PREP-Eval : Pre-registration and REporting Protocol for AI Evaluations

MARÍA VICTORIA CARRO, University of Genoa, IT and FAIR, IALAB, University of Buenos Aires, ARG
RYAN BURNELL, The Alan Turing Institute, UK
CARLOS MOUGAN, AI Office - European Commission, EU
ANKA REUEL, Stanford University, US
WOUT SCHELLAERT, AI Office - European Commission, EU
OLAWALE ELIJAH SALAUDEEN, Massachusetts Institute of Technology, US
LEXIN ZHOU, Princeton University, US
PATRICIA PASKOV, Oxford Martin AI Governance Initiative, UK
ANTHONY G. COHN, University of Leeds, UK
JOSÉ HERNÁNDEZ-ORALLO, Cambridge University, UK and Universitat Politècnica de València, ES

Evaluation is an integral part of the development cycle of any AI system. As AI grows more sophisticated and general, evaluations are becoming increasingly complex and broad in scope, requiring larger multidisciplinary teams, longer timescales, and more elaborate evaluation tools and techniques. Moreover, the opportunity to deploy these general-purpose AI systems across various commercial and public-sector contexts and the potential risks associated with these deployments have created a burgeoning need to evaluate the suitability and safety of these models for use in different contexts. Despite the growing focus on evaluations, there is no generally established protocol or methodology for conducting AI evaluations, beyond the specific practices of frontier AI developers. Here we aim to address this gap by presenting the "Pre-registration and REporting Protocol for AI Evaluations" (`PREP-Eval`), a step-by-step guide for planning and conducting AI evaluations that complements existing transparency tools such as model cards and evaluation factsheets. We draw on insights from analogous practices in fields such as software testing, data mining, and psychology, and incorporate a pre-registration requirement that facilitates the documentation and justification of deviations from the original plan, helping to identify and avoid selective reporting. Our protocol is designed to support a wide range of stakeholders, including frontier AI developers, third-party evaluators, oversight bodies, and newcomers to the field of AI evaluation, but it is particularly valuable for small or medium-sized research or industry teams that are developing new AI tools or integrating existing models into novel applications and may lack established evaluation pipelines. We demonstrate the application of `PREP-Eval` across six diverse use cases and anticipate that `PREP-Eval` will be further consolidated and improved through iterative feedback and collaboration with the broad AI community.

CCS Concepts: • **General and reference → Evaluation**; Reliability; *Measurement*; Computing standards, RFCs and guidelines; • **Computing methodologies → Artificial intelligence**.

Authors' Contact Information: María Victoria Carro, 6381013@studenti.unige.it, University of Genoa, Genova, IT and FAIR, IALAB, University of Buenos Aires, Buenos Aires, ARG; Ryan Burnell, rburnell@turing.ac.uk, The Alan Turing Institute, UK; Carlos Mougan, Carlos.MOUGAN@ec.europa.eu, AI Office - European Commission, Brussels, EU; Anka Reuel, anka.reuel@stanford.edu, Stanford University, US; Wout Schellaert, Wout.SCHELLAERT@ec.europa.eu, AI Office - European Commission, Brussel, EU; Olawale Elijah Salaudeen, olawale@mit.edu, Massachusetts Institute of Technology, US; Lexin Zhou, lz5066@princeton.edu, Princeton University, US; Patricia Paskov, ppaskov@rand.org, Oxford Martin AI Governance Initiative, UK; Anthony G. Cohn, A.G.Cohn@leeds.ac.uk, University of Leeds, UK; José Hernández-Orallo, jorallo@upv.es, Cambridge University, UK and Universitat Politècnica de València, València, València, ES.

Manuscript submitted to ACM

Additional Key Words and Phrases: AI Evaluations, Pre-registration, Safety Evaluation, AI Regulation

1 Introduction

Recent progress in AI has led to a marked broadening of the scale and scope of system evaluations. Over the past few years, a key driver of this shift has been the rise of large-scale, general-purpose systems, such as OpenAI's GPT or Google's Gemini, which are designed without a specific use case or task in mind but can be adapted to a wide variety of applications. As developers, researchers, and the broader community seek to understand the capabilities and limitations of these systems, evaluations constitute a meaningful tool for measurement and estimation across a wide array of domains and contexts.

