Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Paradigms of AI Evaluation: Mapping Goals, Methodologies and Culture

John Burden[1] , **Marko Tešić**[1] , **Lorenzo Pacchiardi**[1] and **José Hernández-Orallo**[1,2]
[1]Leverhulme Centre for the Future of Intelligence, University of Cambridge
[2]VRAIN, Universitat Politècnica de València
{jjb205, mt961, lp666}@cam.ac.uk, jorallo@upv.es

Abstract

Research in AI evaluation has grown increasingly complex and multidisciplinary, attracting researchers with diverse backgrounds and objectives. As a result, divergent evaluation *paradigms* have emerged, often developing in isolation, adopting conflicting terminologies, and overlooking each other's contributions. This fragmentation has led to insular research trajectories and communication barriers both among different paradigms and with the general public, contributing to unmet expectations for deployed AI systems. To help bridge this insularity, in this paper we survey recent work in the AI evaluation landscape and identify six main paradigms. We characterise major recent contributions within each paradigm across key dimensions related to their goals, methodologies and research cultures. By clarifying the unique combination of questions and approaches associated with each paradigm, we aim to increase awareness of the breadth of current evaluation approaches and foster cross-pollination between different paradigms. We also identify potential gaps in the field to inspire future research directions.

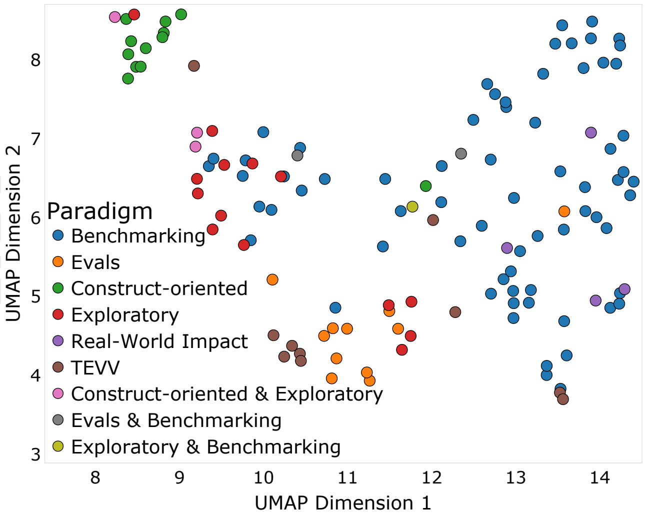

Figure 1: Two-dimensional projection of the surveyed papers based on our framework's dimensions. The coordinates were obtained using UMAP projection [McInnes *et al.*, 2018] of the Jaccard-distance matrix. Each point represents a surveyed paper, whose colour indicates the paradigms it belongs to. We find clusters of papers corresponding to different paradigms. The interactive plot and the code to reproduce it are available here.

1 Introduction

In recent years, Artificial Intelligence (AI) has advanced rapidly and gained public prominence. In particular, general-purpose AI systems, such as Large Language Models (LLMs), have become widely deployed for real-world applications across various sectors [Maliugina, 2024]. As AI adoption grows, so does the need to understand AI systems capabilities [Hernández-Orallo, 2017] and the risks they pose [Hendrycks *et al.*, 2021b] to ensure responsible deployment.

At the same time, the technical expertise required to effectively use state-of-the-art AI models has decreased, enabling a broader range of researchers and practitioners to engage in AI evaluation. This has brought new perspectives to the field, with evaluators from a multitude of disciplines beyond AI, including Cognitive Science, Psychology, Economics, Social Sciences, Software and Safety Engineering contributing to the diversity of methodologies and insights.

While the influx of varied expertise is essential to evaluating the complex and multifaceted impact of modern AI systems on society, it has created a fragmented AI evaluation landscape, marked by a lack of standardisation and limited cross-collaboration. As a result, research efforts have become insular and fail to incorporate advancements from other communities. Communication challenges arise, with key terms often conveying distinct connotations in different communities. For instance, AI developers and regulators rely on evaluations to determine if a system is safe to deploy [Phuong *et al.*, 2024]. In contrast, AI adopters are more concerned with understanding if a system can automate specific tasks within their organisation [Asgari *et al.*, 2024]. Both communities, however, refer to "capability evaluations".

Beyond scientific research and industry practices, underappreciating the breadth of AI evaluation may impact policy, as various policy initiatives ubiquitously rely on "eval-

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

uation”. For instance, the EU AI Act states that developers of a general-purpose AI model with “high impact capabilities evaluated on the basis of appropriate technical tools and methodologies” [European Union, 2024, Art. 51(1)] are subject to additional requirements, such as “perform[ing] model evaluation [...] with a view to identifying and mitigating systemic risks” [European Union, 2024, Art. 55(1)]; thus, “evaluation” refers, in the same document, both to measurement of capabilities and identification of risks. Ongoing efforts[1] identifying concrete evaluation tools to satisfy those requirements should consider the full spectrum of AI evaluation practices and understand their terminological and methodological differences, to select the appropriate set of techniques for distinct regulatory purposes.

To address this, in this paper we survey the landscape of AI evaluation by collecting 125+ papers representative of the various existing approaches[2] and identify six main *paradigms*. We define an AI evaluation paradigm as a conceptual framework that groups together studies with similar methodologies, evaluation goals and underlying assumptions that influence the collection of evidence. These paradigms are not rigid, universally agreed-upon categories but rather our attempt to map and structure the AI evaluation landscape. By delineating them, we aim to provide a useful perspective for researcher and practitioners, helping them navigate and critically assess different evaluation approaches.

