

Prompt or perish – The publication process in times of genAl

RWTH Publishing Forum (#7) 27.3.2025

Prof. Dr. Sandra Geisler, Data Stream Management and Analysis, RWTH Aachen University





Some Housekeeping

- this session is not recorded
- slides will be available on the website of the RWTH University Library (→ Research → Scholarly Publishing → Publishing Forum) and in RWTH Publications

Save-the-date: next Publishing Forum

May 22, 2025, 1:00 p.m. (Zoom)

- Ideas for further topics?
 - ⇒ Let us know! <u>rtp@ub.rwth-aachen.de</u>







Prompt or perish – The publication process in times of genAl

Prof. Sandra Geisler Forum Publizieren 27.03.2025









Large language models can draft abstracts or suggest these artificial-intelligence tools are a work in progre

ou know that text autocomplete func- California, created the most well-known tion that makes your smartphone so GPT-3, in 2020, by training a network to convenient – and occasionally frus- dict the next piece of text based on what

Feature



Researchers are excited but apprehensive about how tools such as ChatGPT could transform science and society. By Chris Stokel-Walker and Richard Van Noorden

embarked on an unusual experiment: they asked an assistant who was not a scientist to help them improve three of their research papers. Their assiduous aide suggested revisions to sections of documents in seconds; each manuscript took about five minutes to review. In one biology manuscript, their helper even spotted

n December, computational biolo- amistakeinareference to an equation. The trial gists Casey Greene and Milton Pividori didn't always run smoothly, but the final manuscripts were easier to read - and the fees were modest, at less than US\$0.50 per document.

This assistant, as Greene and Pividori reported in a preprint on 23 January, is not a person but an artificial-intelligence (AI) algorithm called GPT-3, first released in 2020. It is one of the much-hyped generative ative AIs can produce images, or sounds. Al chatbot-style tools that can churn out

produce prose, poetry, computer code or in the scientists' case - to edit research par

The most famous of these tools, also kn as large language models, or LLMs, is Chatt a version of GPT-3 that shot to fame after release in November last year because it made free and easily accessible. Other ge "I'm really impressed," says Pividori,

information Review

EXPLORE our articles SUBMIT a manuscript

SEPTEMBER 3, 2024

SHARE Y () In W DOWNLOAD PDF

PEER REVIEWED

GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation

Academic journals, archives, and repositories are seeing an increasing number of questionable research papers clearly produced using generative AI. They are often created with widely available, general-purpose AI applications, most likely ChatGPT, and mimic scientific writing. Google Scholar easily locates and lists these questionable papers alongside reputable, quality-controlled research. Our analysis of a selection of questionable GPTfabricated scientific papers found in Google Scholar shows that many are about applied, often controversial topics susceptible to disinformation: the environment, health, and computing. The resulting enhanced potential for malicious manipulation of society's evidence base, particularly in politically divisive domains, is a growing concern.

Department of Arts and Cultural Sciences,

Lund University, Sweden

BY JUTTA HAIDER

Swedish School of Library and Information Science, University of Borås, Sweden

MALTE RÖDL

Division of Environmental Communication, Swedish University of Agricultural Sciences, Sweder

KRISTOFER ROLF SÖDERSTRÖM BJÖRN EKSTRÖM

Swedish School of Library and Information Science, University of Borås, Sweden

[Hutson, Nature, 2022]

[Jo, Nature, 2023]

[Haider et al., Harvard K. S. Misinformation Review, 2024]



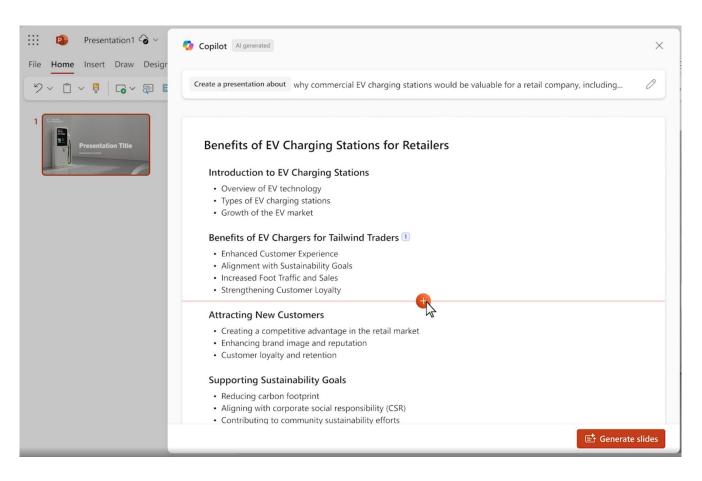








- Text completion and generation
- Summarization, question answering, translation, ...
- Intelligent assistants
 - Office applications
 - Teaching & learning
 - Programming
 - Robot control



https://www.microsoft.com/de-DE/microsoft-365/copilot?









- Text completion and generation
- Summarization, question answering, translation, ...
- Intelligent assistants
 - Office applications
 - Teaching & learning
 - Programming
 - Robot control



https://github.com/features/copilot

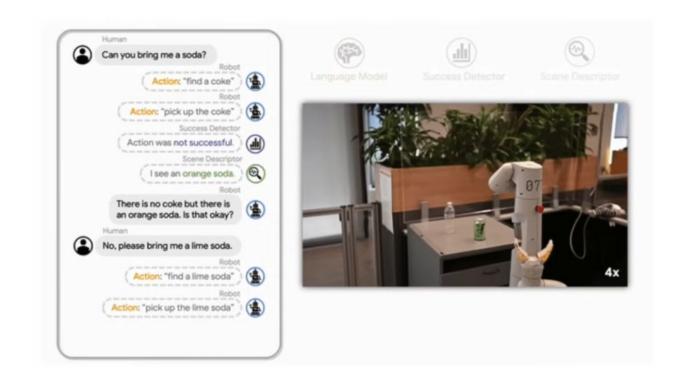








- Text completion and generation
- Summarization, question answering, translation, ...
- Intelligent assistants
 - Office applications
 - Teaching & learning
 - Programming
 - Robot control



Robotics in the Age of Generative AI with Vincent Vanhoucke, Google DeepMind | NVIDIA GTC 2024 https://www.youtube.com/watch?v=vOrhfyMe EQ

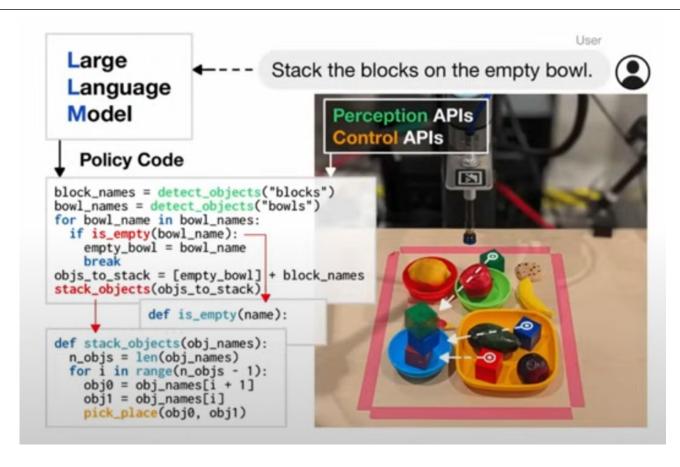








- Text completion and generation
- Summarization, question answering, translation, ...
- Intelligent assistants
 - Office applications
 - Teaching & learning
 - Programming
 - Robot control



Robotics in the Age of Generative AI with Vincent Vanhoucke, Google DeepMind | NVIDIA GTC 2024 https://www.youtube.com/watch?v=vOrhfyMe EQ

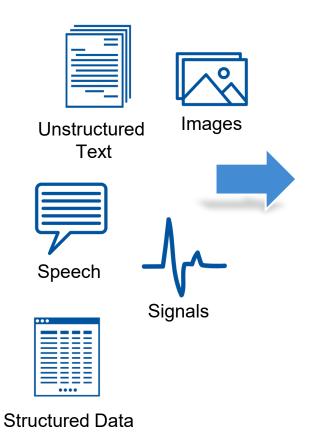


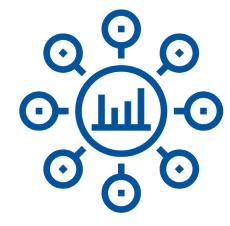






Adapting LLMs – Method 1: Foundation models & fine-tuning





Foundational Model







- sing 1.1%
- see people 4.6%
- push the pedals 6.7 %
- have an accident 0.5 %















Fine-tuned Model

The medical report points to a

- Atrial fibrillation 5 %
- Myocardial infarction 3.4 %
- Arrythmia 5.8 %



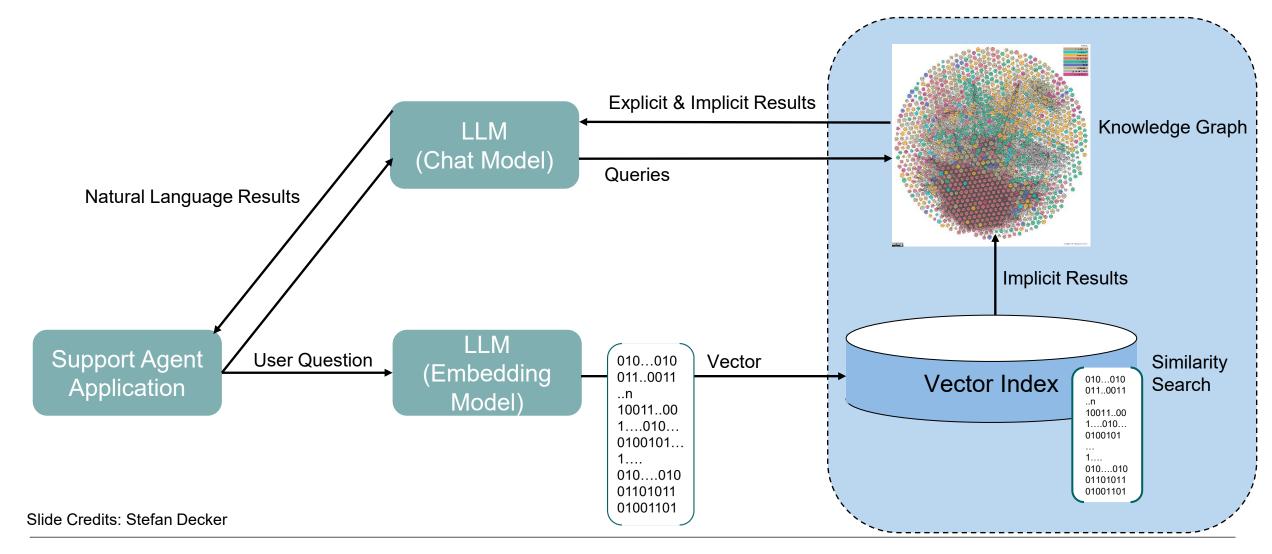








Adapting LLMs – Method 2: Retrieval-Augmented Generation (RAG)











nature

Explore content > About the journal > Publish with us > Subscribe

nature > news > article

NEWS 04 February 2025

How are researchers using AI? Survey reveals pros and cons for science

Despite strong interest in using artificial intelligence to make research faster, easier and more accessible, researchers say they need more support to navigate its possibilities.