In addition, as AI systems evolve toward architectures with greater autonomy, multimodal perception, and increasingly complex ways of interacting with environments, they give rise to a multitude of questions for evaluations to address. For example, how reliably can a code-writing agent adapt to changing specifications or user requirements? Is a delivery robot safe to deploy in an area of the city? Given a new route, what is the probability that a self-driving car will reach its destination safely? Can a new language model specialised in chemistry be used maliciously? Is a customer-service bot treating all (protected) groups in the same way? Answering these diverse questions requires similarly diverse tools and methods, inciting an evolving landscape of evaluations.

However, despite the growing complexity and importance of these evaluations, there is little guidance on how AI evaluations should be conducted [150], especially by small or medium-sized teams with modest budgets adapting or deploying AI systems in sociotechnical environments. A great deal of recent work has focused on what should be evaluated and the benchmarks (e.g., [14, 27, 54]), but relatively little has focused on how to plan and conduct the evaluation. While there is broad agreement on the general desiderata that evaluations should satisfy (i.e., external validity, internal validity and reproducibility), and the limitations of benchmarks [39, 58, 89], there remains an urgent need for a concrete step-by-step protocol that organisations can follow. *This methodological gap leads to evaluations that are poorly thought-out, inefficient, or potentially ineffective.* The increasing use of evaluation for decision-making and policymaking creates a window of opportunity to strengthen evaluation infrastructure, frameworks, and processes to promote more robust methods and more reliable evidence.

To feel this gap, we present `PREP-Eval`, a Pre-registration and REporting Protocol for AI Evaluations, consisting of well-delineated project-methodology stages with a pre-registration outcome after the 'project plan' (Figure 1). While backwards arrows are not shown, it is always possible to go back one or more stages if the process reveals that some earlier choices need to be revisited. However, it is important that each evaluation project starts with clear goals (stage 1) and is followed by well-thought design and plan (stages 2 and 3), not carved in stone, but pre-registered. Following modern, agile methodologies, the protocol is meant to be iterative, as a first cycle of evaluation may motivate a refinement or extension of goals and objectives, triggering a new cycle.

The protocol aims to provide a broad tool for planning, pre-registration and documentation for any evaluation project involving any kind of AI, from specialised tools to general-purpose AI (GPAI), from well-known models in new scenarios to frontier models in old scenarios. While humans interacting of AI are usually subject of the evaluations as well (impact, uplifting, etc.), the emphasis is put on the evaluation of AI. In some situations, it could serve as requirements for an evaluation contract. It is designed to allow evaluators to document their evaluation protocol ahead of time and later report this plan when they present their evaluation results. Doing so would allow the community to better understand why certain evaluation decisions were made and verify that the reported results align with the original plan. The approach is consistent with and complementary to other efforts to increase transparency in AI, such as

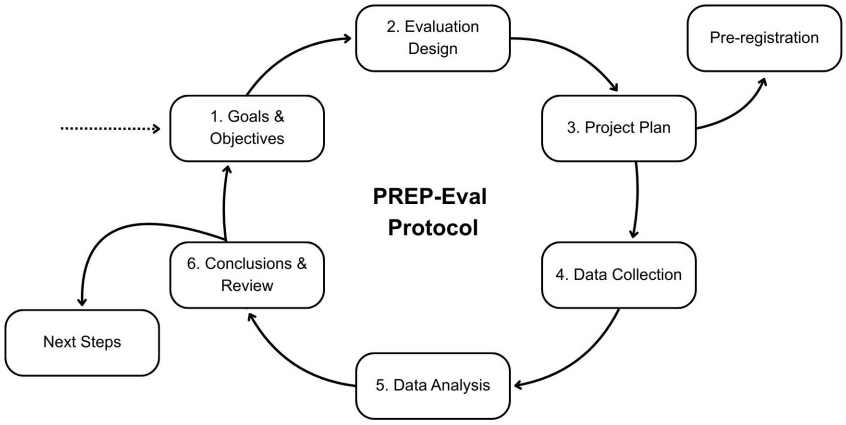

Fig. 1. PREP-Eval lifecycle. Each stage comes with its own substages and documentation (Appendix A). The protocol emphasises the relevance of the first stages (1 to 3), and pre-registration serves several purposes, such as taking these three first stages seriously, before rushing into the other stages and ensuring that the influence of the analysis on an eventual revision of goals is transparent.

Model (and System) Cards (which aim to standardise reporting about the architecture and training of a model; [94]), the Machine Learning Reproducibility checklist (which aims to improve reporting of experimental findings; [107]), the Report Cards, Eval Factsheets or Evaluation Cards (which aim to document or audit AI evaluations, [15, 41, 146, 151]), and the Voluntary Reporting Templates and Safety Cases (which aim to report the risks and mitigations of an AI system, or provide a structured, evidence-based justification that a system is sufficiently safe [2, 19, 31]).