To operationalise this framework, we systematically annotated the collected papers according to a series of dimensions of analysis related to goals, methodologies and culture[3]. This annotation allows us to clarify the questions and approaches adopted by each paradigm, as well as how they differ from one another. Our contributions are twofold: (1) we provide researchers with a broad overview of existing AI evaluation practices and identify main AI evaluation paradigms; (2) by highlighting differences between paradigms, we facilitate comparison, collaboration and knowledge exchange across research communities.

The paper is organised as follows: Sec. 2 introduces the dimensions of analysis, while Sec. 3 describes the identified paradigms. Sec. 4 discusses the role these paradigms play within the AI evaluation ecosystem. In Sec. 5, we identify gaps in the AI evaluation landscape and explore ways to leverage existing methodologies to advance the field. Finally, Secs. 6 and 7 discuss the limitations of our survey and present our conclusions.

1.1 Scope

For our analysis, we define AI evaluation as *the process of measuring and anticipating the behavioural properties of AI systems and their societal impact to inform decisions about their use*. In our survey, we consider work on the evaluation of any kind of AI system, component or algorithm. However, we interpret our definition as excluding explainability and mechanistic interpretability, which aim to get an understanding of the inner workings of AI systems. While such an understanding may be used to anticipate its behaviour [Casper *et al.*, 2024], we decide to concentrate on work directly measuring the behavioural properties. We refer interested readers to existing high-quality surveys on interpretability and explainability [Yang *et al.*, 2023b; Bereska and Gavves, 2024]. Moreover, we do not consider purely theoretical papers presenting conceptual frameworks for evaluation [Bengio *et al.*, 2024], methodologies to reduce the computational cost of existing evaluations [Kaplan *et al.*, 2020; Polo *et al.*, 2024] and work studying how humans react or think about AI systems [Steyvers *et al.*, 2025].

1.2 Previous Analyses of AI Evaluation

Several recent surveys have explored AI evaluation from specific perspectives. Many of these are focused on LLMs, overviewing benchmarking [Chang *et al.*, 2024] or redteaming [Lin *et al.*, 2025], studying how ethical aspects are evaluated [Lyu and Du, 2025] or analysing the literature under a verification and validation perspective [Huang *et al.*, 2024]. Other surveys include Ruah *et al.,* [2024], which focuses on the safety of generative AI systems, and [2023], which considers responsible AI principles. In contrast, our work focuses on a wide variety of AI systems and evaluation approaches.

Other work discussed the quality of evaluation practice and instruments [Hernandez-Orallo, 2020; Burden, 2024], for instance touching on reproducibility [Burnell *et al.*, 2023], statistical rigour [Gorman and Bedrick, 2019], validity [Subramonian *et al.*, 2023] and representativeness and fairness [Bergman *et al.*, 2023; Göllner and Tropmann-Frick, 2023]. We do not focus on these issues here, as they apply broadly to all AI evaluation tools.

Bieger *et al.* [2016] proposed a normative framework for evaluating adaptive general-purpose AI, outlining the purposes of evaluation, the properties to be measured, and the challenges involved. In contrast, we take a descriptive approach to AI evaluation, mapping the landscape by identifying the advantages and limitations of different methodologies without prescribing any particular approach. More recently, Cohn *et al.* [2022] defined different “facets” of evaluation instruments. While there is some overlap with our dimensions of analysis, several of their facets refer to validity and consistency, properties that are broadly relevant across all paradigms and are hence excluded from our analysis. While Cohn *et al.* [2022] apply their framework to 23 evaluation works, neither Bieger *et al.* [2016] nor Cohn *et al.* [2022] identify distinct paradigms of evaluation or characterise their defining features. Further, these works do not survey the larger and rapidly evolving AI evaluation landscape shaped by recent advancements in AI.

[1] Such as the Code of Practice for general-purpose AI models, https://digital-strategy.ec.europa.eu/en/policies/ai-code-practice.

[2] We aimed to provide a representative view of AI evaluation rather than an exhaustive catalogue. Hence, we focused on (1) maintaining diversity in the selection of research, (2) emphasising work from the last five years. We have, however, included certain earlier works that have stood the test of time and are still widely used to evaluate modern AI systems. This balance allows us to accurately survey the current landscape of AI evaluation praxis.

[3] We are unable to reference all the surveyed papers within the allocated page limit; the complete annotated list is available here.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

2 Dimensions of Analysis

We explore *goals*, *methodologies* and *cultures* in AI evaluation. These three factors shape the way evaluations are designed, applied, and interpreted, allowing us to highlight the diversity of approaches, clarify underlying assumptions, and identify gaps across different evaluation paradigms.

We break down these three factors into key dimensions (each with a discrete set of possible values) that we use to map the landscape of AI evaluation. In some cases, a single work may be assigned multiple values for the same dimension. For example, an evaluation may assess both performance and safety (two distinct values within the Indicator dimension) using the same tasks. Fig. 1 shows the clustering of surveyed papers based on these dimensions and highlights the relationship between these clusters and assigned paradigms.

2.1 The Goals

Evaluation can be marked by the type of insight sought. We make use of the following dimensions:

Indicator (performance, fairness, safety, robustness and reliability, behavioural features, cost). One possible goal of AI evaluation is to determine the *performance* of an AI system, namely, its ability to successfully complete tasks. Another important consideration is *fairness*: the extent to which demographic groups are treated differently by the AI system. Alternatively, *safety* encompasses concerns such as preventing the generation of manipulative or toxic content, mitigating harmful outcomes whether explicitly prompted by a malicious actor or arising from the system's design and behaviour (alignment). Other evaluation tools focus on the *robustness and reliability* of an AI system, that is, the extent to which its behaviour is affected by factors that are unrelated to the task at hand and how the system fails in presence of anomalies and edge cases. Some evaluations analyse other *behavioural features* (such as preferences, tendencies, reasoning patterns, etc.) of models in response to inputs. Finally, the *cost* of using a system (whether monetary, environmental, or ethical) can also be a primary indicator used to evaluate a system.