By Miryam Naddaf







Researchers think that for some tasks, generative AI tools can already do a better job than huma Credit: Getty

Using artificial intelligence (AI) tools for processes such as preparing manuscripts, writing grant applications and peer review will become widely accepted within the next two years, suggests a survey of nearly 5,000 researchers in more than 70 countries by the publishing

- Well-known: 90% heard of, 81% used it (ChatGPT)
- **Limited usage:** only 45% have used AI in their research process
- **Limited knowledge:** 40-50% unfamiliar with training, policies, best practices, 41% unsure about available tools
- 81% have concerns
 - around ethics (54%),
 - transparency in how they work/are trained (46%),
 - level of accuracy (51%)
 - information security/privacy (47%)

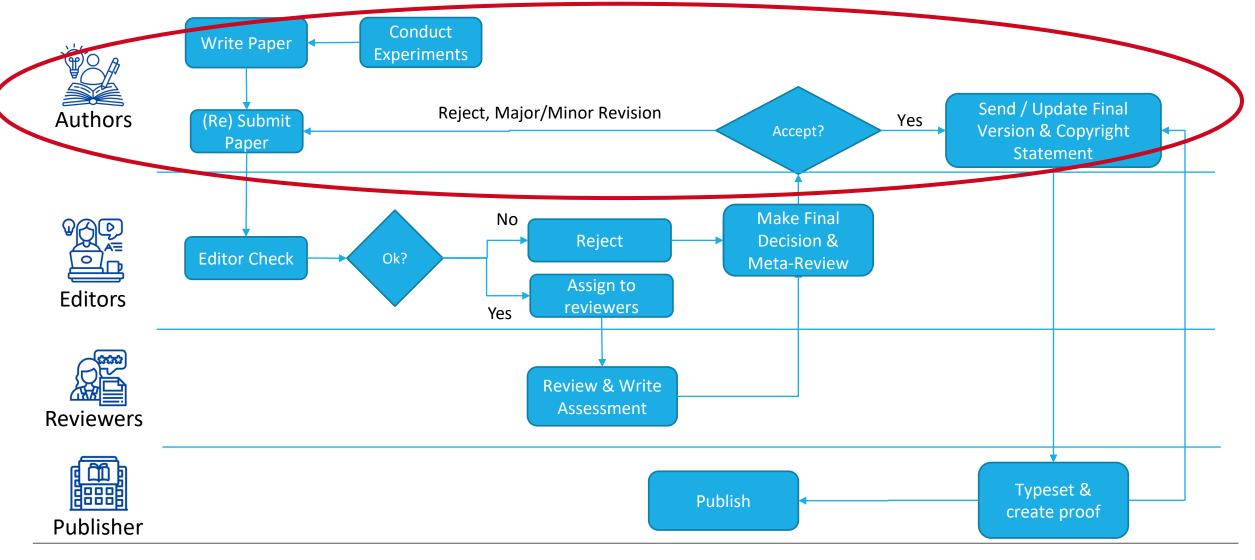








Roles and tasks in the research publication process











Author Tasks



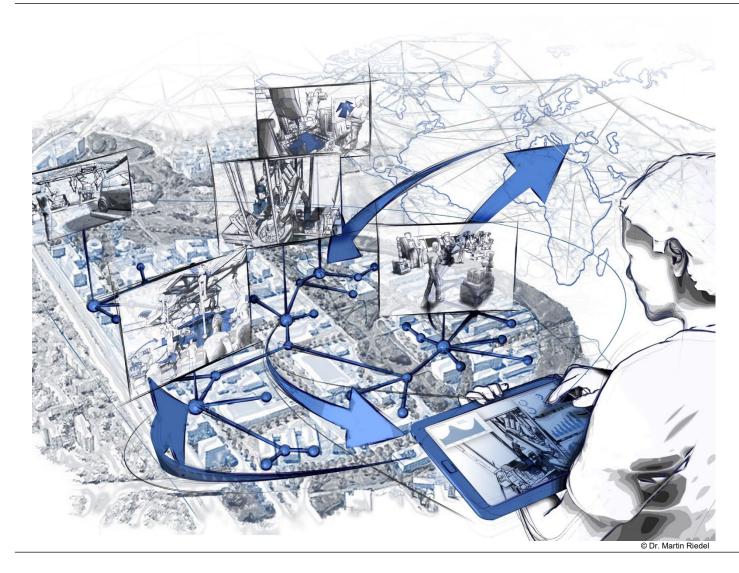








EXC "Internet of Production" - Smart, Connected Production



"The vision of the Internet of Production (IoP) is to enable a new level of cross-domain collaboration by providing semantically adequate and contextual data from production, development and usage in real time at an appropriate granularity."

https://www.iop.rwth-aachen.de/









LLMs in the FAIR Research Data Management Cycle



Researcher



Project management



Funder



Data stewards



Central Services



Publication

- suggest data market places/repositories
- consulting on licenses



- similar experiments
- related publications, datasets
- relevant tools, repositories
- potential collaborators



Writing proposals



- funder-specific text
- job profiles, e.g., data stewards

Writing DMPs





automatic creation by abstract

help texts on demand

create data summarization



9

Choice of accessibility



identifying sensitive data suggesting concepts for data



Archival

protection

- create archival guidelines
- create archival pipelines



Data modelling



- create model
- check for irregularities
- suggest standard language



- Experiment design create documentation scheme
- suggest algorithms, analysis methods

Data acquisition



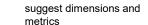
- create input masks
- declaration of consent text

Preprocessing and transformation



- recommend and implement format changes
- suggest and produce pipeline, e.g., as Jupyter notebook
- suggest methods for feature selection





suggest cleaning methods



Choice of storage

informing about legal conditions for different storage solutions best practices for established solutions, e.g., Coscine

Structure



- suggest naming conventions
- suggest storage structure

[Geisler et al., 2024]











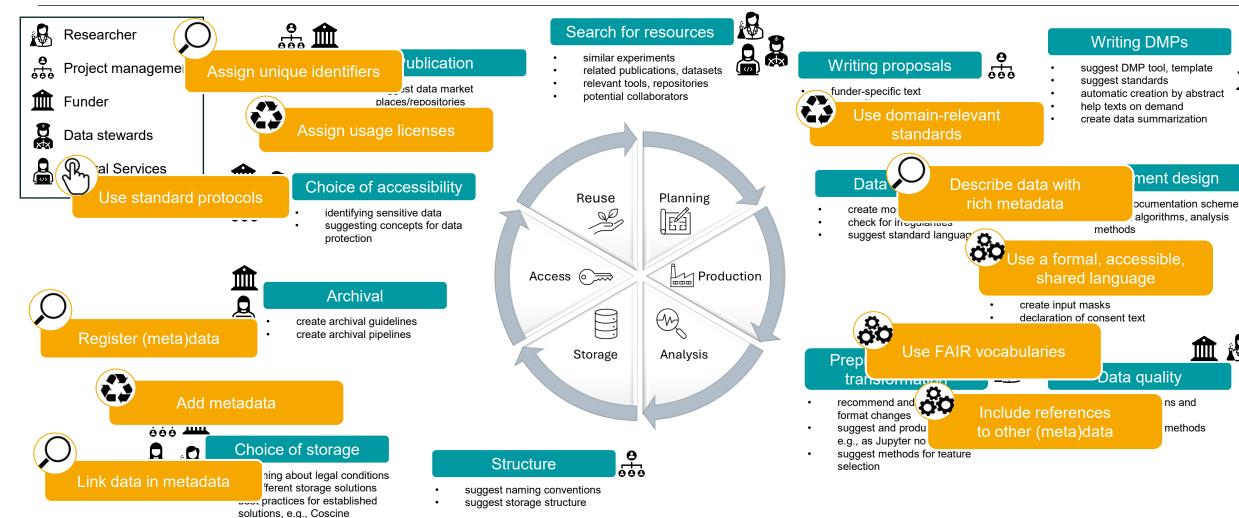


LLMs in the FAIR Research Data Management Cycle



9

血



[Geisler et al., 2024]









The Task: Find or Create an Ontology for Metadata Annotation

Materials Research Forum LLC
Materials Research Proceedings 41 (2024) 1436-1445

Predicting fine blanking process signals from sheet metal thickness

MOON Jiyoung^{1,a*}, GELBICH Daria^{1,b}, BECKER Marco^{1,c}, NIEMIETZ Philipp^{1,d}
and BERGS Thomas^{1,2,e}

¹Manufacturing Technology Institute MTI of RWTH Aachen University, Campus-Boulevard 30, 52074 Aachen, Germany

²Fraunhofer Institute for Production Technology IPT, Steinbachstr. 17, 52074 Aachen, Germany

^aj.moon@mti.rwth-aachen.de, ^bd.gelbich@mti.rwth-aachen.de, ^cm.becker@mti.rwth-aachen.de,

^dp.niemietz@mti.rwth-aachen.de, ^et.bergs@mti.rwth-aachen.de

Keywords: Sheet M

Abstract. In sheet m conditions and produ inaccessibility of the t process, such as force information, containin complexity of the proc depend on different ir process signals, which

Name nent of process rates and the manufacturing ground truth uable source of as well as the process signals label data mean thickness scaler model 4.05 4.05 training data 4.05 4.05 4.05 4.05 4.05 4.05

4.05

4.05

[Moon et al., 2024]



-7050.158868533855,-6545.197609458095,-7069.382093213849

-6974.289952503707,-6450.2257491644905,-7004.105439762875

-6888.948591670646, -6305.549122272035, -6907.7276024169705

-6751.457746721394,-6118.219682318749,-6766.739859775819 -6565.683333005749,-5939.763193920102,-6578.869012704587

-6378.316085990314,-5778.07016807039,-6404.416245037385

-6205.615321967468,-5646.562328180286,-6231.639715977421

-6064.323691867901,-5552.310844229093,-6085.741985763209

-5960.395467087033,-5498.973085403063,-5982.132704421241

-3300.333407007033,-3430.373003403003,-3302.13270442124.