Importantly, even when evaluations are not conducted as part of a formal scientific endeavour, and adherence to traditional scientific standards is not the primary goal, applying this protocol can still be valuable and beneficial. First, it helps the actor conducting the evaluation to plan it carefully. Evaluations can be resource-intensive, and even expert teams can encounter unexpected challenges. A structured protocol allows these challenges to be anticipated, ensuring that the evaluation remains on track and that any potential obstacles are addressed in a timely manner, alleviating potential time pressures. Second, being transparent and explicitly justifying decisions increases trust among both users and stakeholders. Finally, documenting deviations from the original plan and lessons learned in a standardized way facilitates knowledge transfer and collective expertise building, particularly for actors at the forefront of innovation in evaluation methodologies.

In the following sections, we first identify the reasons why an AI evaluation protocol is needed, then outline existing best practices and recommendations for AI evaluations, explaining how our proposal complements these approaches, integrates ideas from protocols in other fields, while providing a novel and necessary contribution towards more robust and reliable AI evaluations. Finally, we present the main phases and subphases of PREP-Eval, and provide various examples that are extended in the appendix. We close the paper with a discussion about the current limitations of the present version and the route map for the protocol in the years to come.

2 The Need for an AI Evaluation Protocol

AI evaluations are gradually moving away from the classic static benchmark in favour of experiments that incorporate techniques from software engineering (e.g. CheckList; [118]), psychology and cognitive science (e.g. [32, 59, 64, 70, 77,

88, 96, 133]); psychometrics (e.g. [76, 86, 138, 154]), cybersecurity (e.g. [43, 47, 106]), and other paradigms [20]. These techniques allow for much more robust inferences about model capabilities and propensities; however, as they draw on cross-disciplinary methods that increase methodological complexity, they may also necessitate larger multidisciplinary teams, longer timescales, and more elaborate evaluation tools. This more involved approach raises a growing need for clear guidance on how to plan, conduct, and document evaluations [21, 115].

This need is further exacerbated by growing number of the actors needing to conduct evaluations. The impressive capabilities of large language models (LLMs) have prompted many organisations with different levels of experience in working with AI systems to try to harness these models for many use cases. However, determining the suitability of a given model for a particular use case is far from trivial. Often, it requires organisations to perform their own system evaluations, or at the very least to be able to access and interpret evaluations conducted by other groups.

Evaluation is also becoming a key focus of government oversight bodies and groups interested in AI safety. Concerns about the potential harms of AI systems have led to calls for independent evaluations assessing the generation of misinformation, biased outputs, harmful behaviours and misuse [7, 82, 129]. This particular view of safety "evals" employs extreme-case analysis, identifying specific worst-case scenarios with the aim of providing assurance about a system's safety or gaining understanding on what causes the unsafe behaviour [20]. Moving forward, these evaluations are likely to form a crucial element of model audits, regulatory oversight, and certifications [95, 111, 113, 115, 124].

Notably, regulations such as the EU AI Act incorporate evaluations in several key provisions [39, 103]. More recently, the Code of Practice [18, 40, 60] requires signatories to conduct state-of-the-art model evaluations relevant to the systemic risk in question, with their design being informed by independent sources of information [108]. In parallel, one of the objectives of the U.S. National AI Action Plan [132] is to build a robust evaluation ecosystem, and the National Institute of Standards and Technology (NIST) has recently released a series of evaluation programs and reports [2, 98, 122]. In addition, the National Technical Committee 260 on Cybersecurity of Standardisation Administration of China has published safety assessment requirements for companies developing generative AI systems [99].

Even (or especially) if each stakeholder has their own evaluation methodology, we need interoperability and mutual understanding, especially for auditing and regulation. This requirement is at odds with most evaluations lacking process-oriented documentation [114, 141]. Without clear, step-by-step documentation, it is difficult for policymakers, researchers, and members of the public to understand what an evaluation shows, how robust and valid the findings are, and how much the results should be trusted. There are many choices evaluators must make regarding the methods, analyses, and metrics they use that can affect the results [55]. Yet these decisions are rarely explained fully, making it difficult to know what the results can tell us about a system's capabilities and propensities, and ultimately performance and safety. This trend is especially problematic for evaluations conducted by commercial system developers, who have a strong incentive to present their AI systems in a positive light. This could be done through selectively reporting only benchmarks for which the system performs well, by conducting evaluations on training data, by drawing comparisons with other systems only when the comparisons are favourable, or by selecting the analysis methods or reporting metrics that are most flattering to the system. Without clear documentation and justification of these kinds of decisions made *during* the evaluation, it is difficult for outsiders to know how well the reported results fully reflect the strengths and weaknesses of an AI system.