Distribution summarisation (aggregate, extreme, functional, manual inspection). Ideally, AI evaluation would fully describe the distribution of subject behaviour conditioned on all possible input values. Yet, in practice, many approaches are limited to reporting summary statistics. Most typical are *aggregate* metrics over a set of inputs (e.g., mean accuracy in benchmarks). Alternatively, the *extremes* of the distribution over a set of inputs are sometimes reported. This includes both worst-case (e.g., the possibility of accidents) and best-case scenarios (e.g., best performance over a set of prompts). More refined is a *functional* description mapping variations in the input (such as task difficulty) to statistics of the conditional distribution of behaviour. This can allow for anticipation and explanation of behaviour (e.g., representing performance as a function of task difficulty). In contrast, a few works instead perform a *manual inspection* of the system's behaviour (e.g., describing performance failures in a qualitative manner) rather than summarising the distribution.

Subject (system, component, algorithm). AI evaluation often focuses on self-contained *systems* that can function without additional components or adaptations (e.g., ChatGPT, a planner or a translator). However, the subject of an evaluation study can also be individual *components* designed to be integrated in other systems, often requiring specialised interfaces (such as specific non-linearity functions, the computation of embeddings, or a SAT solver) or *algorithms* expressed in programming languages or pseudocode (for instance, Stochastic Gradient Descent and Naive Bayes). Both components and algorithms are typically assessed based on their impact across often multiple systems that employ them.

2.2 The Methodologies

A methodology is a collection of practices, principles, and guidelines used to conduct an evaluation. We consider the following dimensions as key factors for categorising methodologies:

Measurement (observations, constructs). Evaluations can report direct *observations* from measurement instruments (e.g., 'System X achieved a score of $x\%$ on a benchmark') or explicitly model latent *constructs*—underlying factors that explain an AI system's observed behaviour (e.g.,'System X demonstrates low arithmetic capability').

Task origin (operation, sample, design). AI systems can be evaluated in real-world *operation* (e.g., testing an autonomous vehicle on public roads or assessing whether an LLM improves productivity in a study involving human workers). Alternatively, evaluation can be conducted using a *sample* of tasks drawn from a distribution representative of real-world usage (e.g., a sample of journeys). In some cases, tasks may be created by *design* (e.g., on a circuit used as a testbed or simulated environment), either because it is difficult, unsafe or unethical to perform evaluation in real-world cases, or because synthetic tasks are thought to enable a more accurate measurement of the desired property.

Protocol (fixed, procedural generation, adaptive, interactive). The evaluation process may unfold in different ways. A common approach involves testing an AI system on a *fixed* set of tasks, defined by the initial input provided to the AI system (even if the interaction between subject and evaluation unfolds differently due to either stochastic evaluation environment or an agent's decisions). Some evaluations use *procedural generation* to create new tasks at test-time for each tested system, following a predefined distribution. In other cases, the generation of tasks or the choice amongst a fixed set of instances is *adaptive*, chosen based on previous behaviour of the system on other tasks (e.g., adjusting the difficulty of questions in response to its answers to previous questions). Finally, evaluations can be *interactive*, where humans guide the process in real-time, probing the system's behaviour through sequential interactions.

Reference (objective, rubric, subjective, no reference). In some cases, the AI system's outputs are compared to an *objective* value, such as a gold standard correct answer. In other cases, a *rubric* is applied, either by a human or an AI scorer, to assess outputs that cannot be objectively compared to a determined value. Alternatively, evaluation may rely on *subjective* feedback or judgement (preferences, moral values, etc.) provided by human users at test time or encoded into an automated system (such as a reward model). In some cases, there is *no reference* answer for comparison.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Task mode (identification, generation). The evaluation may require the subject to *identify* an answer from a predetermined set of options (e.g., selecting a multiple-choice answer or a class label) or to *generate* a novel output (e.g., a numerical value in a continuous range or free-form text).

2.3 The Cultures

By cultures, we refer to the people involved in the evaluation process, their norms, interests and terminology. We categorise evaluation cultures with the following dimensions:

Evaluators (researchers, deployers, regulators). Different people may create and conduct an evaluation. We define *researchers* as those motivated by understanding and improving AI systems from a scientific perspective, focusing on fundamental insights and advancements. In contrast, *deployers* are primarily concerned with commercial viability, customer satisfaction and competitive advantage; they want to determine whether a system is suitable for deployment, assess pricing strategies, ensure safety and enhance brand reputation. Finally, *regulators* seek guarantees and information on safety, ethics, risks to society, and legal compliance. Our definition considers employees of private companies or regulators conducting research for scientific understanding as researchers.

Motivation (comparison, understanding, assurance). Motivation determines focal points for the evaluation, how the results are interpreted, and the suite of techniques it leverages. System *comparison* aims to determine the most suitable subject from a set of candidates for a given scenario, assess whether systems can replace or assist humans, and track the progress of successive generations. Another reason for the development of an evaluation tool is seeking *understanding* of what causes AI systems to behave in the way they do. This could be via learning to recognise the cognitive processes the AI systems exhibits, or identifying what aspects of a specific input lead to certain types of behaviour. Finally, evaluation may aim to provide *assurance*, determining the conditions of reliable operation or a bound on undesirable behaviour, either using empirical methods or formal proofs.

