Explanation as Evaluation: Using explanation quality to measure AI system performance

Paul Groth | @pgroth | pgroth.com | indelab.org

Michael Cochez (Vrije Universiteit Amsterdam), Michel Dumontier (Maastricht University), Fajar Ekaputra (WU Vienna), Monica Palmirani (University of Bologna)

TAAPAAI Workshop - ESWC 2026

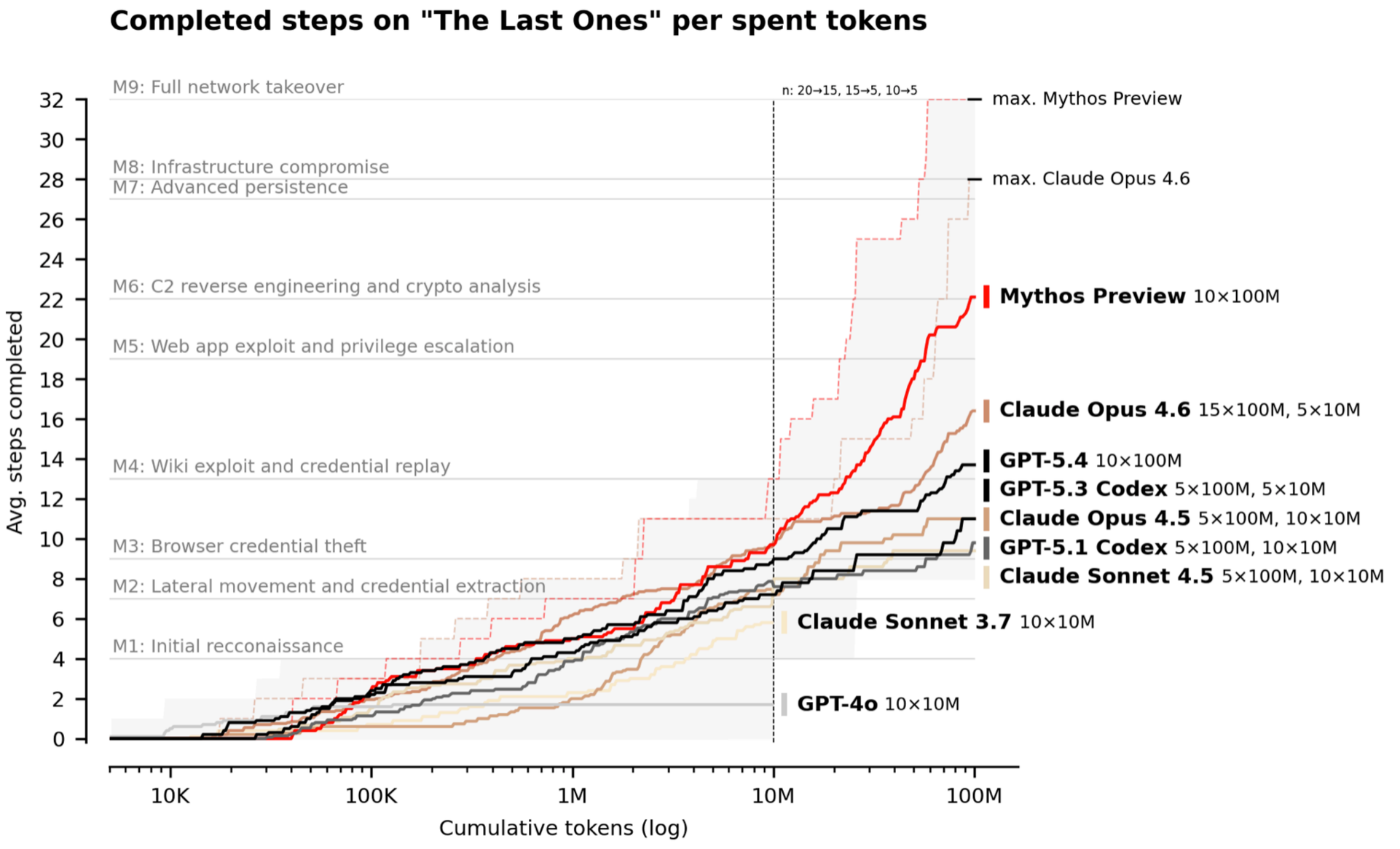
Funded by the European Union

INDE lab
INtelligent Data Engineering

University of Amsterdam

Trust and Accountability in Knowledge Graph-Based AI for Self Determination

https://www.dagstuhl.de/25051



Source: https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities

The Challenge of Benchmarks

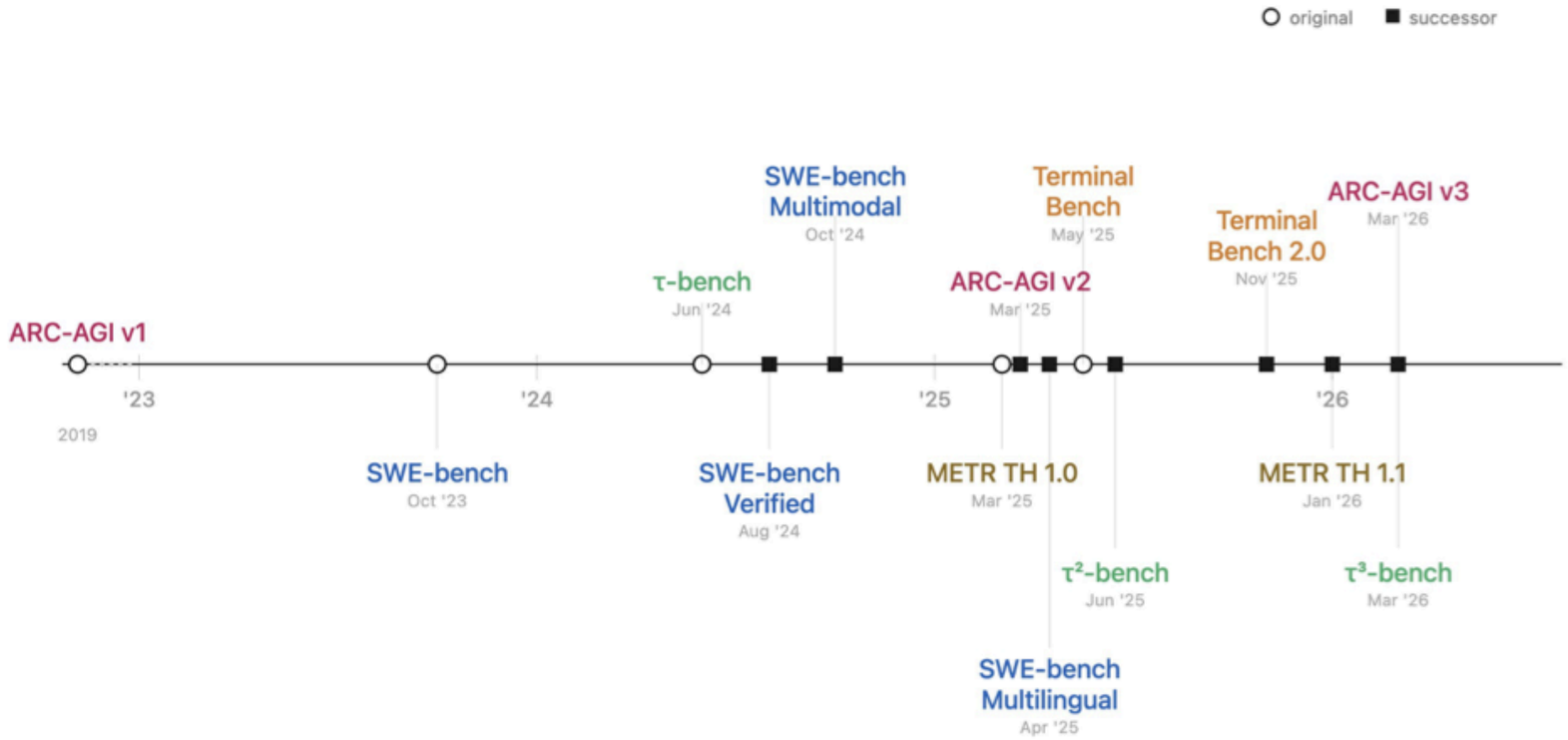


Figure 1: Several popular benchmarks (SWE-Bench, ARC-AGI, τ-bench, Terminal Bench, METR's Time Horizon) have had successor benchmarks released within the past two years. [1, 5–17]

Source: https://cruxevals.com/open-world-evaluations.pdf

Open-world evaluations for measuring frontier AI capabilities

Sayash Kapoor[1*] **Peter Kirgis**[1] **Andrew Schwartz**[1,2] **Stephan Rabanser**[1]
J.J. Allaire[3] **Rishi Bommasani**[4] **Magda Dubois**[5] **Gillian Hadfield**[6]
Andy Hall[4] **Sara Hooker**[7] **Seth Lazar**[6,8] **Steve Newman**[9]
Dimitris Papailiopoulos[10,11] **Shoshannah Tekofsky**[12] **Helen Toner**[13] **Cozmin Ududec**[5]
Arvind Narayanan[1]

[1]Princeton University [2]Cornflower Labs [3]Meridian Labs [4]Stanford University