Take, for instance, the most recent model system cards from some of the largest frontier AI developers (e.g. [4, 50, 102]). Although they provide a certain level of accountability, all of them include very limited information about why the reported benchmarks and human evaluations were selected, whether the model was tested on other, unreported benchmarks, why some behaviours were evaluated while others were not, or whether the test data overlap with the

training data [148]. In general, benchmark selection is justified primarily by "popularity", with limited additional rationale. Without more information, it is impossible to know whether the findings show a representative view of the model's behaviour or whether they are selectively presented in a way that overstates model performance and safety.

These issues are not unique to AI. For example, the incentives for developers to report positive results are analogous to the incentives for drug developers to show that their drugs are effective. In medicine, these incentives created a culture in which ineffectiveness or evidence of side effects were frequently buried or minimised [38, 67, 91, 120, 143]. Likewise, other scientific fields such as psychology and physics have been wrestling with evidence that incentives for researchers to report positive, surprising findings frequently led to questionable research practices such as selective reporting, p-hacking, and even data fabrication, polluting the literature with findings that do not replicate [34, 68, 126]. If left unchecked, these incentives risk having a similar effect on AI [107].

Fortunately, other fields offer established solutions. In medicine, most jurisdictions require the methods and analysis plans for clinical trials to be registered and approved ahead of time by relevant regulatory bodies. Similarly, many publication venues and funding bodies in the sciences such as psychology and physics have adopted policies that require or incentivise researchers to document and register their research plans ahead of time in a process known as "pre-registration". These processes help prevent questionable research practices such as selective reporting and p-hacking, and tend to reduce false positives [16, 26, 120]. In addition to providing important transparency, pre-registration serves as a useful planning tool that helps researchers clearly think through the details of their methods and analysis plan to ensure these are well-thought-out before any expensive data collection is conducted. We can also take inspiration from engineering, which regularly tests prototypes and final products for both performance and safety, from car manufacturing to software engineering [142]. Likewise, testing is an integral part of software engineering, and data mining and data science have introduced methodologies to guide the development of similar kinds of projects [28, 85, 123].

Given the effectiveness of these approaches in other disciplines, and the existence of similar problems in AI [3, 21, 39, 58, 89, 131, 135], we suggest a similar approach could be taken in AI evaluation. We therefore propose here a standardised evaluation protocol that can provide both clear guidance about the steps that should be taken in evaluation, and as a template for documenting and reporting evaluations. The protocol can act as a guide to help researchers and evaluators structure their evaluation projects and a checklist to ensure best practices, paving the way for the development of a standard for anticipative reporting that could be adopted by industry, third-party evaluators, publication venues, governments (e.g., AISIs) or regulatory bodies (e.g., the EU AI Office). The protocol itself can also serve as a common reference point for discussions about evaluation methods and improve understanding of key components of evaluations.

Table 1 summarises some main reasons for a standard protocol for AI evaluation. A protocol should go beyond post-hoc reporting and should accommodate the wide ecosystem of researchers and evaluators, from small teams in academia, startup providers integrating frontier models for new applications, frontier developers themselves, and independent evaluation organisations assessing systems across a range of parties. It is important that the protocol is there before starting the evaluations and guides those teams with limited experience (or reminds more experienced evaluators) in the process of evaluation before it starts, to determine what steps need to be taken and in which order.

In this paper we present `PREP-Eval`, which is, to our knowledge, the first comprehensive, step-by-step protocol for AI evaluation. However, the protocol draws inspiration from ideas and initiatives that have proved very useful in AI itself and other disciplines. The structure of `PREP-Eval` is similar to CRISP-DM [28, 85], a widely-adopted protocol for data-mining and data-science projects that lays out a sequence of phases, each of which describes in detail the tasks and outcomes of that phase. The protocol also draws from methodologies in software testing, and cybersecurity—which focus

Table 1. Main reasons and associated methodological problems PREP-Eval aims to address.