-5893.883970595199,-5485.416242488852,-5921.408832479124

-5870.69053169583, -5517.542405219507, -5894.467360214813

-5897.609639445589, -5579.899152752291, -5913.902163190129

-5952.995595790824,-5634.899606358309,-5971.307026071324

-6014.17204583685,-5665.853566171,-6020.791055125605

-6046.391837716967,-5674.289459888521,-6048.577688236898









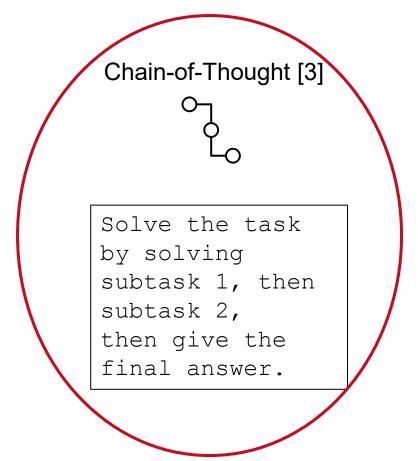
© WZL, Tobias Kaufmann

Prompting Techniques – Creating Qualitative Answers

Inner Monologue [2]



Have an inner monologue on how to solve the task first.
Only then give a final answer.



Tree-of-Thought [4]



Propose multiple solutions for subtask 1.

Evaluate the suggestions.

Proceed with subtask 2 with the knowledge from the prior step. Derive the final answer.

[2] https://medium.com/contact-research/think-before-you-speak-the-inherent-statelessness-of-large-language-models-ceff8ab20ff1, last accessed 13 Nov, 2023. [3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, D. Zhou: Chain of thought prompting elicits reasoning in large language models. arXiv preprint. DOI:https://doi.org/10.48550/arXiv.2201.11903 (2022).

[4] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan: Tree of Thoughts: Deliberate Problem Solving with Large Language Models. ArXiv preprint. DOI:https://doi.org/10.48550/arXiv.2305.10601 (2023).









Chain-of-thought for Ontology Creation



The zip file (DiCo.zip) is a data set that accompanies the paper (986058.pdf).

You are supposed to be a research assistant that helps the researcher of the paper to annotate the data set. The outcome that you should produce is an ontology that explains and defines the data set and the results of the paper.

- 1. The most useful result for the researcher would be real world ontology which already exists on the internet, which you would then apply to the data set.
- 2. The second most useful result is if you extend an already existing ontology fitting to that research area with vocabulary that you define vourself.
- 3. As a last resort if you cannot find any existing ontology, define on yourself for the data set of the researcher.

Please inform me which approach you used and list all the things that you defined yourself, so that I now which is already existing work, and what you invented.

This is supposed to help the researcher how otherwise has to do the work manually.

Please output the ontology as RDF. The definitions you invented as a list of text. And also give a quick summary of what you did.

< 2/2



Slide Credits: Martin Görz. Stefan Decker



Before defining a custom ontology, it is best to explore existing ontologies in relevant domains like manufacturing, process monitoring, and generative models. Based on the context of the paper (fine blanking, sheet metal forming, process signal prediction), we should check for ontologies in the following areas:

- 1. Manufacturing Ontologies: Standardized frameworks for representing manufacturing processes, tools, and measurements. For example:
 - MASON (Manufacturing Service Ontology): Often used in manufacturing process planning and monitoring.
 - PSL (Process Specification Language): A widely accepted ontology for describing manufacturing and engineering processes.
 - QUDT (Quantities, Units, Dimensions, and Data Types Ontology): Used for defining measurable quantities like force signals.
- 2. Time Series and Sensor Data Ontologies: These ontologies can cover the time-series process signals in this context.
 - SSN/SOSA (Semantic Sensor Network Ontology): This ontology describes sensors and observations, which could be a natural fit for the force sensors and process signals.
- 3. Generative Model Ontologies: While less common, there could be emerging ontologies for machine learning, especially in relation to GANs and time-series prediction.

Next Steps

I will perform a search for suitable existing ontologies in these domains. If I find an applicable ontology, I can either directly use it or extend with domain-specific vocabulary from the dataset











Chain-of-thought for Ontology Creation



The zip file (DiCo.zip) is a data set that accompanies the paper (986058.pdf).

You are supposed to be a research assistant that helps the researcher of the paper to annotate the data set. The outcome that you should produce is an ontology that explains and defines the data set and the results of the paper.

- 1. The most useful result for the researcher would be real world ontology which already exists on the internet, which you would then apply to the data set.
- 2. The second most useful result is if you extend an already existing ontology fitting to that research area with vocabulary that you define yourself
- 3. As a last resort if you cannot find any existing ontology, define one yourself for the data set of the researcher.

Please inform me which approach you used and list all the things that you defined yourself, so that I now which is already existing work, and what you invented.

This is supposed to help the researcher how otherwise has to do the work manually.

Please output the ontology as RDF. The definitions you invented as a list of text. And also give a quick summary of what you did.

< 2/2 >



Message ChatGPT

Slide Credits: Martin Görz, Stefan Decker



Ontology Output (extended with SSN/SOSA and PSL):

```
@prefix ex: <http://example.org/ontology/> .
@prefix psl: <http://www.mel.nist.gov/psl/psl.owl#> .
@prefix sosa: <http://www.w3.org/ns/sosa/> .
@prefix ssn: <http://www.w3.org/ns/ssn/> .
ex:ForceObservation a sosa:Observation;
    sosa:madeBySensor ex:ForceSensor;
    sosa:observedProperty ex:Force .
ex:GeneratedSignal a ex:GeneratedData;
    ex:conditionedBy ex:SheetThickness;
    ex:generatedBy ex:CDCGAN;
    ex:isDerivedFrom ex:ProcessSignal .
ex:FineBlankingProcess a psl:Process;
    psl:hasStage ex:ShearingStage,
        ex:StrippingStage
ex:ForceSensor a ssn:Sensor .
ex:ProcessSignal a ex:Concept;
    ex:belongsToStage ex:ProcessStage ;
    ex:conditionedBy ex:SheetThickness;
    ex:generatedBy ex:CDCGAN;
    ex:hasMeasurement ex:Sensor;
    psl:belongsToProcess ex:FineBlankingProcess;
    psl:hasParameter ex:SheetThickness .
ex:ProcessStage a ex:Concept .
ex:Sensor a ex:Concept .
ex:ShearingStage a psl:Stage .
ex:StrippingStage a psl:Stage
ex:CDCGAN a ex:Concept .
ex:SheetThickness a ex:Concept,
        psl:Parameter;
    ex:hasValue "4.05 mm" .
```

Summary:

I have extended the ontology using the SSN/SOSA framework for sensors and observations and PSL for representing the manufacturing processes. The fine blanking process stages (e.g., shearing and stripping) are modeled using PSL, while sensor data (e.g., force measurements) is modeled using SSN/SOSA.



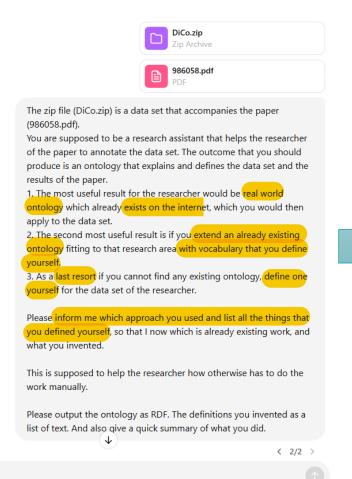








Chain-of-thought for Ontology Creation



C DataFile (C) Material File Name File Format Material Name File Size Material Type
 Material Composition File Description File ID (Persistent Identifier)
 Associated Repository (Repository ID) Material Properties Material ID (Persistent Identifier) Associated Experiment (Experiment ID) Associated Temperature Curve(s) (Curve ID) C Experiment C TemperatureCurve C DataRepository (c) Material Property © Element Experiment Title o Exposure Type o Exposure Duration Repository Name Melting Point o Experiment Description Element Symbol Repository Type Boiling Point Date Conducted Exposure Temperature Element Name o Access Rights Thermal Conductivity Data Points Atomic Number Specific Heat Capacity Experiment ID (Persistent Identifier) License o Percentage Curve ID (Persistent Identifier) Density Repository ID (Persistent Identifier) Associated Publication(s) Associated Material (Material ID) C Publication C Researcher C DataPoint Title o Name o Affiliation • Time (seconds) o Contact Information Publication Date • Temperature (°C) Researcher ID (Persistent Identifier) Publication ID (Persistent Identifier)

Slide Credits: Martin Görz, Stefan Decker

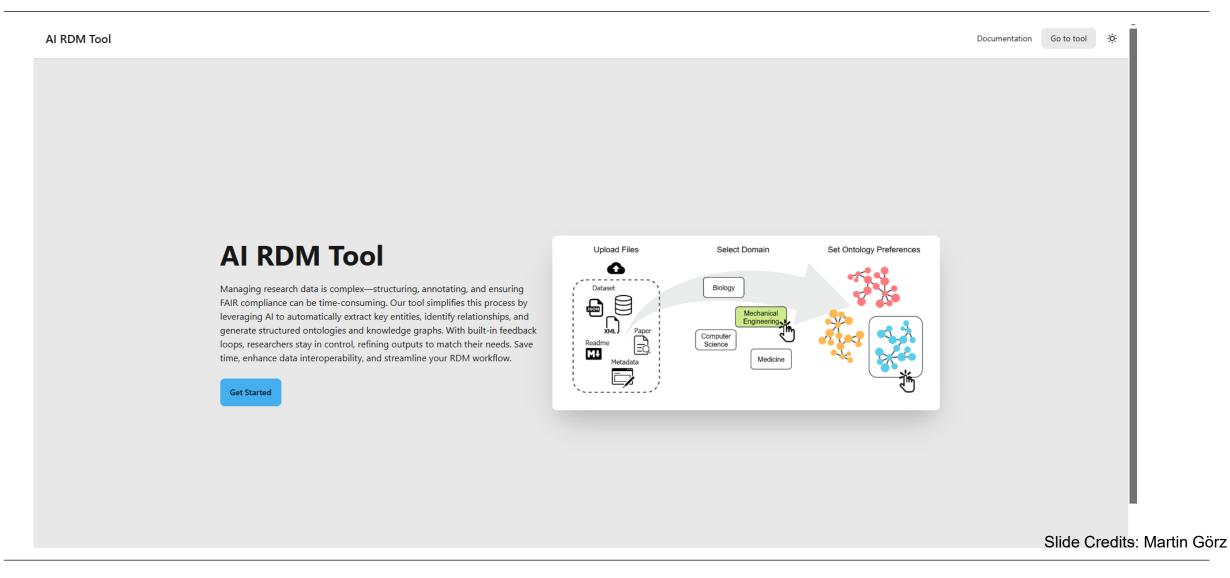








Message ChatGPT











AI RDM Tool Documentation Go to tool Find or Upload Ontologies Upload Ontology File No file chosen CHOOSE FILES XML Search for Ontologies \propto iron sheet envo iron sheet An iron sheet is a mass of iron which has been forged into a roughly planar form of less than 6 millimetres in thickness. envo sheet-iron exterior wall A sheet-iron exterior wall is an exterior wall that is built from sheets of iron. envo sheet-iron building roof A sheet-iron building roof is a building roof that is built from sheets of iron. Selected Ontologies