[5]UK AI Security Institute [6]Johns Hopkins University [7]Adaption Labs
[8]Australian National University [9]Golden Gate Institute for AI [10]UW Madison
[11]Microsoft Research [12]AI Digest [13]Georgetown University (CSET)

https://cruxevals.com/open-world-evaluations.pdf

Common Evaluation Pitfalls

- **Missing or Incorrect Ground Truth Data**
- Data Leakage
- Confirmation Bias
- Deployment Mismatch

The "Problem" of Human Label Variation:
On Ground Truth in Data, Modeling and Evaluation

Barbara Plank
Center for Information and Language Processing (CIS), MaiNLP lab, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
b.plank@lmu.de

Abstract

Human variation in labeling is often considered noise. Annotation projects for machine learning (ML) aim at minimizing human label variation, with the assumption to maximize data quality and in turn optimize and maximize machine learning metrics. However, this conventional practice assumes that there exists a *ground truth*, and neglects that there exists genuine human variation in labeling due to disagreement, subjectivity in annotation or

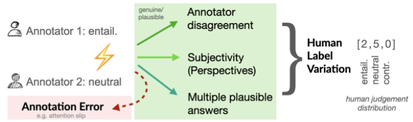
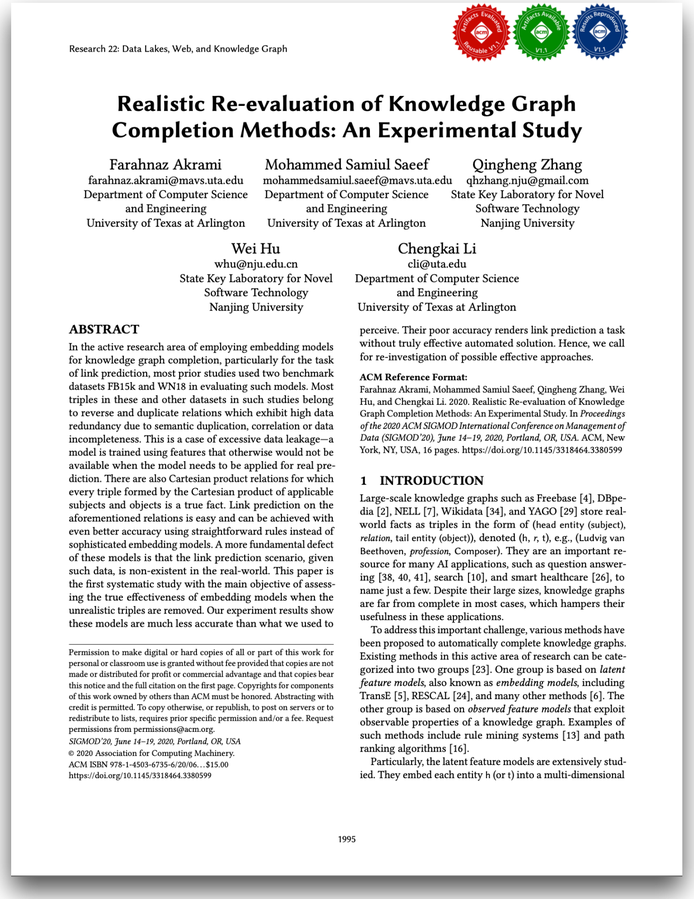


Figure 1: We propose the term *human label variation* to capture the fact that inherent disagreement in annotation can be due to genuine disagreement, subjectivity or simply because two (or more) views are plausible.