Cause	Outcome
Evaluations not well designed or insufficiently scoped	Evaluation projects are inefficient, unfocused, ineffective or opaque, lacking standard project-management practices. - *Example*: In a close analysis of four benchmarks used to evaluate fairness in natural language processing (StereoSet, CrowS-Pairs, WinoBias, and WinoGender), Blodgett et al. (2021) found that all benchmarks revealed severe weaknesses in terms of defining what is being measured [39].
Over-emphasis on pre-release evaluation	Evaluations are focused on capability demonstrations and stylised benchmark tasks, creating blind spots that emerge only under realistic use, or sustained interactions over time. - *Example*: Raji et al. (2022), describe that AI evaluation practices often skip over the question of whether a given system functions, or provides any benefits at all, leading to faulty AI products that are on the market.
Evaluation bias due to non-systematic methodology	Evaluations are not trusted. Results could be misinterpreted, providing a misleading picture of system capabilities and limitations. - *Example*: Ivanova (2023a) analyses two cases –the Winograd Schema Challenge and Theory of Mind evaluation–, which the author characterises as instances of over-attribution, showing impressive model performance that turned out to be substantially worse once relevant methodological controls were introduced.
Insufficient process documentation	Lack of coordination and difficulty for scaling up cooperative evaluations, especially in big or changing teams. - *Example*: In their benchmark quality evaluation framework, Reuel et al. (2024b) find that 17 of the 24 benchmarks analysed do not provide easy-to-run scripts to replicate the results reported in the original paper. This lack of accessibility hinders reproducibility and limits users' ability to scrutinise the benchmarking process in a field where reproducibility is already a significant concern.
Absence of standardisation	Difficulty in sharing the evaluation information with other stakeholders and policy-makers, for auditing and regulation. - *Example*: Bordes et al. (2025) argue that evaluation methodologies lack systematic documentation standards comparable to those used for datasets and models, leading to reduced transparency and reproducibility of AI evaluation. Eval Factsheets solve some of the problems but do not ensure standardised *processes* and can just wash questionable practices *a posteriori*, giving a false sense of quality.

largely on failures—as well as from psychological evaluation and the scientific method—which focus on understanding behaviour and capabilities. The following section covers related work in all these areas.

3 Related Work

Evaluation in AI tries to explain and infer AI behaviour by analysing data. Many evaluative endeavours in AI and other areas share similarities in the way experimental research questions are produced and answered by the data, but approach the process differently. We review these approaches here and what we draw from each of them.

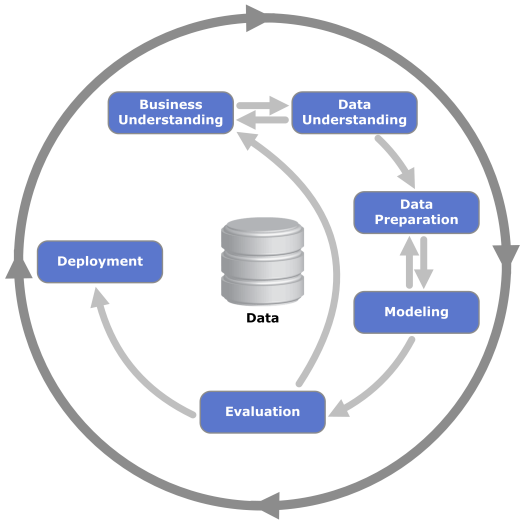

Fig. 2. CRISP-DM life cycle (image taken from Wikimedia with CC licence).

3.1 Evaluation Processes in Data Analysis

To develop PREP-Eval, we draw here on ideas and protocols from a field that is defined by analysing empirical data, such as data mining, data science and data analytics, even if these areas fall under the umbrella of "analysis" rather than "evaluation".

Actually, one the protocols used in data analysis will be our main inspiration. The CRoss-Industry Standard Process for Data Mining (CRISP-DM), shown in Figure 2, provides a historical reference point for a standardised AI-evaluation protocol. Created by a large consortium in the late 1990s [28], CRISP-DM and its many evolutions [85] prioritise simplicity, pragmatism, and proportionality; as a result, they are still the gold standard protocol for planning and documenting goal-oriented data-science projects, prevailing over more complex approaches despite substantial changes in the discipline over the intervening decades.