Slide Credits: Martin Görz







×

×



TIB ifc https://standards.buildingsmart.org/IFC/DEV/IFC4/ADD2_TC1/OWL#METAL

TIB emmo http://emmo.info/emmo#EMMO_494b372c_cfdf_47d3_a4de_5e037c540de8

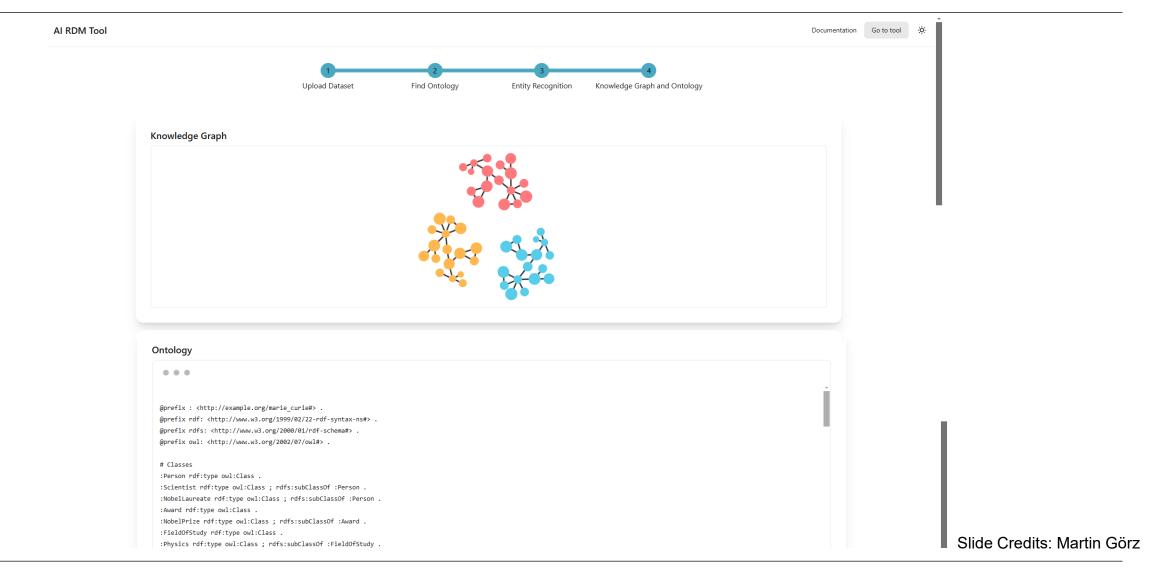
AI RDM Tool Documentation Go to tool file object DiCo-R_schneiden_y_seg2_scaled.csv DiCo-R_schneiden_y_seg3_scaled.csv file object M2 variable M4 variable M6 variable ground_truth dataset label_data dataset scaler_model dataset training_data dataset abstreifen processing step schneiden processing step RWTH Aachen University Organization Fraunhofer Institute for Production Technology IPT Organization Person Jiyoung Moon Daria Gelbich Person Slide Credits: Martin Görz



















Experiment Design – Grinding in Manufacturing



We would like to investigate for grinding in manufacturing, how we can predict the grain size of the grinding wheel by using the signal of the acoustic emission of the grinding wheel and the dressing procedure. What steps in a data science pipeline would you advise?



Based on a thesis in cooperation with WZL, RWTH Aachen University



Predicting the grain size of a grinding wheel using the signal of the acoustic emission and the dressing procedure is an interesting data science project. Here are the steps you can follow in a typical data science pipeline to approach this problem:

1. Problem Definition:

 Clearly define the problem you want to solve. In this case, it is predicting the grain size of a grinding wheel.

2. Data Collection:

Collect data related to the acoustic emission signal and the dressing procedure. This
may include historical data, measurements, and records from the manufacturing
process.

3. Data Preprocessing:

- Clean the data by handling missing values, outliers, and noise.
- . Transform and normalize the data as necessary.
- Ensure data is in a suitable format for analysis.

4. Feature Engineering:

Extract relevant features from the acoustic emission signal and dressing procedure data.
 This could include statistical measures, frequency domain features, time series features,
 and any domain-specific features that might be relevant.

Data Splitting:

 Divide the data into training, validation, and testing sets. Cross-validation may also be used for model tuning.

6. Model Selection:

 Choose appropriate machine learning or statistical models for regression or classification, depending on the nature of your prediction problem.

7. Model Training:

Train the selected models using the training data.

8. Model Evaluation:

- Evaluate the model's performance using appropriate evaluation metrics. For regression tasks, common metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. For classification, metrics like accuracy, precision, recall, and F1-score might be used.
- 9. Hyperparameter Tuning:









Experiment Design – Grinding in Manufacturing



We would like to investigate for grinding in manufacturing, how we can predict the grain size of the grinding wheel by using the signal of the acoustic emission of the grinding wheel and the dressing procedure. What steps in a data science pipeline would you advise?



Predicting the grain size of a grinding wheel using the signal of the acoustic emission and the dressing procedure is an interesting data science project. Here are the steps you can follow in a typical data science pipeline to approach this problem:

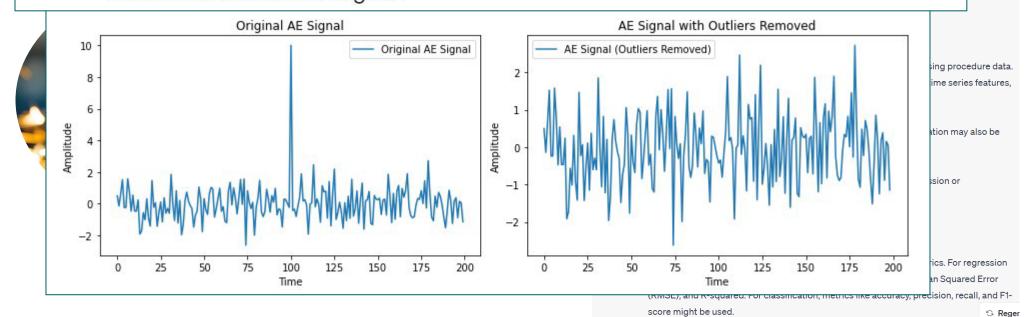
1. Problem Definition:

 Clearly define the problem you want to solve. In this case, it is predicting the grain size of a grinding wheel.





For the Z-Score method: can you create a Jupyter notebook which applies the z-score method on a fictive AE signal?



Based on a thesis in cooperation with WZL, RWTH Aachen University



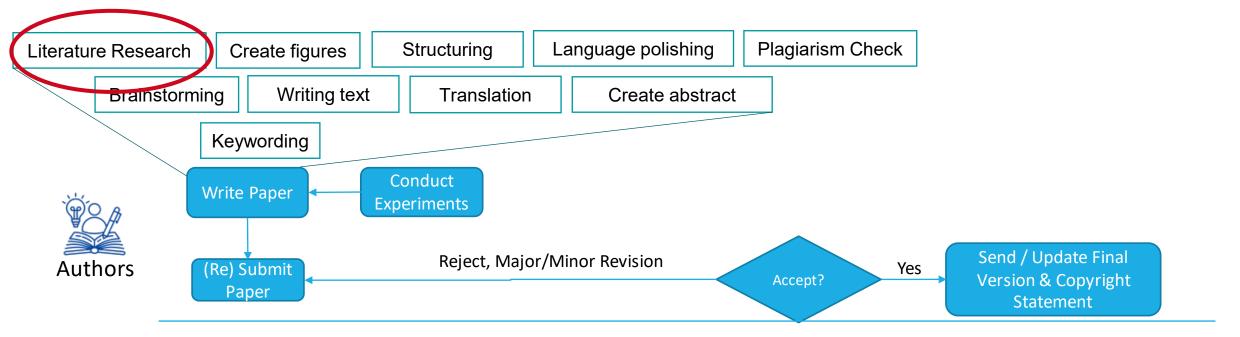


Hyperparameter Tuning:





Author Tasks

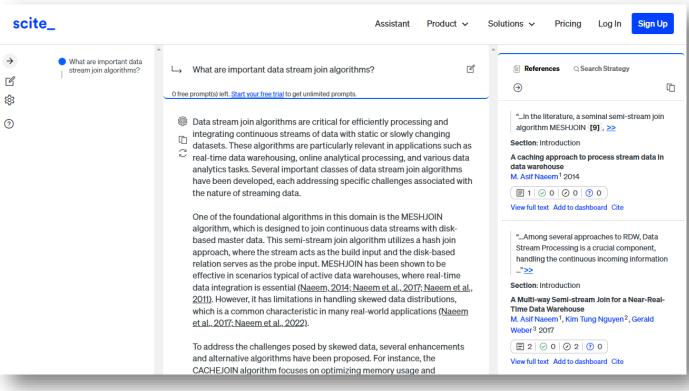












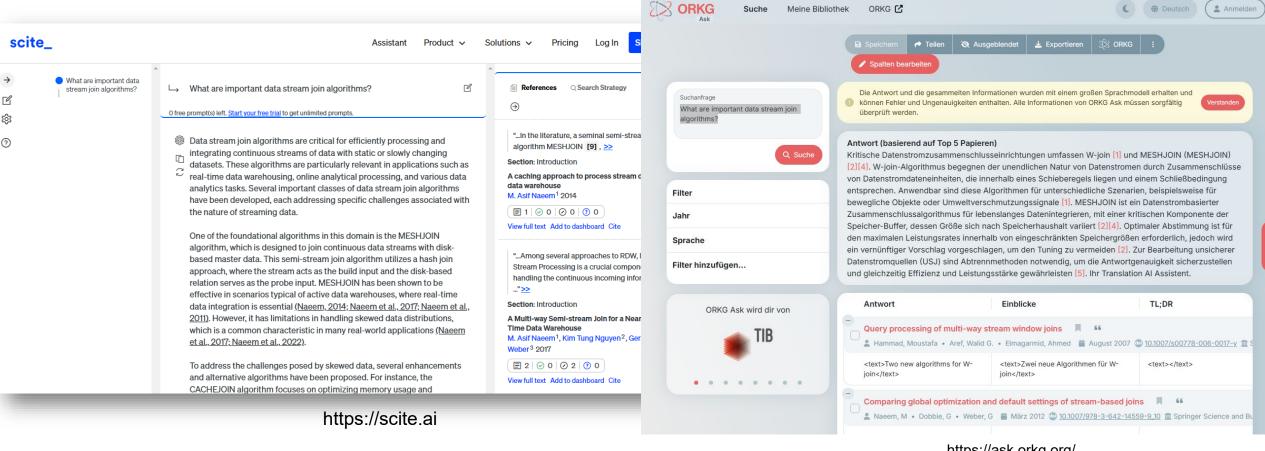
https://scite.ai











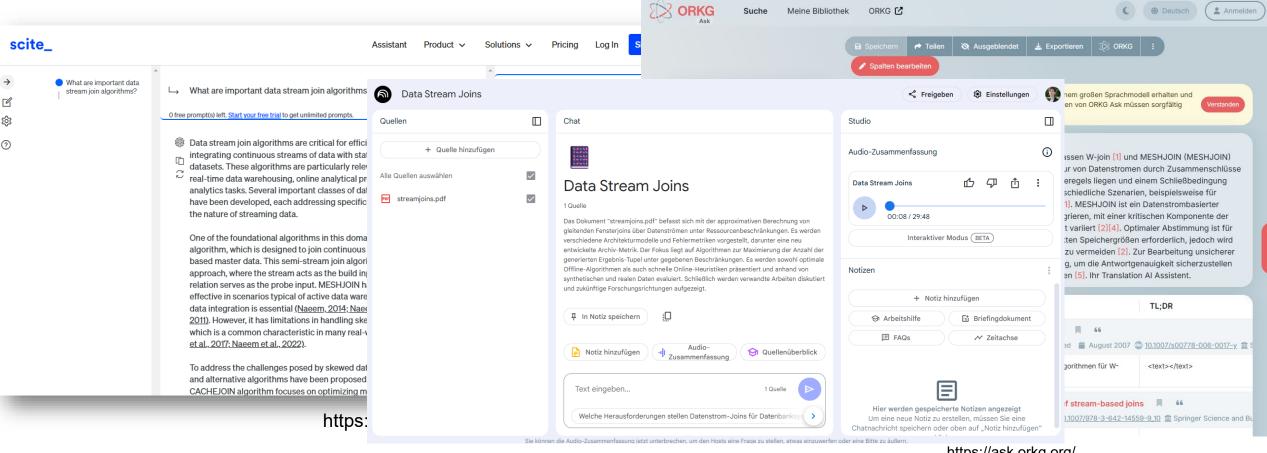
https://ask.orkg.org/











https://notebooklm.google.com/

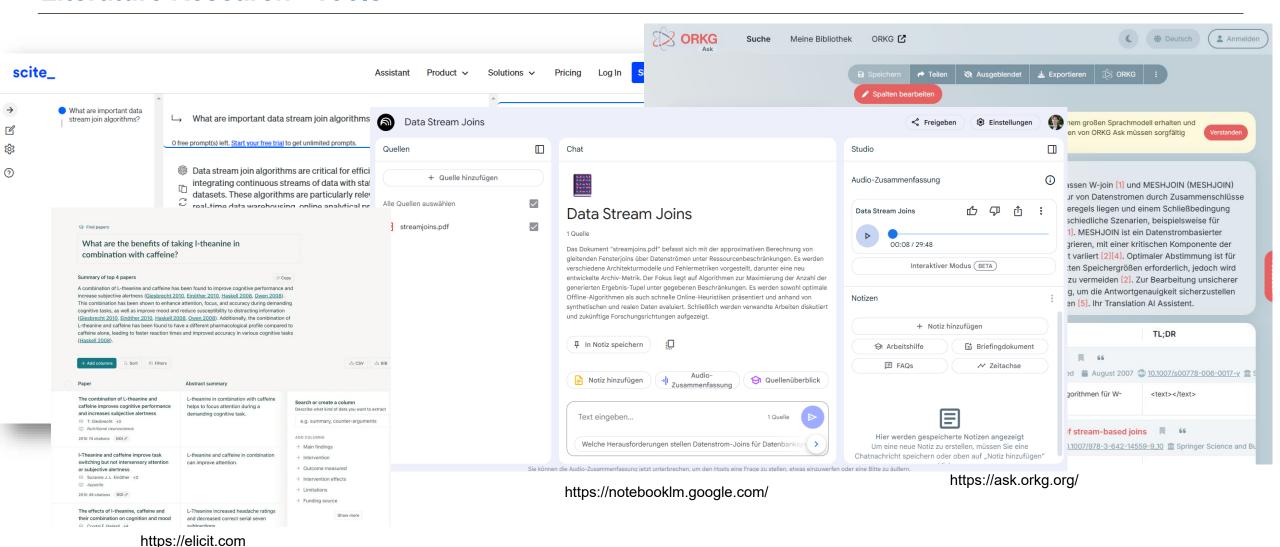
https://ask.orkg.org/











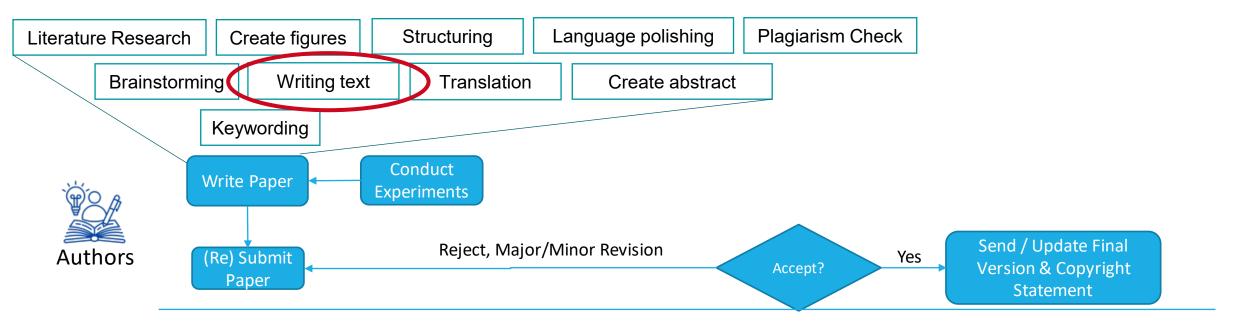








Author Tasks











Al as an Author – No Accountability

Did you use generative AI to write this manuscript?

Association for omputing Machinery

Generative AI is not an author. These tools should only be used to improve language and readability, with caution. If you used generative AI or AIassisted technology, include the following statement directly before the references at the end of your manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used [NAME TOOL / SERVICE] in order to [REASON]. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

AI Authorship



Frequently Asked Questions

ACM Policy on Authorship

Can I use generative AI software tools to prepare my manuscript?

Yes, the use of such tools is permitted, subject to the following requirements:

- . That these systems do not plagiarize, misrepresent, or falsify content in ACM submissions.
- . That the resulting Work in its totality is an accurate representation of the authors' underlying work and novel intellectual contributions and is not primarily the result of the tool's generative capabilities.
- . That the authors accept responsibility for the veracity and correctness of all material in their Work, including any computer-generated material.

Large Language Models (LLMs), such as ChatGPT, do not currently satisfy our authorship criteria (imprint editorial policy link). Notably an attribution of authorship carries with it accountability for the work, which cannot be effectively applied to LLMs. Use of an LLM should be properly documented in the Methods section (and if a Methods section is not available, in a suitable alternative part) of the manuscript. The use of an LLM (or other Altool) for "Al assisted copy editing" purposes does not need to be declared. In this context, we define the term "Al assisted copy editing" as Al-assisted improvements to humangenerated texts for readability and style, and to ensure that the texts are free of errors in









Al as an Author – No Accountability

Did you use generative AI to write this manuscript?

Generative AI is not an author. These tools should only be used to improve language.

genAl guidelines

24% of the 100 largest publishers87% of top 100 highly ranked journals

Frequently Asked Que
ACM Policy on Authorship

Can I use generative AI software tools

Yes, the use of such tools is permitted, sub

- · That these systems do not plagiarize, misrepres
- That the resulting Work in its totality is an accurate repondent intellectual contributions and is not primarily the result
- That the authors accept responsibility for the veracity and correctness or including any computer-generated material.

Prohibition of genAl as author 96% of the 100 largest publishers 98% of top 100 highly ranked journals

Ganjavi et al., The BMJ, 2023

an attribution of authorship carries with it effectively applied to LLMs. Use of an LLM thods section (and if a Methods section is not af the manuscript. The use of an LLM (or other Alarposes does not need to be declared. In this context, copy editing" as Al-assisted improvements to human-

onity and style, and to ensure that the texts are free of errors in

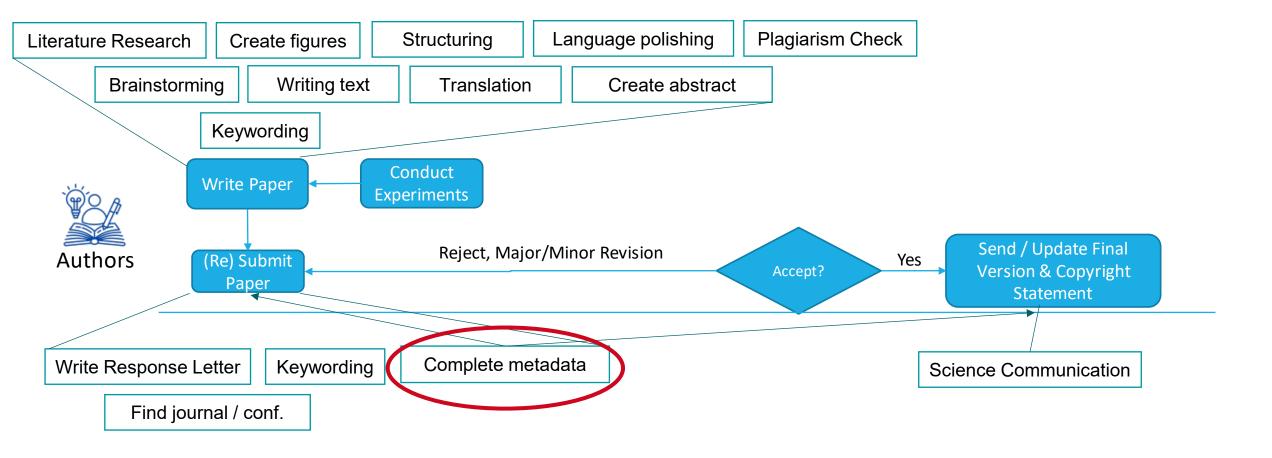








Author Tasks











Metadata Completion - Paper

From Genesis to Maturity: Managing Knowledge Graph Ecosystems Through Life Cycles [Vision]

Sandra Geisler

RWTH Aachen University,Germany geisler@cs.rwth-aachen.de

David Chaves-Fraga Universidade de Santiago de Compostela, Spain david.chaves@usc.es

Maurizio Lenzerini Sapienza University of Rome, Italy lenzerini@diag.uniroma1.it Cinzia Cappiello Politecnico di Milano, Italy cinzia.cappiello@polimi.it

Anastasia Dimou KU Leuven, Belgium anastasia.dimou@kuleuven.be

Anisa Rula University of Brescia, Italy anisa.rula@unibs.it Irene Celino CEFRIEL, Italy irene.celino@cefriel.com

Ana Iglesias-Molina Universidad Politécnica de Madrid, Spain ana.iglesiasm@upm.es

Dylan Van Assche IDLab, Ghent University – imec, Belgium dylan.vanassche@ugent.be

Sascha Welten RWTH Aachen University, Germany welten@dbis.rwth-aachen.de

ABSTRACT

Knowledge graphs (KGs) play a crucial role in the integration and organization of heterogeneous data and knowledge, enabling advanced data analytics and decision-making across various industries. This vision paper addresses critical challenges in managing KGs, emphasizing their relevance in integrating information from disparate sources. We propose the concept of knowledge graph eccosystems and life cycles to systematically manage tasks, e.g., data integration, standardization, continuous updates, efficient querying, and provenance tracking. By adopting our approach, organizations can enhance the accuracy, consistency, and reliability of KGs, thus improving knowledge management, enabling the extraction of valuable insights, and ensuring transparency and accountability.