Discipline (AI, Psychology, Security, Economics, etc.). The people developing and conducting an evaluation may belong to different disciplines, such as *AI*, *psychology*, *security*, *economics* or others. Their different cultural perspectives and scholarly norms may reflect differences in the methods employed for the evaluation.

3 Evaluation Paradigms

Based on our survey, we identified six main paradigms of AI evaluation. In the following subsections we describe each paradigm, leveraging insights from our annotation exercise to operationalise the distinctions between them. In Table 1, we present the most common values for each dimension across paradigms. Where possible, we have chosen names for these paradigms in line with community convention.

3.1 Benchmarking Paradigm

Benchmarking can be traced back to the Common Task Framework of the 1980s [Koch and Peterson, 2024]. A key tenet of benchmarking is that by providing constant test conditions, any variation in a system's responses compared to other systems can be attributed to its intrinsic characteristics. Benchmarking, then, involves evaluating AI systems by testing them on standardised sets—or distributions—of instances and summarising and comparing their performance using different *aggregate* measures. The focus is typically on evaluating the *observed performance* of a system or component (although benchmarks also frequently test for fairness, safety and robustness). Benchmarks are often marked against an *objective* reference, with *identification* (e.g., multiple-choice questions) being a common task mode. They are widely applied across a wide range of AI systems, including LLMs, RL agents, image classifiers, and more. Benchmarks are inherently built in order to *compare* systems, providing a tool for selecting the most appropriate AI systems or tracking progress. Archetypal examples in the Benchmarking Paradigm include ImageNet [Deng *et al.*, 2009], the Arcade Learning Environment [Bellemare *et al.*, 2013] and MMLU [Hendrycks *et al.*, 2021a]. Benchmarks also arise in the form of competitions (e.g., the annual RoboCup competition [Kitano *et al.*, 1997]) where entrants compete to demonstrate competence at a particular task.

Benchmarking offers several inherent advantages. First, the use of a standardised test provides a level playing field for all systems being evaluated, while the use of well-defined, objective metrics reduces the ambiguity and allows easy tracking of performance indicators over time. Benchmarks are often publicly available, increasing the transparency of the evaluation process; this can be further improved by reporting the results at the instance level [Burnell *et al.*, 2023]. However, the benchmarking paradigm also has notable limitations. Open sharing of benchmarks can lead to data contamination [Zhou *et al.*, 2023], with test data being used for training. Relatedly, repeated testing on the same benchmark may induce overfitting, optimising for specific tests rather than generalised to broader tasks [Fang *et al.*, 2023]. Further, benchmark results are inherently tied to a particular distribution of test items, limiting the generalisability of the measurement [Raji *et al.*, 2021]. Further discussions on the limitations of benchmarking can be found in Liao *et al.* [2021].

3.2 Evals Paradigm

The Evals Paradigm focuses on system *safety*, often operationalised as a system's failure to comply with safety specifications or its tendency to exhibit harmful behaviour during tasks that carry risks (e.g., the extent to which it demonstrates so-called "dangerous capabilities" [Shevlane *et al.*, 2023; Phuong *et al.*, 2024]), although *fairness* is sometimes considered too. This Paradigm often employs *extremes* analysis, identifying specific (worst) cases. The aim is providing *assurance* on a system's safety or gaining *understanding* on what causes the unsafe behaviour. A common methodology within the Evals Paradigm is "red-teaming", an *interactive* and adversarial process where humans or automated agents attempt to "break" or provoke the system to *generate* undesirable responses. Red-teaming is commonly employed by model *deployers* (e.g., OpenAI [OpenAI, 2023], Anthropic [Anthropic, 2023]). Generally, tasks are *designed* in an adversarial way. Evals are predominantly applied to general-purpose AI systems such as LLMs, for which assessing potential risks

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Dimension	**Benchmarking**	**Evals**	**Construct-Oriented**	**Exploratory**	**Real-World Impact**	**TEVV**
Archetype	[Deng *et al.*, 2009]	[Ganguli *et al.*, 2022]	[Guinet *et al.*, 2024]	[Berglund *et al.*, 2024]	[Collins *et al.*, 2023]	[Yang *et al.*, 2023a]
Indicators	Performance	**Safety**/Fairness	-	Performance/Behaviour	Cost/Fairness/**Performance**	Safety/robustness and reliability
Dist. summ.	**Aggregate**	Extreme/Aggregate	**Functional**	Aggregate/Manual Inspection	**Aggregate**	Extreme
Subject	-	**System**	**System**	System	**System**	**System**
Measurement	**Observed**	**Observed**	**Latent**	**Observed**	Observed	Observed
Task origin	-	**Design**	Design	Design	-	Design/Operation
Protocol	Fixed	Interactive	-	-	-	-
Reference	Objective	Subjective	-	-	Subjective/Rubric	Objective
Task Mode	-	**Generation**	-	-	**Generation**	-
Evaluators	Researchers	Researchers, Deployers	**Researchers**	Researchers	-	-
Motivation	Comparison	Assurance/Understanding	Understanding	**Understanding**	**Comparison**	**Assurance**
Disciplines	-	Security, Bio	Cognitive Science	-	Social Sciences	Control Theory
Raw Number	72	13	15	18	4	10
Percentage	57%	10%	12%	14%	3.2%	7.9%

Table 1: For each paradigm we identify, the table shows an archetypal paper and the values of different dimensions that mostly characterise that paradigm. Bolded entries indicate values that were present across the vast majority of the considered papers within that paradigm. Empty entries indicate that a dimension was not informative for that paradigm. Also included are the number and percentage of papers identified as belonging to each paradigm (note that a few papers bridge multiple paradigms and are therefore double-counted above).

is particularly relevant due to their broad (and unpredictable) range of applications. Since red-teaming is often conducted by human evaluators, there is typically no objective reference against which the system's responses are compared. Instead, evaluation relies on *subjective* human judgment. Archetypal papers from the Evals paradigm are Ganguli *et al.* [2022]'s work on red-teaming or Kinniment *et al.* [2024]'s examination of LLM's ability to self-replicate.