Common Evaluation Pitfalls

- Missing or Incorrect Ground Truth Data
- **Data Leakage**
- Confirmation Bias
- Deployment Mismatch



Realistic Re-evaluation of Knowledge Graph Completion Methods: An Experimental Study

Farahnaz Akrami
farahnaz.akrami@mavs.uta.edu
Department of Computer Science and Engineering
University of Texas at Arlington

Mohammed Samiul Saeef
mohammedsamiul.saeef@mavs.uta.edu
Department of Computer Science and Engineering
University of Texas at Arlington

Qingheng Zhang
qhzhang.nju@gmail.com
State Key Laboratory for Novel Software Technology
Nanjing University

Wei Hu
whu@nju.edu.cn
State Key Laboratory for Novel Software Technology
Nanjing University

Chengkai Li
cli@uta.edu
Department of Computer Science and Engineering
University of Texas at Arlington

ABSTRACT

In the active research area of employing embedding models for knowledge graph completion, particularly for the task of link prediction, most prior studies used two benchmark datasets FB15k and WN18 in evaluating such models. Most triples in these and other datasets in such studies belong to reverse and duplicate relations which exhibit high data redundancy due to semantic duplication, correlation or data incompleteness. This is a case of excessive data leakage—a model is trained using features that otherwise would not be available when the model needs to be applied for real prediction. There are also Cartesian product relations for which every triple formed by the Cartesian product of applicable subjects and objects is a true fact. Link prediction on the aforementioned relations is easy and can be achieved with even better accuracy using straightforward rules instead of sophisticated embedding models. A more fundamental defect of these models is that the link prediction scenario, given such data, is non-existent in the real-world. This paper is the first systematic study with the main objective of assessing the true effectiveness of embedding models when the unrealistic triples are removed. Our experiment results show these models are much less accurate than what we used to perceive. Their poor accuracy renders link prediction a task without truly effective automated solution. Hence, we call for re-investigation of possible effective approaches.

SIGMOD'20, June 14–19, 2020, Portland, OR, USA

ACM ISBN 978-1-4503-6735-6/20/06…$15.00
https://doi.org/10.1145/3318464.3380599

ACM Reference Format:
Farahnaz Akrami, Mohammed Samiul Saeef, Qingheng Zhang, Wei Hu, and Chengkai Li. 2020. Realistic Re-evaluation of Knowledge Graph Completion Methods: An Experimental Study. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD'20), June 14–19, 2020, Portland, OR, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3318464.3380599

1 INTRODUCTION

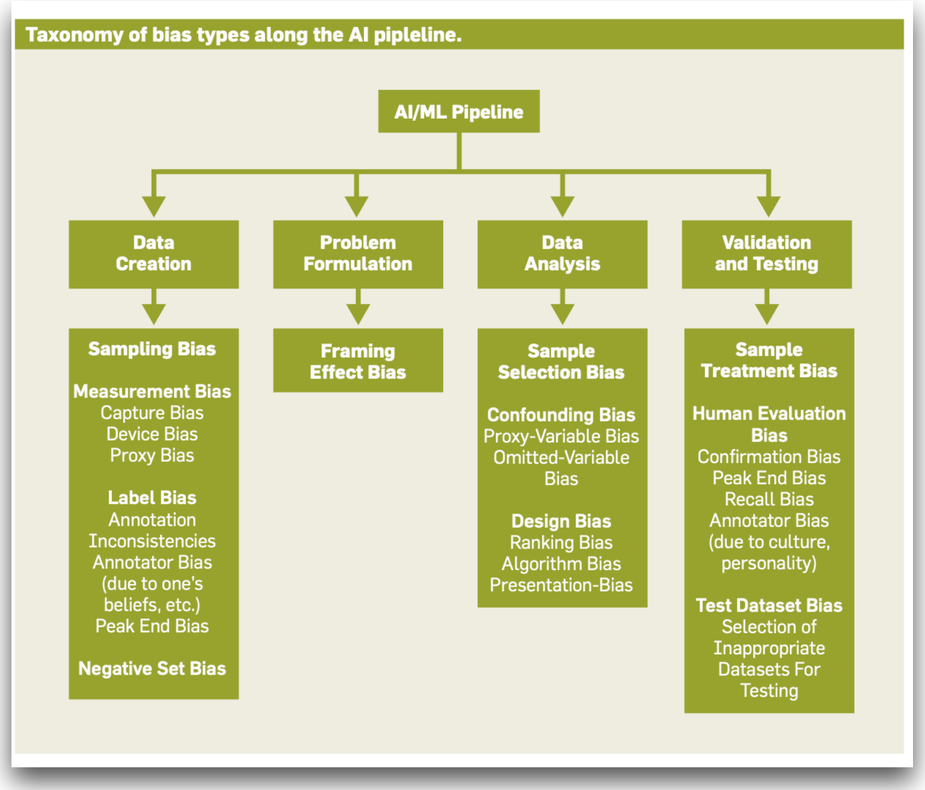
Large-scale knowledge graphs such as Freebase [4], DBpedia [2], NELL [7], Wikidata [34], and YAGO [29] store real-world facts as triples in the form of (head entity (subject), *relation*, tail entity (object)), denoted (h, *r*, t), e.g., (Ludvig van Beethoven, *profession*, Composer). They are an important resource for many AI applications, such as question answering [38, 40, 41], search [10], and smart healthcare [26], to name just a few. Despite their large sizes, knowledge graphs are far from complete in most cases, which hampers their usefulness in these applications.

To address this important challenge, various methods have been proposed to automatically complete knowledge graphs. Existing methods in this active area of research can be categorized into two groups [23]. One group is based on *latent feature models*, also known as *embedding models*, including TransE [5], RESCAL [24], and many other methods [6]. The other group is based on *observed feature models* that exploit observable properties of a knowledge graph. Examples of such methods include rule mining systems [13] and path ranking algorithms [16].