KEYWORDS

Knowledge Management, Ecosystems, Knowledge Graphs

ACM Reference Format

Maria-Esther Vidal TIB Hannover, Leibniz University Hannover, Germany vidal@l3s.de

1 INTRODUCTION

Nowadays, sharing of high-quality data within data ecosystems is essential to fostering collaboration, efficiency, innovation, and competitiveness among ecosystem stakeholders [39]. However, data ecosystems, e.g., in healthcare and biomedical research, are highly complex, involving a wide range of stakeholders and critical information is often dispersed across multiple, disparate sources. This fragmentation complicates access to the data necessary for generating insights and enabling advanced applications [10]. In such ecosystems, data is inherently heterogeneous, multi-modal, voluminous, and sensitive, presenting challenges related to interoperability and reusability. Additionally, the knowledge needed to describe and contextualize this data is often fragmented, potentially ambiguous, and distributed across extensive ontologies and taxonomies, which often lack mappings between them. A substantial amount of knowledge often remains implicit, captured only in individual expertise and not documented [34]. Therefore, the challenge of harnessing distributed data and knowledge is significant, complex, and extends well beyond the healthcare domain [1].

Knowledge graphs (KGs) provide a robust solution by integrating data from disparate sources into a cohesive data structure providing unified knowledge and data. This integration enables comprehensive insights across data ecosystems. Many articles [27, 42, 48] and









Metadata Completion - Paper

Please complete the metadata of the paper above to describe it



Here is the complete metadata for the paper based on the provided content:

Metadata for the Paper

- Title: From Genesis to Maturity: Managing Knowledge Graph Ecosystems Through Life Cycles [Vision]
- Authors:
 - Sandra Geisler (RWTH Aachen University, Germany)
 - Cinzia Cappiello (Politecnico di Milano, Italy)
 - Irene Celino (CEFRIEL, Italy)
 - David Chaves-Fraga (Universidade de Santiago de Compostela, Spain)
 - Anastasia Dimou (KU Leuven, Belgium)
 - Ana Iglesias-Molina (Universidad Politécnica de Madrid, Spain)
 - Maurizio Lenzerini (Sapienza University of Rome, Italy)
 - Anisa Rula (University of Brescia, Italy)
 - Dylan Van Assche (IDLab, Ghent University imec, Belgium)
 - Sascha Welten (RWTH Aachen University, Germany)
 - Maria-Esther Vidal (TIB Hannover, Leibniz University Hannover, Germany)
- Conference: Very Large Databases Conference 2025 (VLDB'25)
- Publication Venue: Proceedings of the VLDB Endowment, Volume 18

Date: August 25–29, 2025

• Location: London, United Kingdom

Publisher: Association for Computing Machinery (ACM)

• DOI: [Placeholder, since the actual DOI is not visible in the provided text]

• Keywords: Knowledge Management, Knowledge Graphs, Knowledge Graph Ecosystems

Abstract:

Knowledge graphs (KGs) play a crucial role in integrating and organizing heterogeneous data, enabling advanced data analytics and decision-making. This vision paper addresses challenges in managing KGs, proposing the concept of knowledge graph ecosystems and life cycles. It introduces methods for data integration, ontology standardization, continuous updates, efficient querying, and provenance tracking. By adopting a structured life cycle approach, organizations can improve the accuracy, consistency, and reliability of KGs, ultimately enhancing knowledge management, insight extraction, transparency, and accountability.

- ACM Classification:
 - Information systems → Data management systems (Graph-based database models, knowledge base management)
 - Computing methodologies → Artificial intelligence (Knowledge representation and reasoning), Machine learning (Graph-based learning and inference)
 - Theory of computation → Semantics and reasoning (Ontology management and evolution)
 - Applied computing → Life and medical sciences (Computational healthcare informatics)











Metadata Completion - ORCID

retrieve the ORCID of Sandra Geisler, RWTH Aachen



Sandra Geisler's ORCID is 0000-0002-8970-6282. ORCID.ORG She is a junior professor for Data Stream Management and Analysis at RWTH Aachen University since September 2021. DBIS.RWTH-AACHEN.DE Her research interests include data streams, data lakes, data quality, health informatics, and data integration. SCHOLAR.GOOGLEDE





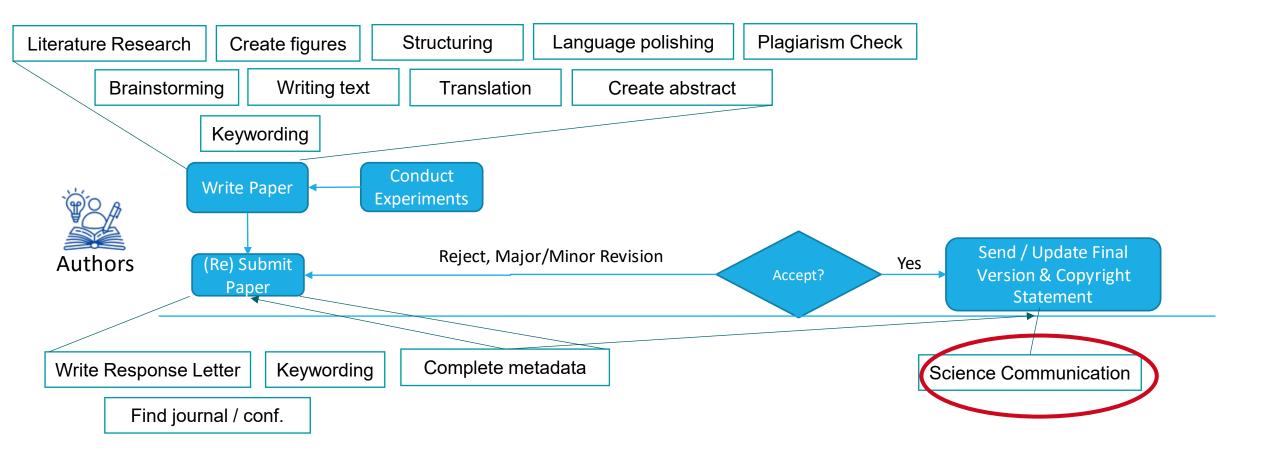








Author Tasks



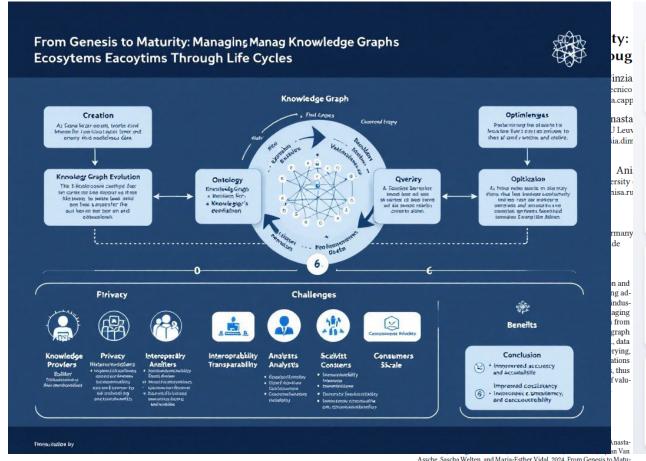




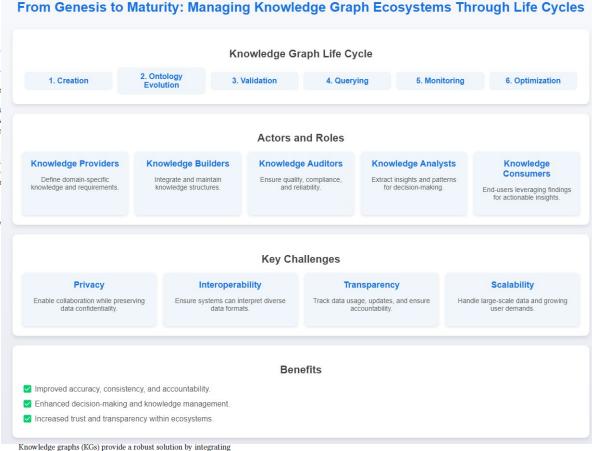




Science Communication – A Visual Abstract



Assche, Sascha Welten, and Maria-Esther Vidal. 2024. From Genesis to Maturity: Managing Knowledge Graph Ecosystems Through Life Cycles [Vision]. In Proceedings of Very Large DataBases Conference 2025 (VLDB '25). ACM, New York, NY, USA. 8 pages. https://doi.org/10.1145/nnnnnnnnnnnnnnnn