Evals offer a systematic way to identify flaws in AI systems. The failures uncovered by Evals are concrete and actionable, making them amenable to mitigation through additional training, fine-tuning, or reinforcement learning from human feedback (RLHF) [Christiano *et al.*, 2017]. However, a major limitation of Evals is that failing to identify flaws does not indicate that a system is safe for deployment (absence of evidence is not evidence of absence). Many evaluated systems (often LLMs) are poorly understood black boxes and can be highly sensitive to small input variations. The ad-hoc and subjective nature of Evals evaluations risks not rigorously accounting for these subtle changes, limiting their reliability as comprehensive safety assessments.

3.3 Construct-Oriented Paradigm

The Construct-Oriented Paradigm leverages system responses to quantitatively measure underlying "constructs" that describe the system's behaviour at an abstract level. These constructs are typically based on existing theories of cognitive traits or capabilities, but they can also be solely inferred from system responses (e.g., using factor analysis). To measure these *latent* constructs, this paradigm often employs a *functional* link with the observed system behaviour and *designed* tasks carefully controlling for confounding factors, possibly adapting tasks from the cognitive sciences literature. This paradigm primarily aims to gain a better *understanding* of self-contained *systems*, particularly their *performance*. The works in this paradigm are mostly authored by *researchers* with a background in *psychology/cognitive science*. Archetypal examples are Guinet *et al.* [2024], which applies psychometric methods to infer the ability of LLMs to solve 8th-grade mathematics tests, and Momennejad *et al.* [2024], which proposes a protocol for evaluating cognitive capabilities in LLMs, operationalising these capabilities through variations in specific tasks to systematically investigate how these variations influence LLM responses.

The Construct-Oriented paradigm has the advantage of providing measurements that can be robust to variations in the test set and that can be more readily transferred from evaluation settings to real-world deployment scenarios. However, developing evaluation instruments within this paradigm is considerably challenging: it requires strong domain expertise and detailed mathematical modelling of cognitive phenomena. A limitation of this approach is its reliance on existing human psychology theories which may not always provide suitable accounts for complex or capabilities and traits of general-purpose systems.

3.4 Exploratory Paradigm

The Exploratory Paradigm begins by forming a hypothesis—often inspired by anecdotal observations—about a system's behaviour. Similarly to how psychologists test hypothesis for human and animal cognition, a set of tasks capturing key features—such as reasoning steps, memory requirements or generalisation patterns—is *designed* to systematically isolate the phenomenon in different scenarios and exclude alternative explanations. The findings arising from testing the *system* on those tasks, mostly in terms of *observed performance* and occasionally *manual inspection* of *behavioural features*, are combined (and, in some cases, compared to humans) to provide evidence supporting the considered hypothesis qualitatively describing a system's "cognitive processes". The exploratory approach contrasts with the Construct-Oriented paradigm, which usually relies on pre-established theoretical frameworks that are complemented by the quantitative measurements of constructs. The aim to *understand* the AI system's cognitive processes is what chiefly distinguishes this paradigm from Benchmarking (which also focuses on observed behaviour, but for the purpose of comparing systems). Here, the AI system's results on tests are only important insofar as they support or refute the hypothesis of interest. Work in this paradigm is chiefly conducted by *researchers* in *AI* or *psychology/cognitive science* and they have been mostly applied to LLMs; archetypal

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

works include The Reversal Curse [Berglund *et al.*, 2024] and MeltingPot [Agapiou *et al.*, 2022].

The Exploratory paradigm offers the ability to propose and test hypotheses and provide glimpses into AI systems' inner mechanisms at the representational level of analysis [Marr, 1982]. By examining behavioural traces, this paradigm can provide novel insights into cognitive processes that may not be captured by aggregate metrics alone. This paradigm also has limitations: first, making well-designed tests requires substantial expertise and effort; moreover, each work in this paradigm typically relies on a few bespoke tests to explore narrow hypotheses, making it challenging to synthesise the findings into a comprehensive account of an AI system's overall behaviour and tendencies. This, together with the focus on observed indicators—rather than developing models of latent constructs—can limit generalisability, making it difficult to draw broader conclusions about a system's properties.

3.5 Real-World Impact Paradigm

While most evaluation paradigms seek to assess specific properties of AI systems, the Real-World Impact (RWI) Paradigm measures the impact of AI systems when deployed in the real world. This paradigm leverages techniques from the social and clinical sciences by running (randomised controlled) trials where, most often, assistance by an AI *system* to humans is considered as an "intervention" whose effect must be quantified, in terms of a change in *aggregate performance* on the considered task relative to the system's *cost*. The goal is therefore *comparing* different AI systems, or AI-assisted humans to humans alone. Evaluations are generally carried out *interactively* or on a *fixed* dataset and, due to the complexity of real-world tasks, human *subjective* ratings or *rubrics* may be employed. A common motivation for these evaluations is the *comparison* of systems situated in the context of real-world applications requiring *generation* of novel outputs. An archetypal work from this paradigm is `Math Converse` [Collins *et al.*, 2023], which investigates the perceived helpfulness of LLMs for mathematics. A second representative work is Si *et al.* [2024]'s study of LLM's ability to generate novel research ideas; which observes that human reviewers find LLM-generated ideas to be more novel but less feasible than those crafted by humans.