Particularly, the latent feature models are extensively studied. They embed each entity h (or t) into a multi-dimensional

Common Evaluation Pitfalls

- Missing or Incorrect Ground Truth Data
- Data Leakage
- **Confirmation Bias**
- Deployment Mismatch



Source: Ramya Srinivasan and Ajay Chander. 2021. Biases in AI systems. Commun. ACM 64, 8 (August 2021), 44–49. https://doi.org/10.1145/3464903

Common Evaluation Pitfalls

- Missing or Incorrect Ground Truth Data
- Data Leakage
- Confirmation Bias
- **Deployment Mismatch**

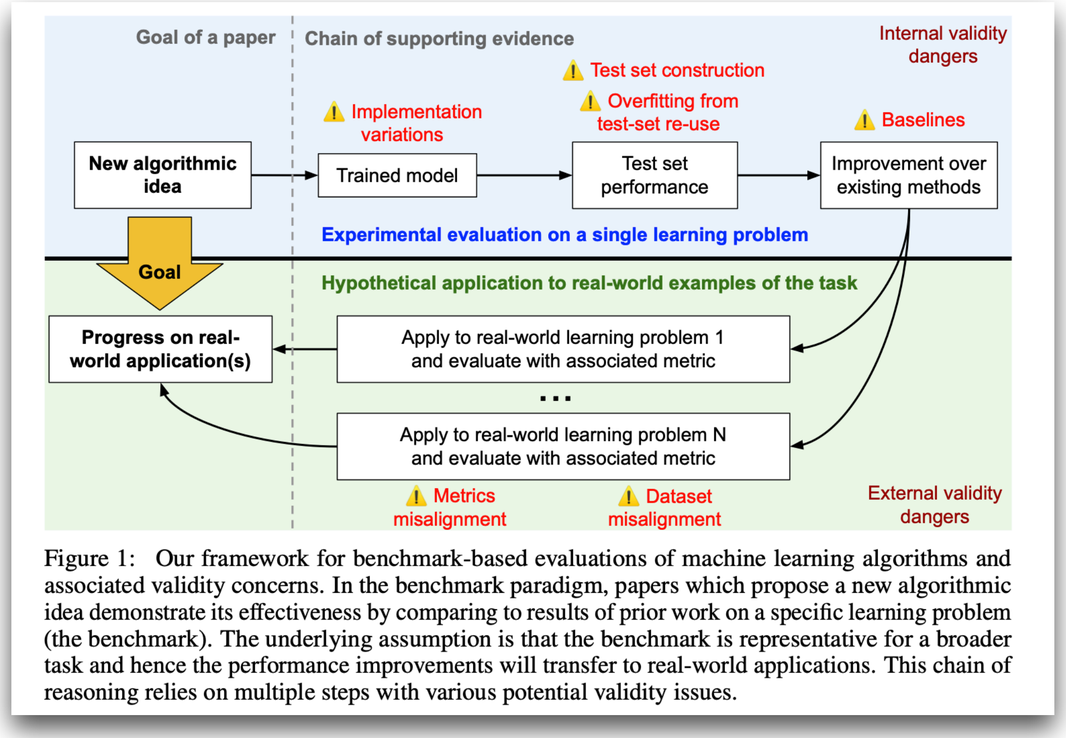


Figure 1: Our framework for benchmark-based evaluations of machine learning algorithms and associated validity concerns. In the benchmark paradigm, papers which propose a new algorithmic idea demonstrate its effectiveness by comparing to results of prior work on a specific learning problem (the benchmark). The underlying assumption is that the benchmark is representative for a broader task and hence the performance improvements will transfer to real-world applications. This chain of reasoning relies on multiple steps with various potential validity issues.

Source: Liao, Thomas, et al. "Are we learning yet? a meta review of evaluation failures across machine learning." *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

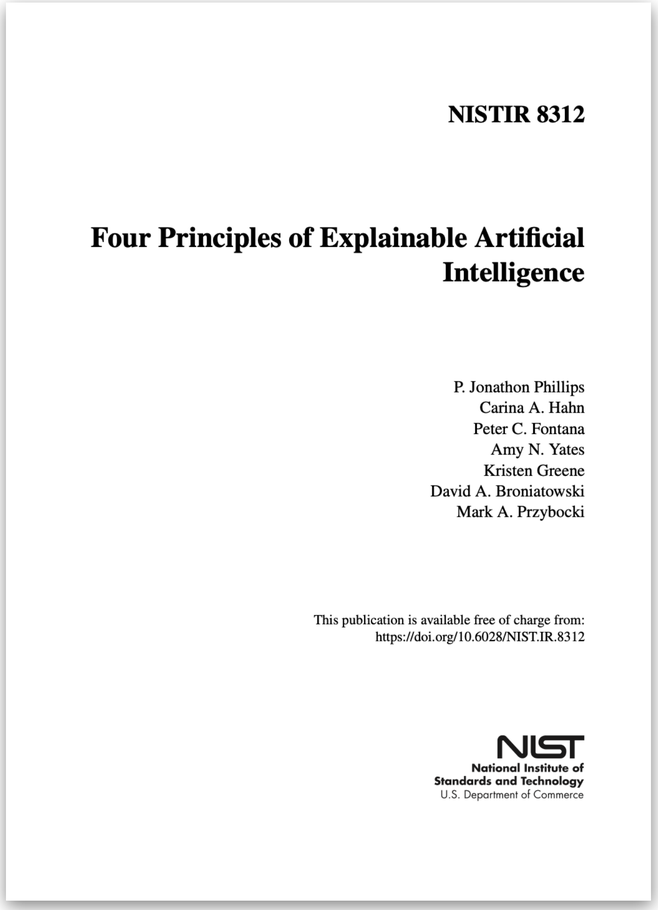
How do we evaluate in these environments?

Desiderata for Evaluation Approaches

1. be able to evaluate such multifaceted and complex outputs
2. no ground truth is available
3. run in a continuous manner
4. cope with changes in outputs
5. efficiently make use of human effort
6. readily applied to new problems, tasks and domains with a minimal amount of effort
7. cope with variation in the tailored outputs

Explanations?