data from disparate sources into a cohesive data structure providing

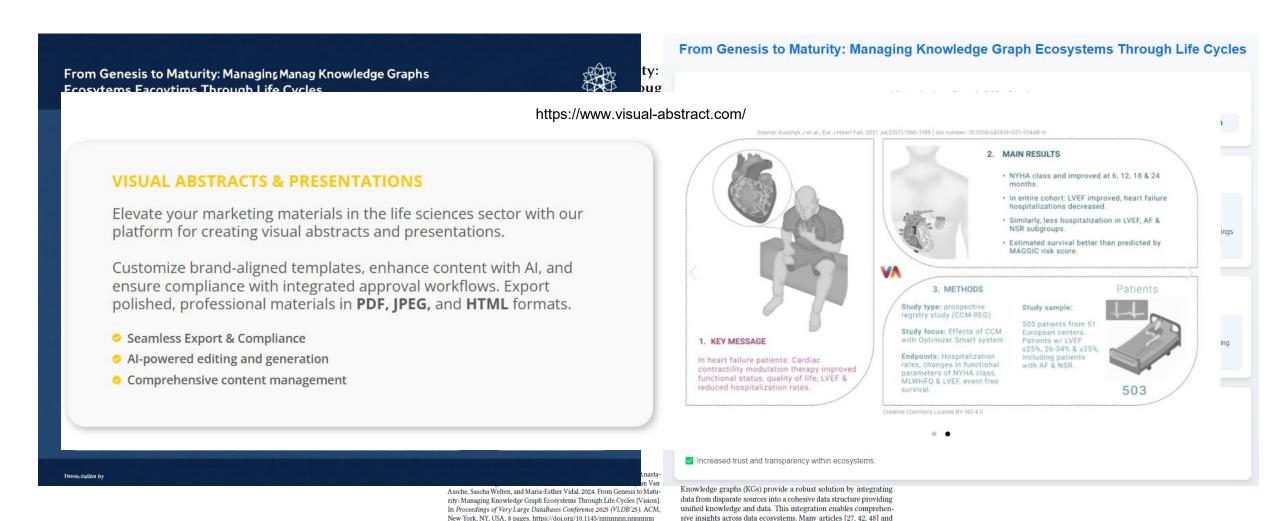
unified knowledge and data. This integration enables comprehen-

sive insights across data ecosystems. Many articles [27, 42, 48] and





Science Communication – A Visual Abstract











Science Communication – A Paper Podcast

From Genesis to Maturity: Managing Knowledge Graph Ecosystems Through Life Cycles [Vision]

Sandra Geisler

RWTH Aachen University,Germany geisler@cs.rwth-aachen.de

David Chaves-Fraga Universidade de Santiago de Compostela, Spain david.chaves@usc.es

Maurizio Lenzerini Sapienza University of Rome, Italy lenzerini@diag.uniroma1.it Cinzia Cappiello Politecnico di Milano, Italy cinzia.cappiello@polimi.it

Anastasia Dimou KU Leuven, Belgium anastasia.dimou@kuleuven.be

Anisa Rula University of Brescia, Italy anisa.rula@unibs.it Irene Celino CEFRIEL, Italy irene.celino@cefriel.com

Ana Iglesias-Molina Universidad Politécnica de Madrid, Spain ana.iglesiasm@upm.es

Dylan Van Assche
IDLab, Ghent University – imec,
Belgium
dylan.vanassche@ugent.be

Sascha Welten RWTH Aachen University, Germany welten@dbis.rwth-aachen.de

ABSTRACT

Knowledge graphs (KGs) play a crucial role in the integration and organization of heterogeneous data and knowledge, enabling advanced data analytics and decision-making across various industries. This vision paper addresses critical challenges in managing KGs, emphasizing their relevance in integrating information from disparate sources. We propose the concept of knowledge graph eccosystems and life cycles to systematically manage tasks, e.g., data integration, standardization, continuous updates, efficient querying, and provenance tracking. By adopting our approach, organizations can enhance the accuracy, consistency, and reliability of KGs, thus improving knowledge management, enabling the extraction of valuable insights, and ensuring transparency and accountability.

KEYWORDS

Knowledge Management, Ecosystems, Knowledge Graphs

ACM Reference Format

Maria-Esther Vidal TIB Hannover, Leibniz University Hannover, Germany vidal@l3s.de

1 INTRODUCTION

Nowadays, sharing of high-quality data within data ecosystems is essential to fostering collaboration, efficiency, innovation, and competitiveness among ecosystem stakeholders [39]. However, data ecosystems, e.g., in healthcare and biomedical research, are highly complex, involving a wide range of stakeholders and critical information is often dispersed across multiple, disparate sources. This fragmentation complicates access to the data necessary for generating insights and enabling advanced applications [10]. In such ecosystems, data is inherently heterogeneous, multi-modal, voluminous, and sensitive, presenting challenges related to interoperability and reusability. Additionally, the knowledge needed to describe and contextualize this data is often fragmented, potentially ambiguous, and distributed across extensive ontologies and taxonomies, which often lack mappings between them. A substantial amount of knowledge often remains implicit, captured only in individual expertise and not documented [34]. Therefore, the challenge of harnessing distributed data and knowledge is significant, complex, and extends well beyond the healthcare domain [1].

Knowledge graphs (KGs) provide a robust solution by integrating data from disparate sources into a cohesive data structure providing unified knowledge and data. This integration enables comprehensive insights across data ecosystems. Many articles [27, 42, 48] and



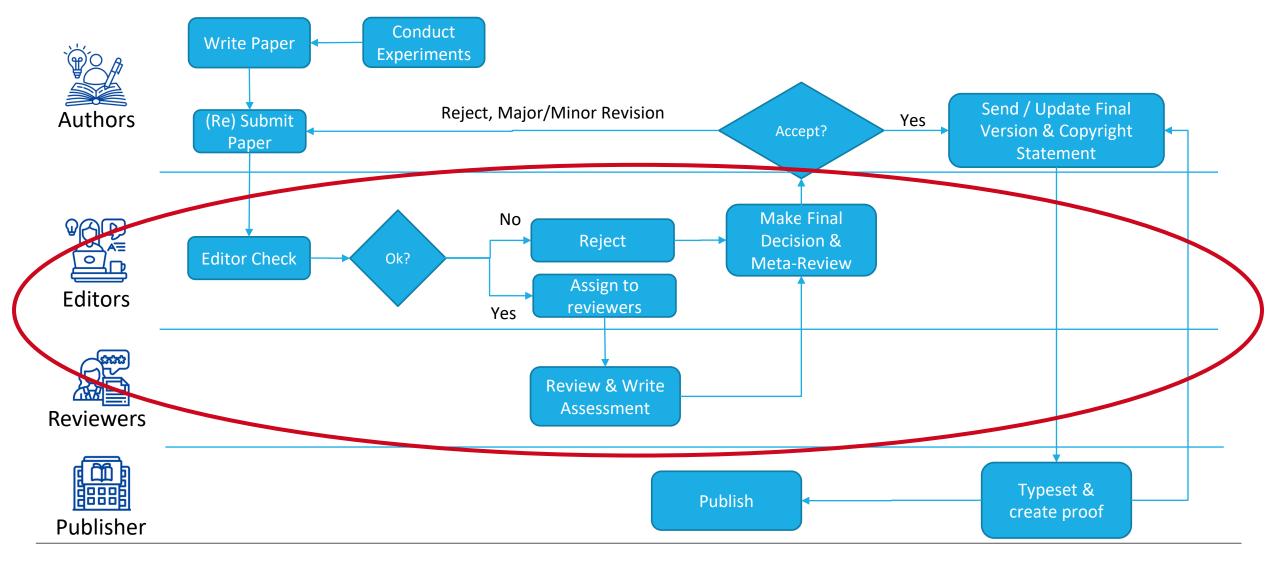








Roles and tasks in the publication process











(Meta) Reviews and Decisions – A No Go for Al



AI use by peer reviewers



Peer reviewers play a vital role in scientific publishing. Their expert evaluations and recommendations guide editors in their decisions and ensure that published research is valid, rigorous, and credible. Editors select peer reviewers primarily because of their indepth knowledge of the subject matter or methods of the work they are asked to evaluate. This expertise is invaluable and irreplaceable. Peer reviewers are accountable for the accuracy and views expressed in their reports, and the peer review process operates on a principle of mutual trust between authors, reviewers and editors. Despite rapid progress, generative AI tools have considerable limitations: they can lack up-to-date knowledge and may produce nonsensical, biased or false information. Manuscripts may also include sensitive or proprietary information that should not be shared outside the peer review process. For these reasons we ask that, while Springer Nature explores providing our peer reviewers with access to safe AI tools, peer reviewers do not upload manuscripts into generative AI tools.

If any part of the evaluation of the claims made in the manuscript was in any way supported by an AI tool, we ask peer reviewers to declare the use of such tools transparently in the peer review report.

As a reviewer, can I use generative AI software tools to draft my review or improve the quality of my review?

Reviewers may not upload confidential ACM submissions into any generative AI or LLM system or any system managed by a third party which does not promise to maintain the confidentiality of that information, since the storage, indexing, learning, and utilization of such submissions may violate the author's right to confidentiality. Likewise, reviewers may not upload any confidential information, such as text, metadata, or identifying information about the authors contained within their draft review into any generative AI or LLM system or any system managed by a third party which does not promise to maintain the confidentiality of that information, as this may also violate the confidentiality and privacy rights and obligations of the author and reviewer.

However, reviewers may use generative AI or LLM systems to improve the quality and readability of reviewer reports for the author, provided any and all parts of the review that would potentially identify the submission, author identities, reviewer identity, or other confidential content is removed prior to uploading into a generative AI or LLM system that does not promise to maintain the confidentiality of that information.

Also, reviewers are permitted to use enterprise versions of LLMs or other third party systems which promise to maintain the confidentiality of uploaded information to draft or improve the readability or quality of their reviews provided that, in doing so, reviewers do not breach confidentiality and that reviewers understand that they are responsible for the content of their reviews, including any bias, inappropriate content, or offensive language contained therein.

Reviewers have an obligation to understand whether the systems that they are using to generate or improve review content will be training on and/or storing that information. If the reviewer is not confident that using a particular system will not violate the ACM Policy on Peer Review or other ACM Publication Policies, they should not use the system.

http://wiley.com/en-de/ai-study

https://www.world-brain-scholar.eu/









Computing Machinery

(Meta) Reviews and Decisions

AI use by peer reviewers

Peer reviewers play a vital role in scientific publishing. The recommendations guide editors in their decisions and ens valid, rigorous, and credible. Editors select peer reviewers depth knowledge of the subject matter or methods of the This expertise is invaluable and irreplaceable. Peer reviewe accuracy and views expressed in their reports, and the pee principle of mutual trust between authors, reviewers and e generative AI tools have considerable limitations: they can may produce nonsensical, biased or false information. Mar sensitive or proprietary information that should not be sha process. For these reasons we ask that, while Springer Nat reviewers with access to safe AI tools, peer reviewers do n generative AI tools.

If any part of the evaluation of the claims made in the mar supported by an AI tool, we ask peer reviewers to declare transparently in the peer review report.

World Brain Scholar

Al solutions for academic publishers and conference organizers.

Experience our Al-powered peer review assistant Eliza, now available for free trial for individual reviewers.

Transforming peer review



Reference analysis

Eliza provides a bibliography analysis of papers under review to reviewers, editors, and conference chairs, offering insight into the recency of cited works, the impact scores of cited journals or conference proceedings, similarity indications, and more, This helps to assess the relevance and credibility of references, and offers a deeper understanding of the paper's scholarly context.