The RWI paradigm has a number of advantages: as AI systems become better at complex tasks, the social sciences provide many established methodologies to evaluate their societal impact, which may not be well estimated in artificial scenarios not directly considering user experience. This paradigm has practical challenges as well: in contrast to other paradigms, conducting experiments with human participants in realistic scenarios adds ethical constraints and logistical complexities. Moreover, this type of research mostly operates on a slower timescale compared to AI, making it hard for the RWI paradigm to timely provide information on new systems. These challenges have likely contributed to RWI being a small paradigm so far (it is the least represented in our sample of papers). Nevertheless, we expect this paradigm to grow significantly in the coming years as AI systems become more capable of performing economically valuable tasks.

3.6 TEVV Paradigm

The Test, Evaluation, Verification, and Validation (TEVV) Paradigm draws on methodologies from formal software verification, with its primary focus on ensuring that AI *systems* behave in a well-defined and predictable way. TEVV is characterised by focusing on *observed extreme* values of *safety* or *robustness and reliability* measures, with the central goal of *assurance*—providing bounds or guarantees for a minimum or average level of performance under various conditions. To achieve this, TEVV often explicitly operationalises the constructs it aims to measure, formally defining them to reduce uncertainty of the measurement process. It then employs either *designed* tasks or *operational* studies, with a variety of protocols. Works in this paradigm commonly deal with Reinforcement Learning and applied fields such as autonomous driving; Yang *et al.* [2023a] and Mussot *et al.* [2024] are archetypal examples of TEVV works in these two fields.

TEVV offers several advantages, the most notable being its ability to provide formal guarantees and robust safety assurances. However, this approach requires a deep understanding of the system and its operational mechanisms. For many state-of-the-art or general-purpose AI systems, such an understanding is often lacking, making TEVV challenging to apply effectively. Indeed, we found that TEVV was one of the least represented paradigms in recent AI venues.

3.7 Evaluations crossing Paradigm Boundaries

Unsurprisingly, we found several evaluation papers do not fit neatly into a single paradigm, but instead bridge multiple paradigms and combine their methodologies (see also Fig. 1). For example, Perez *et al.* [2022]'s model-written evaluations straddle the Benchmarking and Evals paradigms, using LLMs to generate a large and varied set of questions, thus creating a standardised benchmark, that aim to provide safety assurances in the style typical of Evals. Similarly, CogBench [Coda-Forno *et al.*, 2024] bridges the Capability-oriented and Exploratory paradigms. Here, the authors build a cognitive phenotype of LLMs using psychological experiments designed to assess different *constructs*. Simultaneously, they *explore* a number of hypotheses based on anecdotal evidence, including whether RLHF makes LLMs more human-like and the relationship between model size and tendency to exhibit human-like behaviour. These hybrid approaches demonstrate the flexibility of AI evaluation methodologies.

4 The Role of the Different Paradigms

Each of the paradigms we described plays a distinct role within the AI evaluation ecosystem. Each targets a particular type of measurement and fulfils the unique needs of different evaluators. For example, a company developing a safety-critical AI system, such as an autonomous vehicle, would rely on evaluation methodologies that closely follow paradigm to obtain robust guarantees. On the other hand, a deployer of a system in a lower-stakes environment, such as an image classifier for pets, would likely rely on the performance of the system on a benchmark of representative images to determine when the system is ready to deploy. Techniques from multiple paradigms can (and should) be combined when appropriate.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

This is already common among developers of state-of-the-art general-purpose AI systems, where a mix of Benchmarking (to assess capabilities such as reasoning or coding skills) and Evals (to assess safety concerns) is used. Beyond this specific combination, integrating methodologies across different paradigms remains relatively uncommon. While this may be considered as a limitation of the current AI evaluation ecosystem, this also presents an opportunity for improvement.

5 Challenges and Opportunities

Some paradigms, such as Benchmarking, are applied across a wide range of AI systems. However, others tend to be domain-specific: TEVV is mostly applied to embodied, agentic forms of AI, such as RL systems or self-driving cars; similarly, in the papers we surveyed, we found the Evals and Exploratory paradigm were almost exclusively applied to LLMs, although this could be partly due to our focus on recent works. We believe this has occurred due to the way these paradigms emerged in response to specific developments in AI. For example, Evals largely developed as a response to risks from LLMs [Ganguli *et al.*, 2022]. This means that a vast range of existing evaluation approaches remain underutilised in various domains and AI system types. While technically challenging, expanding the applications of different paradigms beyond their typical uses would lead to a more comprehensive understanding of AI systems, their strengths, weaknesses, and broader impacts. While there appears to be growing interest in expanding the range of techniques applied to LLMs, often drawing on methods from TEVV [Huang *et al.*, 2024; OpenAI, 2024], we hope to see this cross-pollination across all domains where AI is evaluated.

Besides expanding the domain of application of each paradigm, great opportunities lie in developing new evaluations bridging different paradigms, as the examples mentioned in Sec. 3.7. By highlighting the possibilities afforded by the paradigms we identified, we hope to inspire researchers to develop new evaluations leveraging the strengths of multiple paradigms for specific questions, to achieve more comprehensive and insightful assessments. At the same time, as discussed in Sec. 4, using multiple paradigms to tackle an individual question from different perspectives is also a powerful but underexploited strategy.

We can integrate our insights with existing discussions on gaps in AI evaluation. For instance, Hutchinson *et al.* [2022] point out the lack of *moral* evaluations in AI development raising important questions about data consent, the dignity of data workers, and the social responsibilities of developers. These factors were not included in our dimensional analysis due to the widespread lack of reporting on such topics. Similarly, Rauh *et al.* [2024] identify a "risk coverage gap" with many ethical and social risks currently insufficiently addressed. In our framework, evaluations addressing these risks would likely fall under the Real-World Impact Paradigm, which we found to be the least developed. Therefore, Rauh *et al.* [2024] and our work both highlight how this niche is unaddressed. Finally, Huang *et al.* [2024] points out the lack of verifications with provable guarantees for LLMs, which also surfaced from our analysis of the TEVV paradigm. In general, by drawing attention to gaps in the space, we hope to encourage researchers to develop new evaluation methodologies that better address issues.