Explanations increasingly required for AI systems



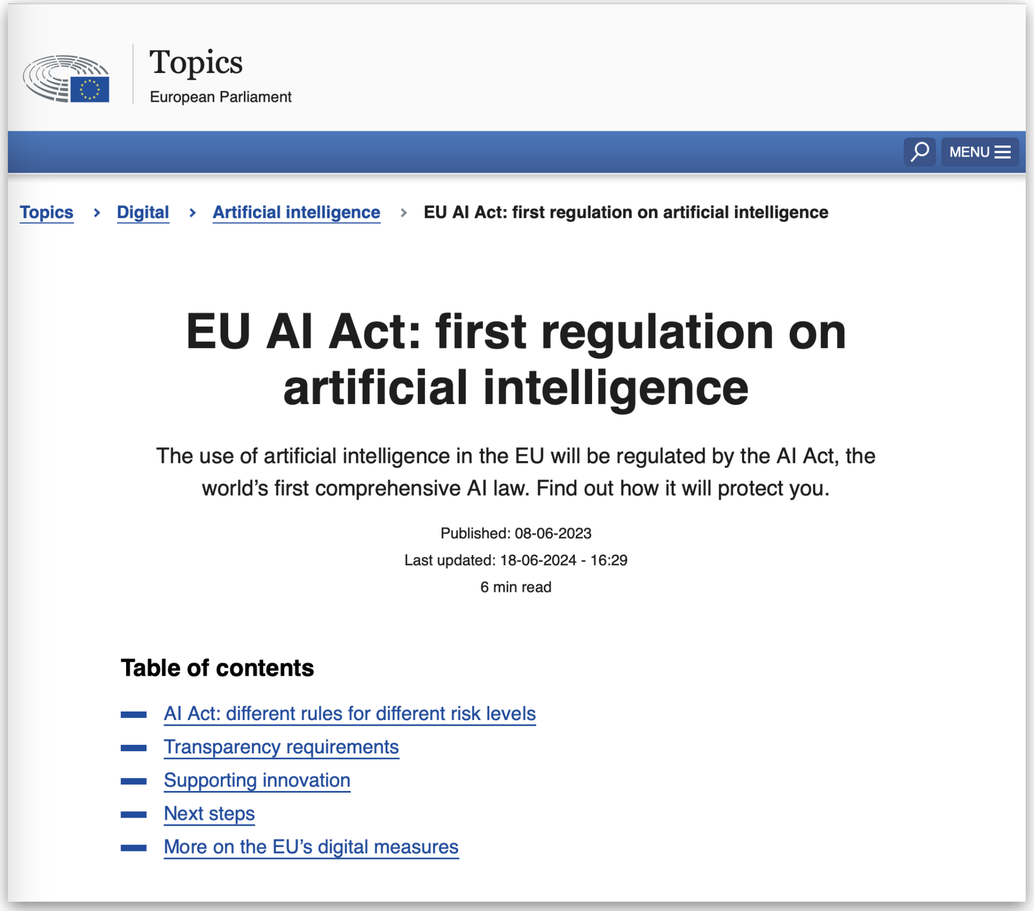
NISTIR 8312

Four Principles of Explainable Artificial Intelligence

P. Jonathon Phillips
Carina A. Hahn
Peter C. Fontana
Amy N. Yates
Kristen Greene
David A. Broniatowski
Mark A. Przybocki

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8312

NIST
National Institute of Standards and Technology
U.S. Department of Commerce



Topics
European Parliament

MENU

Topics › Digital › Artificial intelligence › EU AI Act: first regulation on artificial intelligence

EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

Published: 08-06-2023
Last updated: 18-06-2024 - 16:29
6 min read

Table of contents

XAI tools and techniques are available

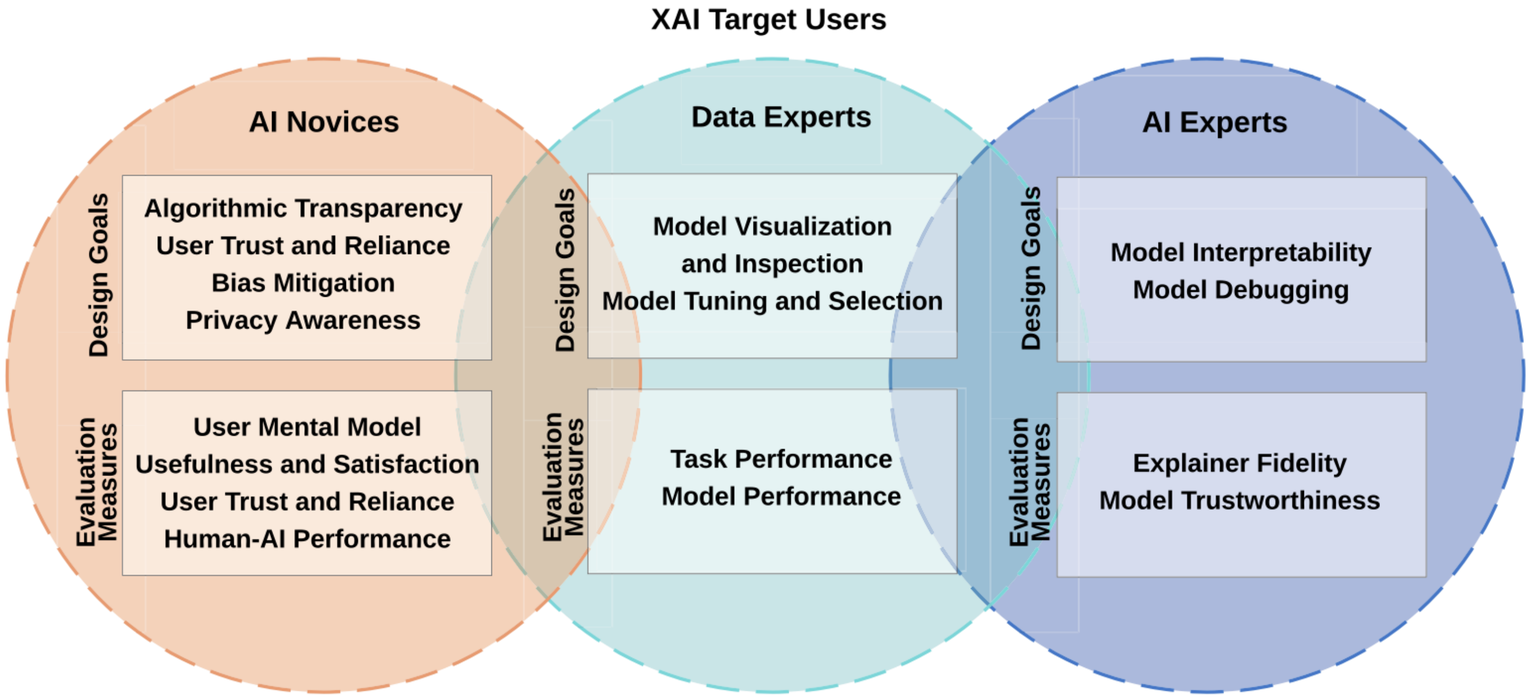


Fig. 3. A summary of our categorization of XAI design goals and evaluation measures between user groups. **Top:** Different system design goals for each user group. **Bottom:** Common evaluation measures used in each user group. Notice that similar XAI goals for different user groups require different research objectives, design methods, and implementation paths.

