context leading to shallow questions Studying the obvious rather than what is of societal importance	Less critical evaluation of fewer methodological choices Cargo cult methodological choices	Over standardization Loss of context of humans studied Biases in synthetic data	Errors due to black boxes in the pipeline	Too much trust without domain understanding	Diminishing rigor and originality and quality of papers
Crosscutting: Lossed skills by researchers					

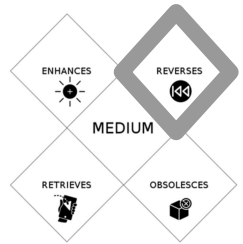


Reverse risks when taken to extremes

Creativity echo chamber

- Individuals within a creative field are primarily exposed to ideas and perspectives that reinforce their own, rather than being challenged by diverse viewpoints

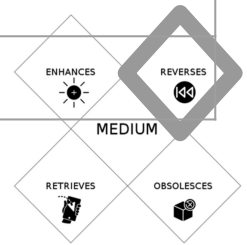
- This can lead to ***homogenized ideas*** and a ***lack of groundbreaking creativity***
- ***But it may also over time lead to more creativity!***



Reverses what was intended when taken to extreme

Research Goals & Questions Formulation	Experimental Design & Methodology	Data Collection	Processing	Analysis & Interpretation	Writing & Dissemination
Loss of deep situated context leading to shallow questions Studying the obvious rather than what is of societal importance	Less critical evaluation of fewer methodological choices Cargo cult methodological choices	Over standardization Loss of context of humans studied Biases in synthetic data	Errors due to black boxes in the pipeline	Too much trust without domain understanding	Diminishing rigor and originality and quality of papers