Data mining involves many similar steps to AI evaluation, including data collection and analysis, making it an excellent starting point. However, several factors mean that CRISP-DM does not perfectly translate to AI evaluation. First, data mining typically involves gathering already existing data and generating a statistical or machine learning model to explain the data, so there is little focus on how to design and conduct experiments. AI evaluation usually requires many experiments or interactions with an AI system to generate the evaluation data that are then analysed. In other words, CRISP-DM is at the modelling level while PREP-Eval is at the meta level.

As shown in Figure 2, CRISP-DM clearly lays out a step-by-step guide for how to conduct and document a data mining project from start to finish. The process specification emphasises the importance of clearly specifying project goals and creating a project plan, and each substage produces an output that is documented.

3.2 Evaluation Processes in Software and Cybersecurity

Given that AI systems are composed of software (running over hardware), it is unsurprising that AI evaluation is related to software verification and testing. Much like software tests, some AI evaluations are designed to test whether

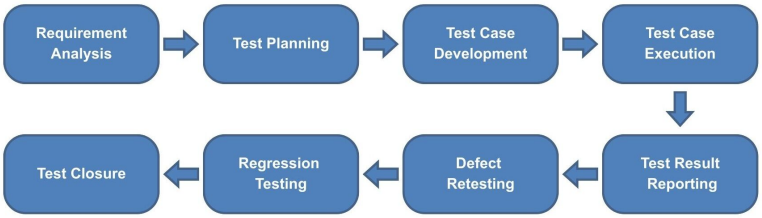

Fig. 3. Software Testing Lifecycle according to Jamil et al. (2016)

a system will fail under certain conditions. However, AI evaluations often have a much broader scope than software testing. Software is usually designed with a specification in mind, which can then be systematically verified and tested. By contrast, the capabilities and other properties governing the behaviour of AI systems are often not well defined or understood. As a result, evaluations are often designed to understand a system better, rather than just to identify failure points. In addition, unlike software tests that have clear pass or fail conditions, there may not be a single "right answer" to many of the tasks AI systems are used for (e.g., generating an image from a prompt).

Nonetheless, it is worth looking to software testing processes for inspiration. There are several standard protocols for software testing, but none is seen as the gold standard [72], but software testing often broadly follows the Software Testing Lifecycle [35, 123] pictured in Figure 3. In this lifecycle we see analysis and planning stages preceding all the other stages, and a very linear flow, at least graphically, although some tasks may overlap a bit and can be merged into larger blocks for some other variants of the software test lifecycle.

In cybersecurity, red teaming implies an adversarial scrutiny of a system or network usually performed by a team of people (the "red" team) that operates adversarially. The objective is to break into, hack, or damage a system, conceiving any possible malicious attack or negligent use of the system.

While red teaming is very similar to stress testing in software testing, which focuses on extreme cases and inputs that may lead to hazards, it encapsulates the effort as a team endeavour. The specific methodology is based on exploring the boundaries and performing many iterations.

3.3 Pre-registration in the Experimental Sciences

In scientific fields such as medicine, psychology, physics and economics, evaluations differ according to the particularities of each domain but share similar concerns about some research incentives that may lead to rushed or tweaked results. These concerns over replicability and questionable research practices have led to the adoption of "pre-registrations" [51, 66, 93, 100, 101, 144, 145]. A pre-registration is a planning document usually created prior to data collection that describes the basic elements of the scientific method applied to a specific project—hypotheses, methods, and analysis plan. Before data collection or inspection begins, this document is uploaded to a time-stamped public repository such as the Open Science Framework (although kept hidden until the researchers decide to make it public). In some cases, pre-registrations are also submitted for peer review, where reviewers may suggest revisions or minor adjustments to the proposed plan before it is finalised [26, 127, 145]. When the data are later submitted for publication, a link to the pre-registration is included so that reviewers and readers are able to check how closely the methods and analysis used in the study align with the pre-registered plan, making it clear which aspects of the reported results were planned and which were more exploratory. In doing so, pre-registrations help prevent and raise awareness of questionable

research practices such as selective reporting and p-hacking [57, 97, 119]. It is important to note that pre-registrations do not preclude researchers from deviating from their plan or conducting exploratory analyses, they simply make these deviations transparent.

Pre-registration fits `PREP-Eval` well because scientific research contains many of the key components of AI evaluations, including experimental design, data collection, and data analysis, as well as similar incentives. In fact, pre-registration has already been suggested for AI [61, 87, 104, 107, 134].

3.4 Evaluation Processes in AI

Ideas from all of the previous domains have been brought to AI evaluation. For instance, red teaming has been extensively applied to LLMs recently [25, 43, 48, 79, 106]. More generally, adversarial testing [74, 109, 149], presents a lifecycle in which evaluations are designed to detect examples where models fail, and modify benchmarks accordingly. This iterative view of evaluation also influences our protocol, `PREP-Eval`.