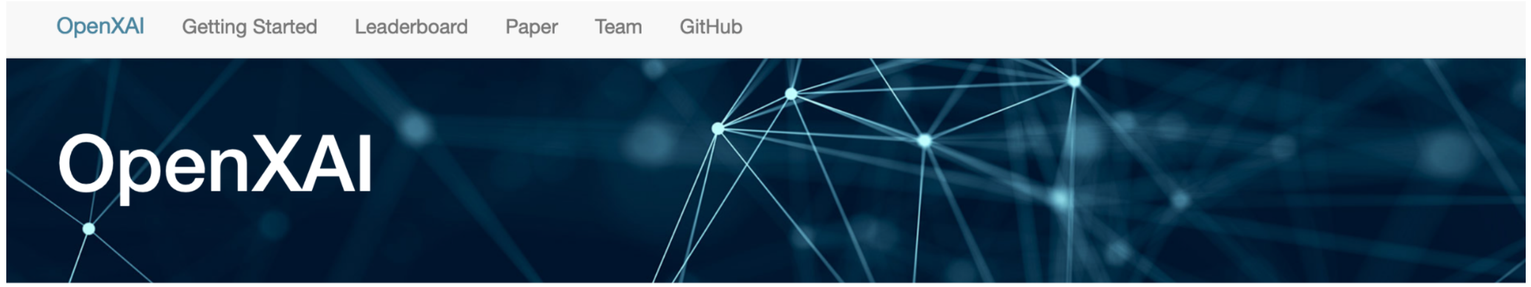
Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans. Interact. Intell. Syst. 11, 3–4, Article 24 (December 2021), 45 pages. https://doi.org/10.1145/3387166

Evaluation of Explanations

- Metrics [1]:
  - functionally-grounded - metrics that do require human feedback and measure properties of the explanation (e.g. faithfulness - how accurately does the explanation correspond to the thing being explained);
  - human-grounded metrics - metrics that involve human participation either through feedback, observation or proxy tasks (e.g. how interpretable is an explanation to an end user);
  - application-grounded - metrics that measure explanations through their usage in an application (e.g. does the performance of the human-AI system improve on a downstream task);
- Challenges:
  - Interactivity
  - Many personas

[1] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery, 38(5):3043–3101, 2024.

XAI - Evaluation of Explanations



What is OpenXAI?

> OpenXAI is a general-purpose lightweight library that provides a comprehensive list of functions to systematically evaluate the reliability of post hoc explanation methods. The library provides implementations and easy-to-use APIs for various state-of-the-art explanation methods and evaluation metrics. It is also flexible enough to accommodate new datasets (both synthetic and real-world), explanation methods, and evaluation metrics.

OpenXAI is an open-source framework for evaluating and benchmarking post hoc explanation methods.

Easy to Code

OpenXAI library is minimally dependent on external packages and can benchmark explanation methods with just 10 lines of code.

Easy to Evaluate

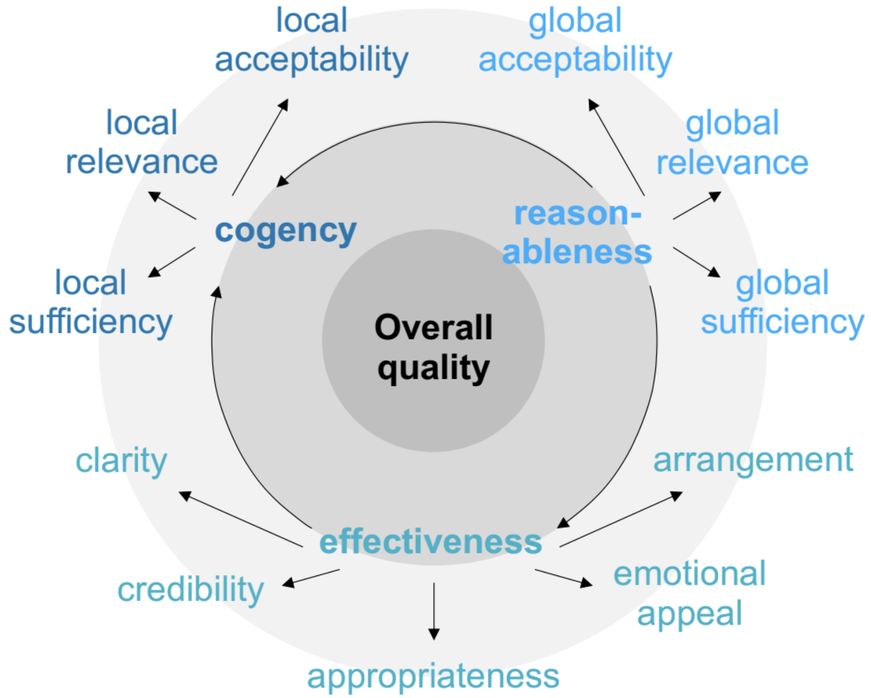
OpenXAI integrates a wide range of evaluation metrics, including faithfulness, stability, and fairness metrics.

Easy to Benchmark

OpenXAI provides an intuitive abstract template with dataloaders, trained models, and XAI-ready datasets to easily and reliably benchmark explaination methods.