Crosscutting: Lossed skills by researchers

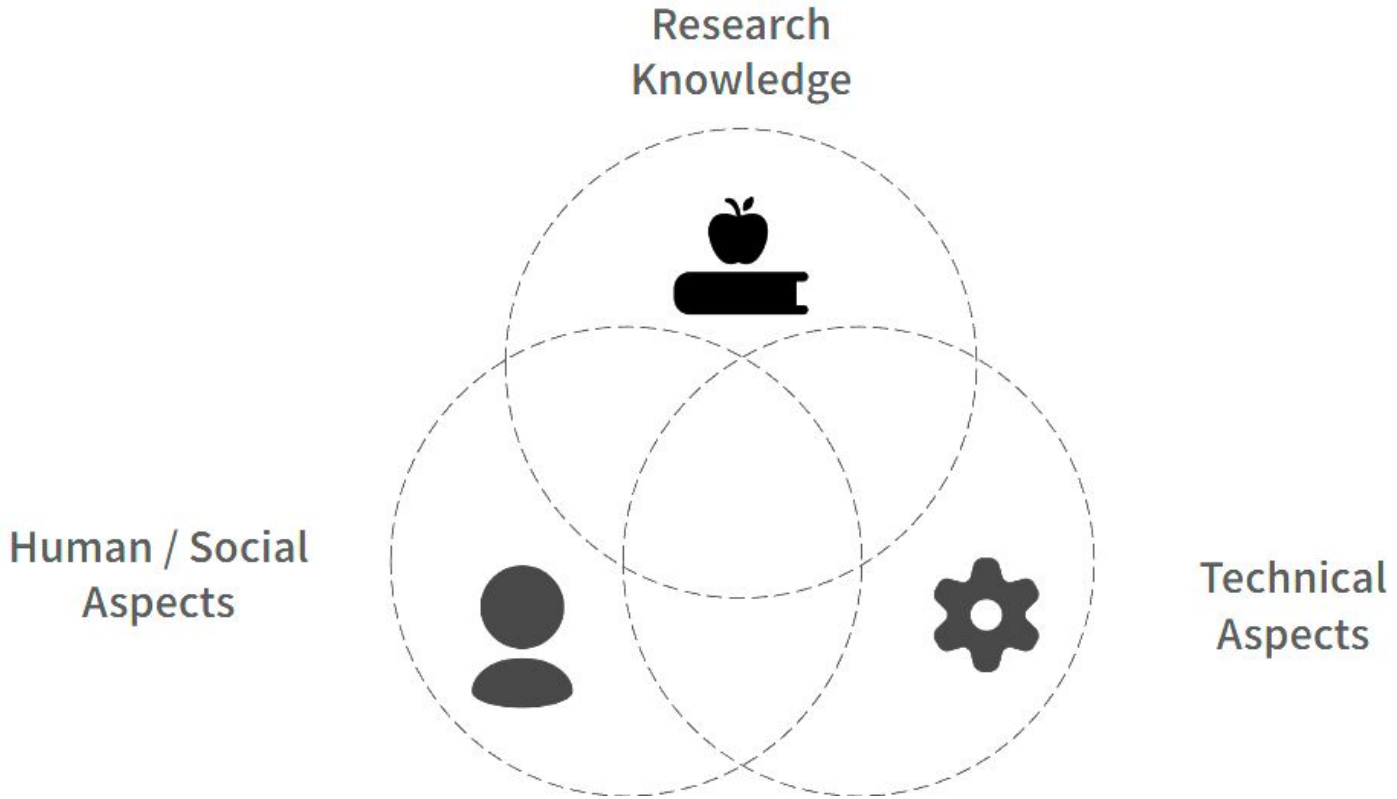


Who-What-How of SE Research

Generative AI in Science

GenAI use in SE Research

Implications?



 | Impact on **who** our research serves?

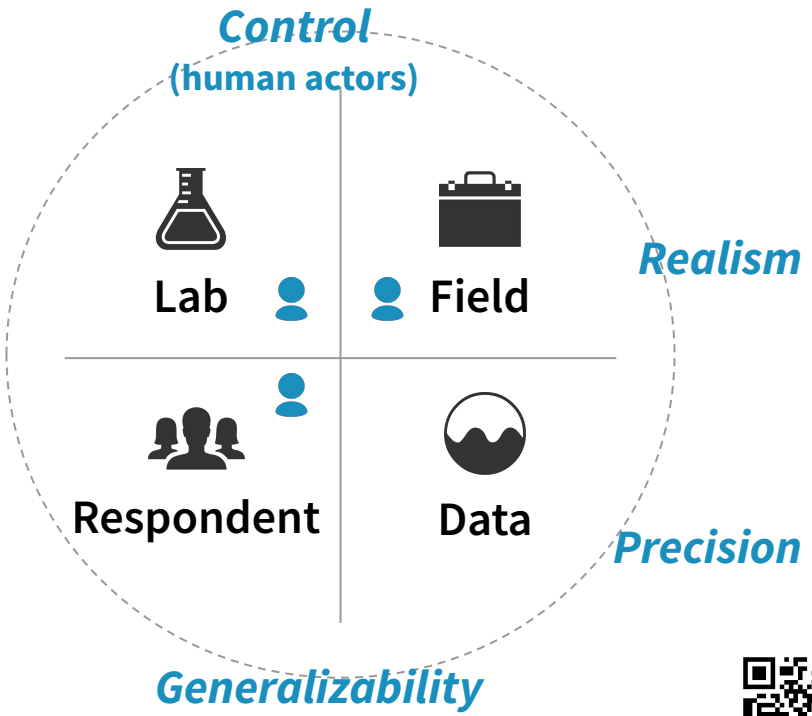
Impact on **what** we do?

Non-Empirical

Formal Theory

Meta

Empirical



Impact on **how** we conduct our studies and mix our methods?

https://bit.ly/wwh-framework

What kinds of "**knowledge**" are still **valid** when GenAI is involved? Is the researcher still the interpreter and driver?

Will the use of GenAI in research reinforce the **biases** in the data being studied (that may also be a focus of the study)?

How will the use of GenAI support **reproducibility** and **transparency**? Do we need to define what is needed?

Researcher-AI collaboration needs to be studied (cognitive fatigue, shifting roles)?

Will we need new **theories** to study impact of GenAI on SE research?

What new **literacies** may researchers need to acquire?

Methodological implications?

GenAI **blurs traditional boundaries** between developers, users, agents, and researchers

GenAI introduces **new phenomena** to study (e.g., dynamics of Human-AI collaboration), in turn leading to new **theories**

GenAI tools influence both the creation and evolution of **software artifacts** and brings **new data** to study (e.g., interaction logs and generated artifacts)

Non-deterministic outputs from AI models and integrating AI-generated data into analyses may cause our **research methods to evolve** to account for the co-creation processes between humans and AI

The use of GenAI in software engineering research may introduce new **validity threats**

A **Paradigm Shift** in SE Research?

https://bit.ly/genAI-ESE
(Treude & Storey)

Conduct **comparison studies** (with real data) with the use of GenAI and without

Develop **benchmarks** and **validation protocols**

Standardize reporting of GenAI use in our research (it should be clear what was used and when/how, document tools/models/prompts used)

Capture our findings in **theories** so we can compare / aggregate our work

We need to study the implications of **GenAI use over time** (topic/method drift etc)

Education and training on core skills, workshops/technical briefings to teach core Researcher-AI role

Calls to Action

Should we get on this train?

Margaret-Anne Storey
https://www.linkedin.com/in/margaret-anne-storey-8419462/