6 Limitations

We aimed to capture the breadth of the AI evaluation landscape by surveying a highly diverse set of works. Given the extent of the field, our broad scope inevitably limited the depth with which we survey each paradigm. A focused investigation of fewer paradigms might reveal additional patterns or relationships that we could not fully explore. Additionally, our selection of papers was based on our analysis of the recent literature, which may introduce bias. Certain evaluation paradigms, such as Benchmarking, may be overrepresented due to trends in the field, particularly the focus on LLMs. To mitigate this, we ensured that the authors of this paper have extensive AI evaluation expertise and deep familiarity with different areas of the ecosystem. This, combined with seeking out evaluation works intentionally different from one another, mitigated our selection bias as much as reasonably possible.

Another constraint lies in how our dimensions capture the nuances of different evaluation techniques. For example, we found the distinction between aggregate and functional distribution summary to be sometimes blurry—there is a fine line between stratifying aggregate performance based on predefined categories (e.g., required capabilities for tasks) and an imprecise functional model. This and other ambiguities can lead to disagreements among raters when assigning dimensions to papers and classifying them into particular paradigms. At the same time, incorporating additional dimensions could offer deeper insights into how different paradigms are characterised and help identify sub-paradigms. Our approach balances granularity of annotation with practical usability. We further found that some dimensions were less informative than anticipated, for example Task Mode (Identification and Generation) was not useful for distinguishing between paradigms where perhaps a more nuanced breakdown of Task Mode would have been. Finally, the AI evaluation landscape is rapidly evolving and new paradigms may emerge or existing ones become more refined. Despite these limitations, we believe the dimensions introduced here provide a valuable foundation for guiding further research in AI evaluation and characterising future works.

7 Conclusion

This survey presents a snapshot of the current AI evaluation landscape, offering insights into prevailing approaches. We categorised over 125 recent or highly influential AI evaluation papers based on our multi-dimensional framework examining goals, methodologies, and research cultures. Through this analysis, we identified six distinct paradigms offering individual perspectives to AI evaluation that contributes to the wider AI evaluation ecosystem, despite the lack of standardisation and occasional inconsistencies in terminology across these paradigms. Our aim with this paper was to bring attention to these different approaches and foster greater cross-pollination between paradigms, ultimately promoting a more integrated and holistic assessment of AI.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Acknowledgements

JB received funding from the Effective Ventures Foundation Long Term Future Fund through Grant ID: a3rAJ000000017iYA. LP received funding from US DARPA (grant HR00112120007, RECoG-AI) and Open Philanthropy. JHO acknowledges the grants SFERA (PID2021-122830OB-C42) and Cátedra ENIA-UPV, TSI-100930-2023-9 funded by Spain's MCIN.

Contribution Statement

JB, MT and LP contributed equally to this work.

References

[Agapiou *et al.*, 2022] John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, et al. Melting pot 2.0. *arXiv preprint arXiv:2211.13746*, 2022.

[Anthropic, 2023] Anthropic. Frontier Threats Red Teaming for AI Safety, 2023.

[Asgari *et al.*, 2024] Ali Asgari, Antonio Guerriero, Roberto Pietrantuono, and Stefano Russo. From testing to evaluation of NLP and LLM systems: An analysis of researchers and practitioners perspectives through systematic literature review and developers' community platforms mining. 2024.

[Bellemare *et al.*, 2013] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[Bengio *et al.*, 2024] Yoshua Bengio, Michael K Cohen, Nikolay Malkin, Matt MacDermott, Damiano Fornasiere, Pietro Greiner, and Younesse Kaddar. Can a Bayesian oracle prevent harm from an agent? *arXiv preprint arXiv:2408.05284*, 2024.

[Bereska and Gavves, 2024] Leonard Bereska and Efstratios Gavves. Mechanistic Interpretability for AI Safety – A Review, April 2024. arXiv:2404.14082 [cs].

[Berglund *et al.*, 2024] Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*, 2024.

[Bergman *et al.*, 2023] A Stevie Bergman, Lisa Anne Hendricks, Maribeth Rauh, Boxi Wu, William Agnew, Markus Kunesch, Isabella Duan, Iason Gabriel, and William Isaac. Representation in AI evaluations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 519–533, 2023.

[Bieger *et al.*, 2016] Jordi Bieger, Kristinn R Thórisson, Bas R Steunebrink, Thröstur Thorarensen, and Jona S Sigurdardottir. Evaluation of general-purpose artificial intelligence: why, what & how. *Evaluating General-Purpose AI*, 2016.

[Burden, 2024] John Burden. Evaluating AI evaluation: Perils and prospects. *arXiv preprint arXiv:2407.09221*, 2024.

[Burnell *et al.*, 2023] Ryan Burnell, Wout Schellaert, John Burden, Tomer D Ullman, Fernando Martinez-Plumed, Joshua B Tenenbaum, Danaja Rutar, Lucy G Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, 2023.

[Casper *et al.*, 2024] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272, Rio de Janeiro Brazil, June 2024. ACM.

[Chang *et al.*, 2024] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[Christiano *et al.*, 2017] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[Coda-Forno *et al.*, 2024] Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. CogBench: a large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9076–9108. PMLR, 21–27 Jul 2024.