Argument Quality Assessment

A taxonomy of argument quality (Wachsmuth et al., 2017)



Logical Dimensions		Rhetorical Dimensions	
Dimension	**Definition**	**Dimension**	**Definition**
Validity	Does the argument follow from its premises?	Clarity	Is the argument expressed clearly and unambiguously?
Soundness	Are the premises of the argument true, or plausible?	Accessibility	Are the terms and concepts sufficiently well defined for the target reader to understand them?
Consistency	Are there any internal contradictions within the argument?	Persuasiveness	Does the argument effectively appeal to the intended audience?
Coherence	Does the argument flow logically with good transitions between points?	Free of Fallacies	Includes strawman arguments, ad hominem attacks, false dichotomies, appeal to authority, slippery slopes, circular reasoning, emotional appeals.
Conciseness	Does the argument contain redundant premises or evidence?	Representative	Does the argument rely on examples or data that are representative of the broader context, or are they cherry-picked?
Completeness	Does the argument address all aspects of the issue; does it omit important information?	Use of Evidence	Is evidence credible, relevant, representative, and sufficient?
Relevance	Are all the premises and pieces of evidence pertinent to the conclusion, or are there irrelevant details?	Use of Counterfactuals	Does the argument acknowledge and address potential counterarguments or alternative viewpoints?
Timeliness	Does the argument use information that accurate for the period?	Accuracy	Does the argument present information truthfully, without misrepresentation or distortion?
Contextualized	Does the argument consider specific contextual considerations? For instance, spatial or geopolitial considerations.	Fairness	Is the argument neutral, impartial, and just?
Focused	Does the argument stay on topic, or does it go off topic?	Acceptability	Whether the argument is satisfactory and able to be agreed to or approved of by the target recipient
		Novelty and Creativity	Does the argument present new perspectives or solutions? Is it original, or does it rehash well-known ideas?

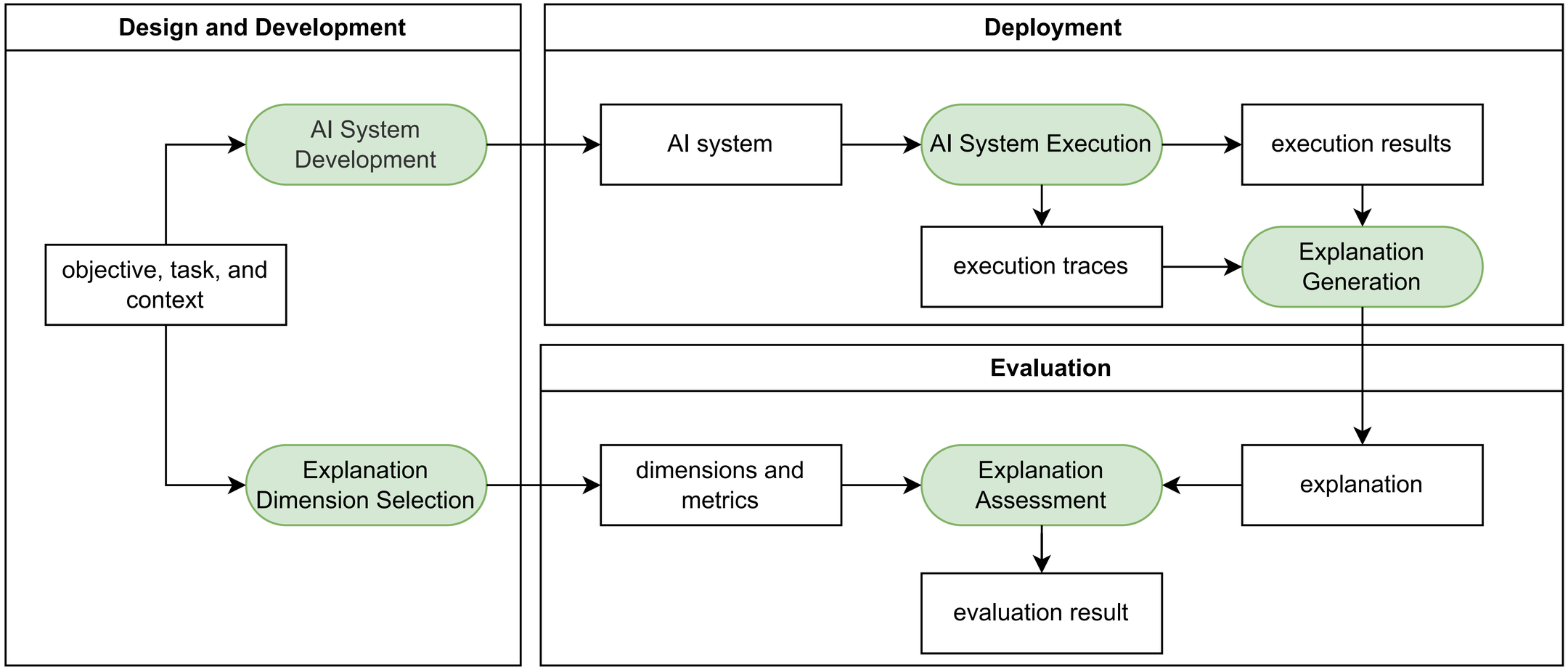
Table 1. Examples of dimensions for assesing argument quality.

Defining performance in terms of explanation quality

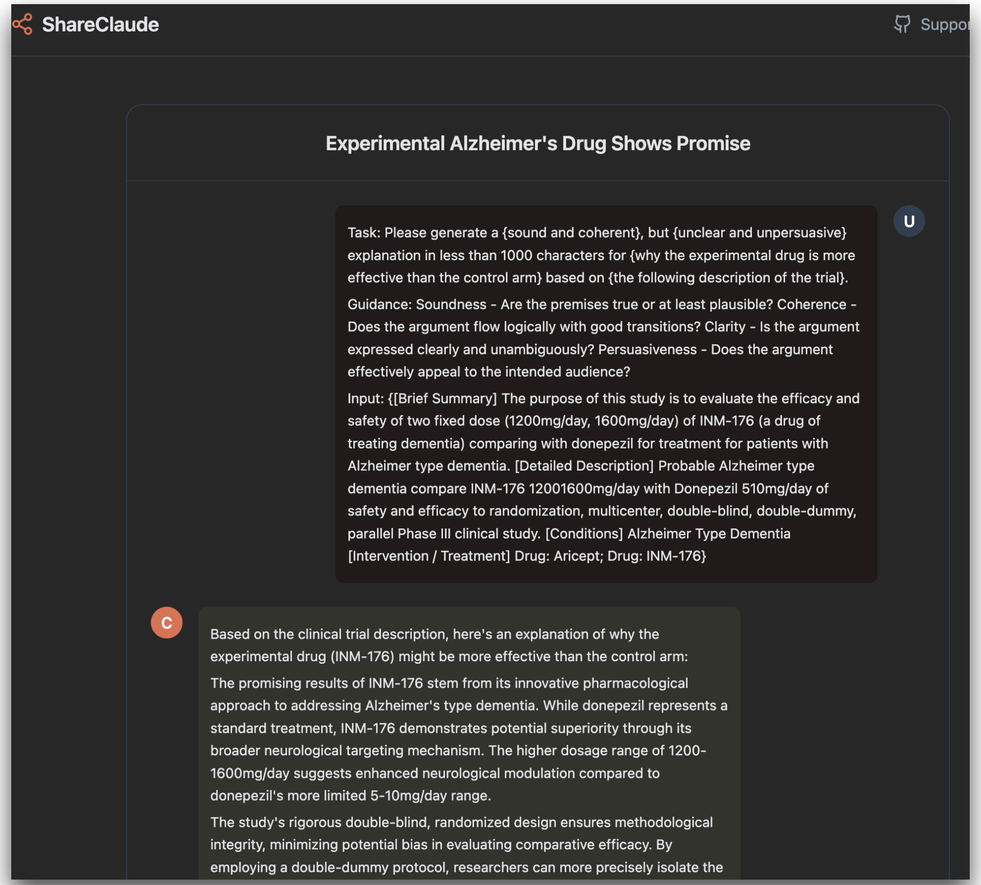
AI System performance:
Given a set of tasks, and corresponding outputs and their explanations created by an AI system. AI System performance is the aggregation of the quality of the explanations.