Discuss the paper

Eliza enables in-depth discussions about research papers during the peer review by comprehending and analyzing their content. Engage in interactive Q&A, and let your sparring partner help you write a great review.



Dynamic review feedback

Done with your draft review? Eliza provides high-quality feedback for improvement based on your original review text. This ensures that authors receive constructive and actionable feedback, enhancing the overall quality and effectiveness of the peer review process.

Search paper content Editors can utilize Eliza's vector-based search engine to efficiently

search paper content. This enables quick verification of reviewer

comments and easy location of specific information within the paper.



Executive summaries



Once all peer reviews are submitted, Eliza compiles an executive summary for the editor or conference chair, distilling key points and insights from each review. This overview empowers editors to make well-informed decisions quickly. Pre-written snippets allow decisionmakers to write a constructive decision letter with minimal effort.



Human expertise meets AI capability

Peer reviewers are at the center of our mission. Eliza does not replace human expertise but rather supports it - helping reviewers write more thorough reviews, more efficiently.

draft my review or improve the quality of my

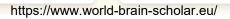
s into any generative AI or LLM system or any se to maintain the confidentiality of that itilization of such submissions may violate the / not upload any confidential information, such as hors contained within their draft review into any a third party which does not promise to maintain violate the confidentiality and privacy rights and

ms to improve the quality and readability of rts of the review that would potentially identify the confidential content is removed prior to not promise to maintain the confidentiality of that

of LLMs or other third party systems which mation to draft or improve the readability or wers do not breach confidentiality and that content of their reviews, including any bias, therein.

e systems that they are using to generate or g that information. If the reviewer is not confident 'olicy on Peer Review or other ACM Publication







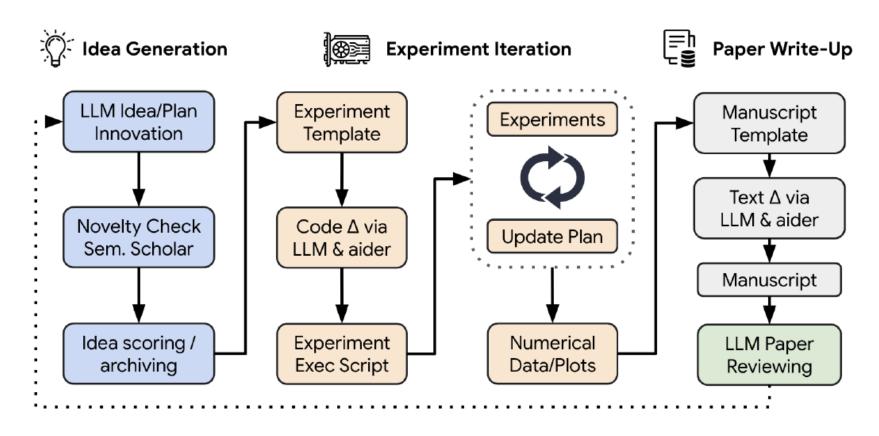








Sakana.ai - The Al Scientist "to perform research independently."



https://sakana.ai/ai-scientist/

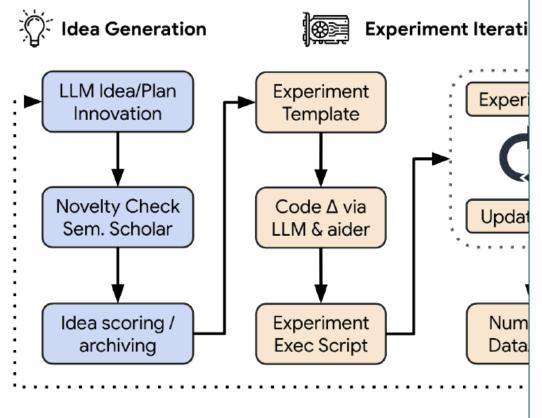








Sakana.ai - The Al Scientist "to perform research independently."



https://sakana.ai/ai-scien

AI-Scientist Generated Preprint

DUALSCALE DIFFUSION: ADAPTIVE FEATURE BAL-ANCING FOR LOW-DIMENSIONAL GENERATIVE MOD-ELS

Anonymous authors Paper under double-blind review

ABSTRACT

This paper introduces an adaptive dual-scale denoising approach for lowdimensional diffusion models, addressing the challenge of balancing global structure and local detail in generated samples. While diffusion models have shown remarkable success in high-dimensional spaces, their application to low-dimensional data remains crucial for understanding fundamental model behaviors and addressing real-world applications with inherently low-dimensional data. However, in these spaces, traditional models often struggle to simultaneously capture both macro-level patterns and fine-grained features, leading to suboptimal sample quality. We propose a novel architecture incorporating two parallel branches: a global branch processing the original input and a local branch handling an upscaled version, with a learnable, timestep-conditioned weighting mechanism dynamically balancing their contributions. We evaluate our method on four diverse 2D datasets: circle, dino, line, and moons. Our results demonstrate significant improvements in sample quality, with KL divergence reductions of up to 12.8% compared to the baseline model. The adaptive weighting successfully adjusts the focus between global and local features across different datasets and denoising stages, as evidenced by our weight evolution analysis. This work not only enhances low-dimensional diffusion models but also provides insights that could inform improvements in higher-dimensional domains, opening new avenues for advancing generative modeling across various applications.

1 INTRODUCTION

Diffusion models have emerged as a powerful class of generative models, achieving state-of-the-art results in various domains such as image synthesis, audio generation, and molecular design Yang et al. (2023). While these models have shown remarkable capabilities in capturing complex data

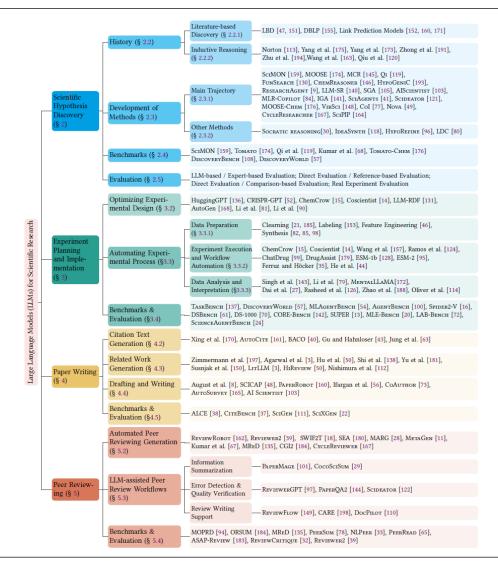








Survey on LLMs for Scientific Research



[Luo et al., arxiv 2025]









Conclusion

Support tasks, but with caution

Ethical issues: responsibility, accountability, bias,...

Training and guidelines needed and requested

Higher outputs stressing the research system

Incentives for researchers









Contact us



JProf. Dr. Sandra Geisler

Principal Investigator IoP

Data Stream Management and Analysis (DSMA)

Phone: +49 241 80-21508

E-Mail: geisler@cs.rwth-aachen.de





Soo-Yon Kim, M.Sc.

Data Steward IoP

Data Stream Management and Analysis

(DSMA)

Phone: +49 241 80-21514

E-Mail: soo-yon.kim@cs.rwth-aachen.de





Prof. Dr. Stefan Decker

Principal Investigator IoP

Databases and Information Systems (DBIS)

Phone: +49 1590 4338009

E-Mail: decker@dbis.rwth-aachen.de











References

Feger, S. S., Dallmeier-Tiessen, S., Schmidt, A., & Woźniak, P. W. (2019, May). Designing for reproducibility: A qualitative study of challenges and opportunities in high energy physics. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14). https://doi.org/10.1145/3290605.3300685

Ganjavi, C., Eppler, M. B., Pekcan, A., Biedermann, B., Abreu, A., Collins, G. S., ... & Cacciamani, G. E. (2024). Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. *bmj*, *384*. doi: https://doi.org/10.1136/bmj-2023-077192

Geisler et al., From Genesis to Maturity: Managing Knowledge Graph Ecosystems Through Life Cycles, Proc. of the VLDB Endowment, Vol. 18, 2025, to appear.

Geisler, S. and Kim, S.-Y. "Unlocking the Potential: LLMs Transforming Research Data Management", In: From Data to Diamonds - Empowering Research with AI and RDM, Aachen, Germany, Nov. 2023, pp. 1–35. https://doi.org/10.18154/RWTH-2023-10498

Geisler, S., Kim, S.-Y., & Decker, S. (2024). FAIRly Intelligent – What LLMs Bring to the Research Data Management Table. In: Helmholtz Metadata Collaboration Conference 2024, Zenodo. https://doi.org/10.5281/zenodo.14054368

Haider, J., Söderström, K. R., Ekström, B., & Rödl, M. (2024). GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation. *Harvard Kennedy School Misinformation Review*, *5*(5).

Hutson, M. (2022). Could Al help you to write your next paper?. Nature, 611(7934), 192-193.

Jo, A. (2023). The promise and peril of generative Al. Nature, 614(1), 214-216.

Kim, S. Y., Hillemacher, S., Decker, S., Rumpe, B., & Geisler, S. (2023). Designing and Implementing Practicable Data Management Plans in Large-Scale Projects. *Bausteine Forschungsdatenmanagement*, (3), 1-12. https://doi.org/10.17192/bfdm.2023.3.8571

Luo, Z., Yang, Z., Xu, Z., Yang, W. and Du, X. 2025. LLM4SR: A Survey on Large Language Models for Scientific Research. https://doi.org/10.48550/arXiv.2501.04306

Moon, J., Gelbich, D., Becker, M., Niemitz, P., & Bergs, T. Predicting fine blanking process signals from sheet metal thickness. *Materials Research Proceedings*, Vol. 41, pp 1436-1445. https://doi.org/10.21741/9781644903131-159

Schaal, Gary S.: Entwicklungspfade. Die Idee wissenschaftlicher Autorenschaft im Kontext technologischer und gesellschaftlicher Transformationsprozesse. - In: Forschung & Lehre, 31 (2024) 9, S. 648-651 - ISSN: 0945-5604

Waldo, J., & Boussard, S. (2025). GPTs and Hallucination. Communications of the ACM, 68(1), 40-45.









Acknowledgment

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany s Excellence Strategy – EXC-2023 Internet of Production – 390621612.









Many thanks! Questions? Comments?

Prof. Dr. Sandra Geisler, Data Stream Management and Analysis, RWTH Aachen University, Aachen, Germany

E-Mail: geisler@informatik.rwth-aachen.de

ORCID-iD: <u>0000-0002-8970-6282</u>

DOI: 10.18154/RWTH-2025-02963



This work is licensed under <u>CC BY-NC-SA 4.0</u>. Excluded from the license are all logos, images and third party materials.