[Cohn *et al.*, 2022] A Cohn, José Hernández-Orallo, Julius Sechang Mboli, Yael Moros-Daval, Zhiliang Xiang, and Lexin Zhou. A framework for categorising AI evaluation instruments. In *Proceedings of the Workshop on AI Evaluation Beyond Metrics co-located with the 31st International Joint Conference on Artificial Intelligence (IJCAI-ECAI 2022)*, volume 3169. CEUR Workshop Proceedings, 2022.

[Collins *et al.*, 2023] Katherine M. Collins, Albert Q. Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B. Tenenbaum, William Hart, Timothy Gowers, Wenda Li, Adrian Weller, and Mateja Jamnik. Evaluating language models for mathematics through interactions, 2023.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

[European Union, 2024] European Union. EU AI Act. *https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689*, 2024.

[Fang *et al.*, 2023] Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on imagenet transfer to real-world datasets? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 25050–25080. Curran Associates, Inc., 2023.

[Ganguli *et al.*, 2022] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[Göllner and Tropmann-Frick, 2023] Sabrina Göllner and Marina Tropmann-Frick. Bridging the gap between theory and practice: Towards responsible AI evaluation. In *CHAI@ KI*, pages 68–76, 2023.

[Gorman and Bedrick, 2019] Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy, July 2019. Association for Computational Linguistics.

[Guinet *et al.*, 2024] Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. Automated evaluation of retrieval-augmented language models with task-specific exam generation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[Hendrycks *et al.*, 2021a] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[Hendrycks *et al.*, 2021b] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*, 2021.

[Hernández-Orallo, 2017] José Hernández-Orallo. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press, 2017.

[Hernandez-Orallo, 2020] Jose Hernandez-Orallo. AI evaluation: On broken yardsticks and measurement scales. In *Workshop on evaluating evaluation of AI systems at AAAI*, 2020.

[Huang *et al.*, 2024] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):175, 2024.

[Hutchinson *et al.*, 2022] Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1859–1876, 2022.

[Kaplan *et al.*, 2020] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[Kinniment *et al.*, 2024] Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. Evaluating language-model agents on realistic autonomous tasks, 2024.

[Kitano *et al.*, 1997] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, and Eiichi Osawa. Robocup: The robot world cup initiative. In *International Conference on Autonomous Agents*, 1997.

[Koch and Peterson, 2024] Bernard J. Koch and David Peterson. From protoscience to epistemic monoculture: How benchmarking set the stage for the deep learning revolution, 2024.

[Liao *et al.*, 2021] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *NeurIPS Datasets and Benchmarks Track*, 2021.

[Lin *et al.*, 2025] Lizhi Lin, Honglin Mu, Zenan Zhai, Minghan Wang, Yuxia Wang, Renxi Wang, Junjie Gao, Yixuan Zhang, Wanxiang Che, Timothy Baldwin, et al. Against the achilles' heel: A survey on red teaming for generative models. *Journal of Artificial Intelligence Research*, 82:687–775, 2025.

[Lyu and Du, 2025] Yujing Lyu and Yanyong Du. The ethical evaluation of large language models and its optimization. *AI and Ethics*, pages 1–14, 2025.

[Maliugina, 2024] Dasha Maliugina. 45 real-world LLM applications and use cases from top companies, 2024.

[Marr, 1982] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982.

[McInnes *et al.*, 2018] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[Momennejad *et al.*, 2024] Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Nebojsa Jojic, Hamid Palangi, Robert Ness, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with cogeval. *Advances in Neural Information Processing Systems*, 36, 2024.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

[Mussot *et al.*, 2024] Vincent Mussot, Eric Jenn, Florent Chenevier, Ramon Conejo Laguna, Yassir Id Messaoud, Jean-Loup Farges, Anthony Fernandes Pires, Florent Latombe, and Stephen Creff. Assurance cases to face the complexity of ML-based systems verification. In *Embedded Real Time System Congress, ERTS'24*, 2024.

[OpenAI, 2023] OpenAI. OpenAI Red Teaming Network, 2023.

[OpenAI, 2024] OpenAI. Introducing the model spec, 2024. Accessed: 2025-01-28.

[Perez *et al.*, 2022] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022.

[Phuong *et al.*, 2024] Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Gregoire Deletang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe, and Toby Shevlane. Evaluating frontier models for dangerous capabilities, 2024.

[Polo *et al.*, 2024] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

[Raji *et al.*, 2021] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the everything in the whole wide world benchmark. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[Rauh *et al.*, 2024] Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, et al. Gaps in the safety evaluation of generative AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1200–1217, 2024.

[Shevlane *et al.*, 2023] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

[Si *et al.*, 2024] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ nlp researchers. *ArXiv*, abs/2409.04109, 2024.

[Steyvers *et al.*, 2025] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, pages 1–11, 2025.

[Subramonian *et al.*, 2023] Arjun Subramonian, Xingdi Yuan, Hal Daum'e, and Su Lin Blodgett. It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. In *Annual Meeting of the Association for Computational Linguistics*, 2023.

[Yang *et al.*, 2023a] Wen-Chi Yang, Giuseppe Marra, Gavin Rens, and Luc De Raedt. Safe reinforcement learning via probabilistic logic shields. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

[Yang *et al.*, 2023b] Wenli Yang, Yuchen Wei, Hanyu Wei, Yanyu Chen, Guan Huang, Xiang Li, Renjie Li, Naimeng Yao, Xinyi Wang, Xiaotong Gu, Muhammad Bilal Amin, and Byeong Kang. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Human-Centric Intelligent Systems*, 3(3):161–188, August 2023.

[Zhou *et al.*, 2023] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your LLM an evaluation benchmark cheater, 2023.