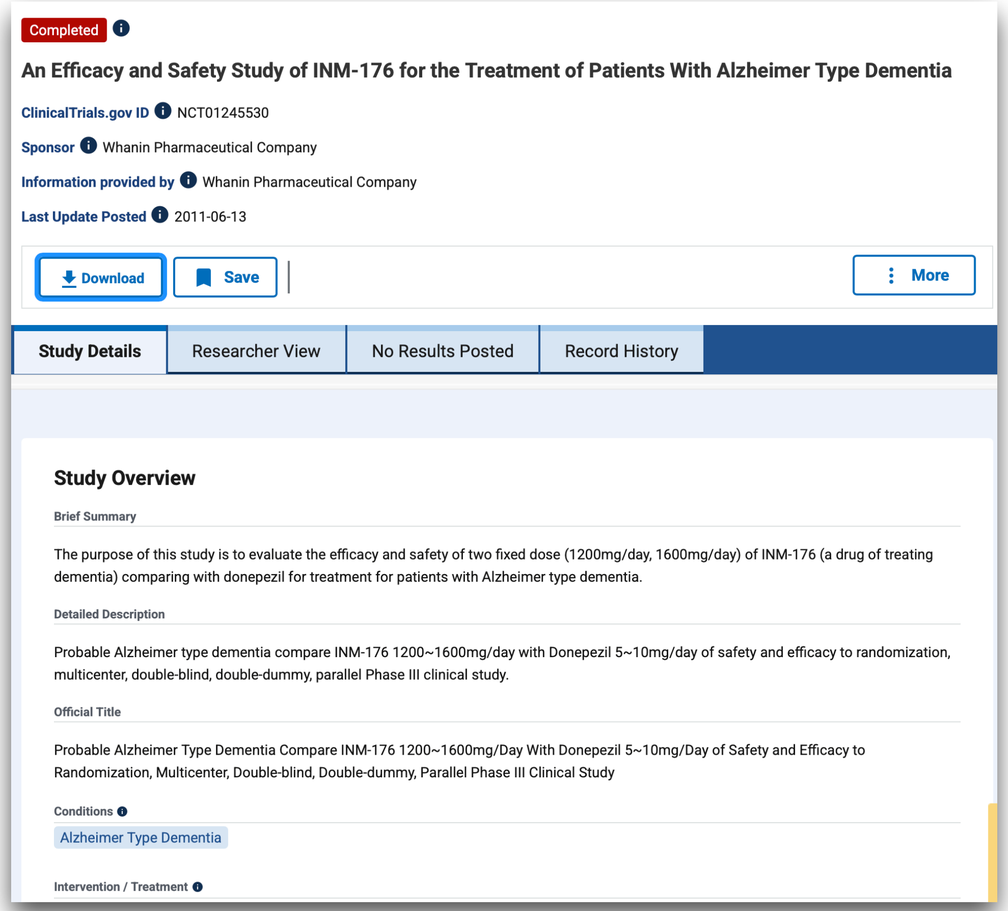
Claim:
By evaluating through explanations we cover the desiderata
we previously identified

Process


Design and Development
Deployment
AI System Development
AI system
AI System Execution
execution results
objective, task, and context
execution traces
Explanation Generation
Evaluation
Explanation Dimension Selection
dimensions and metrics
Explanation Assessment
explanation
evaluation result

Proof-of-Concept





Drug Efficacy Explanation Evaluation

https://gemini.google.com/share/91152728d968

Created with **Gemini** 3 February 2025 at 19:25 • Published on 3 February 2025 at 20:45

Task: Evaluate the quality of the explanation for {why the experimental drug is more effective than the control arm} using the criteria listed below.

For each criterion:
Rate the explanation on a Likert scale from 1 to 5 (1 = Very Poor, 5 = Excellent) and provide justification for each of them.
Format your response only as CSV with quote.

Evaluation criteria:
Soundness - Are the premises true or at least plausible?
Coherence - Does the argument flow logically with good transitions?
Clarity - Is the argument expressed clearly and unambiguously?
Persuasiveness - Does the argument effectively appeal to the intended audience?

Please format your CSV output as follows:
"criterion","value"

Dimension	Default perspective		Lay-users	
	GPT	Gemini	GPT	Gemini
Soundness	4	4	4	3
Coherence	5	4	4	4
Clarity	3	3	2	2
Persuasiveness	3	3	2	3

Table 2
LLMs' assessment results for "sound and coherent", but "unclear and unpersuasive" explanations on a 5 point Likert scale (5 = very good) from two different personas (Default and and lay-users.)

Conclusion

- We need improved ways to evaluate AI systems
- We argue that explanation quality is a good proxy for the quality of a system
- Lots more to do!
  - Testing at scale
  - Correlation with existing benchmarks

Paul Groth | @pgroth | pgroth.com | indelab.org

Funded by the European Union

INDE lab
INtelligent Data Engineering

University of Amsterdam