Recent work has also drawn inspiration from other disciplines to improve or formulate recommendations for AI evaluations (e.g., [104]). For instance, Microsoft (2025) incorporated insights from diverse fields to address governance-related questions, such as determining the most appropriate stages of the lifecycle for conducting evaluations. Other efforts have even extended to incorporate insights from the evaluation of cognition in animals [112]. While these works offer meaningful general recommendations, they do not propose a step-by-step protocol.

Still, the evaluation processes followed in AI evaluation mostly come from practices, standards, and recommendations that have been proposed for evaluating machine learning systems [17, 44, 73], natural language processing (NLP) models [13, 49] and various conceptions of General-Purpose AI (even before this term was not associated with LLMs [8, 10, 59]). As concerns about the reliability of evaluations have grown, similar efforts have emerged to outline best practices or standards for evaluating LLMs [6, 9, 45, 52, 65, 70, 78, 83, 103, 104, 136], generative AI systems more generally [71, 137, 147] and agents [24, 140, 155]. Other guidelines have focused on the evaluation instruments, such as benchmarks and competitions [3, 22, 33, 56, 81, 84, 114] or specific elements of the evaluation process, including reporting, operating conditions and documentation [87, 128, 130]. However, most of these initiatives remain at the level of general principles and conceptual guidance.

Our proposal introduces a step-by-step protocol that operationalises these recommendations through an actionable, sequential, and structured process. Although some existing frameworks include checklists with concrete actions (e.g., [104, 114, 117, 118]), the sequential architecture of our protocol provides distinct advantages. Most notably, it facilitates traceability, establishing a clear audit trail that enables researchers and practitioners to trace each finding or potential error back to its origin in the evaluation pipeline, starting from the design of the evaluation and the allocation of resources. This structure not only organises and streamlines the evaluation process, but also makes it more accessible to practitioners and researchers across a wide range of backgrounds.

Other works have taken a descriptive approach to identifying the elements or dimensions of evaluations that can guide their design [23, 37]. Paskov et al. (2025) propose a series of recommendations for human uplift studies, benchmark evaluations, and cross-cutting recommendations for all evaluation types. Recommendations are organised into four distinct stages of the evaluation process. In the design stage, they recommend pre-registration as a gold standard, similar to McCaslin et al. (2025)'s proposals for chemical and biological benchmarks; and UK AI Security Institute (2024) has likewise noted that it uses pre-registration to accelerate evaluation workflows while maintaining quality. Yet, all of these works provide few details on how pre-registration should be implemented.

PREP-Eval embeds pre-registration throughout the evaluation process, helping identify the specific points at which evaluators might revisit and potentially deviate from their initial plan, a possibility that may encourage adoption. The incorporation of fine-grained, task-specific checklists within each sub-stage helps prevent omissions or makes them transparent, and enables an effective division of labour within evaluation teams. These design choices complement higher-level recommendations by clarifying not only what should be considered, but also *what concrete actions* should be taken. This ensures that methodological decisions are explicitly articulated and justified from the outset, prompting evaluators to be more clear about their decisions, and write them in a language that it could (potentially) be understood in an easier way when revisited by themselves later or by third parties..

4 The PREP-Eval Protocol

The PREP-Eval protocol has six phases as shown in Figure 1. Each phase should be planned and documented to ensure each step is clear and sensible. At the end of phase 3 there is a pre-registration that can be private or public, allowing selected stakeholders to give feedback during or at the end of the project, to check results with the original plan. Pre-registration could be shared with secure third-parties or review boards or the general public, depending on the context and content of the evaluation. At the end of phase 6, a full report can be documented (and potentially released following similar considerations).

Table 2 and 3 present the reference framework for the protocol summarised at the level of phases and tasks. Following the structure of CRISP-DM, we consider a hierarchical breakdown, with phases, generic tasks and specialised tasks in three levels. Here we only show the name of the phases and generic tasks at the first level, while Appendix A contains the complete user guide, which presents a more detailed version of the protocol including the specific outputs and outcomes of each task.

The protocol was intentionally designed to be cyclical in nature. After a full iteration of the cycle, the nature of the protocol leads to a state where knowledge about the system has increased, and new evaluation goals can be established to ignite another iteration. Likewise, as planning progresses, there may be a need to go back to an earlier stage and modify the plan. For example, when looking at available evaluation benchmarks in phase 2.1, a lack of candidate benchmarks may force a rethink of the evaluation goals laid out in phase 1. The protocol can also be used to trace the evolution of model development or deployment over time. An organisation might want to do evaluations continuously over time or at checkpoints. One approach is to use a similar idea to the dev-test split—specify up front the tests that will be used to check progress during development versus tests used once the final model has been developed. Evaluations on the same system can also be performed simultaneously. For instance, one team can evaluate the reasoning capabilities of a language model while another is performing red teaming for safety.

One key aspect of the protocol is the identification of an "estimator" as a generalised notion for the main result of an evaluation. This estimator, in the statistical sense, can range from a position indicator of a metric, such as the result of a benchmark, to a sophisticated cognitive model that captures the capabilities of an AI system to predict performance for new tasks. It can also be an estimator of risk that is extracted as the result of the evaluation, taking into account the context of use and other operating conditions. Accordingly, even in red teaming the goal would not be to find as many issues as *possible* in a system, but to estimate the risk depending on the probability and impact of issues found.

Another key element of the protocol is the distinction between the project goals and the technical objectives, which critically determine phases 1 and 2 of the protocol. Many evaluations start with the question of what to measure, but fail to state the purpose of the evaluation more clearly, which has an influence—and may bias—the whole process.

PREP-Eval Protocol	
Phase 1: Goals and Objectives	
Evaluations can serve many purposes, e.g., identifying system failures, comparing capabilities of two different models, or tracking improvements across versions. Because each goal may require a different methodology, it is essential to clearly define the evaluation's objectives and rationale from the outset.	
1.1 Determine project purpose	Describe the relevant background to the evaluation project, including the terminology, project goals, and success criteria.
1.2 Determine technical objectives	Identify and justify the targets of the evaluation, e.g., an AI system or a new evaluation method; and describe the success criteria of the evaluation in terms of metrics and uncertainty of the estimators.
1.3 Situation assessment	Develop an inventory of resources, identify requirements and constraints, anticipate risks and contingencies, and assess current understanding of the evaluation target(s).
Phase 2: Evaluation Design	
Evaluations could combine standardized benchmarks, red teaming, and human ratings, among other methods. Choosing the right approaches requires careful consideration of the evaluation target(s), system characteristics, and available tools.	
2.1. Identify potential evaluation methods	Familiarize yourself with current evaluation methods, assess maturity and adoption, summarize strengths and limitations and monitor emerging work. For example, for red teaming, identify processes for generating adversarial inputs (automated, manual, or hybrid). For human evaluations, consider crowdsourcing platforms used to collect responses.
2.2. Selection of evaluation methods	Select an evaluation method and rigorously justify your choice. If no available methods are appropriate, design or build new methods.
2.3. Analysis specification	Decide and justify how the evaluation data will be analysed and what estimators will be produced. This could include summary statistics, metrics, error analysis or building prediction models.
Phase 3: Project Plan	
Create a plan to guide the design, execution, and analysis phases, ensuring alignment between goals, methods, and expected outcomes.	
3.1. Create a project plan	Draft an initial project plan. This may include the major stages of the evaluation process, a realistic timeline, the resources required, the expected outputs and deliverables of the project, and any other relevant information gathered during the previous planning phases. Distribute the plan for review and input and consolidate it into a final version.
3.2. Pre-register evaluation	Submit a "pre-registration" of the protocol to a time-stamped repository for potential feedback. This pre-registration, covering up to Phase 3.2, should be complemented later alongside project outputs at the end of phase 6.

Table 2. PREP-Eval Protocol, covering Phases 1–3 (continued in Table 3).

This distinction, along with the pre-registration, will improve trust in the results, which will be beneficial both to the developers (third parties will believe their claims) and to the broader community.

Phase 4: Data Collection	
The complete set of processes for generating or obtaining the experimental data needed for analysis, including the setup and dataset annotations that are needed before running experiments with AI systems, or procuring and processing evaluation data that is already available.	
4.1. Experimental setup, annotations and pilots	Verify data quality and integrity (e.g., if existing datasets will be used, ensure they are accessible), run experimental samples, determine the annotation setup, develop filters and classifiers, and conduct pilot tests. If issues are identified, adjust the protocol as needed and document all changes.
4.2. Full data collection	Run full experiment and obtain full data from the AI system, verifying that the data collected follows the pre-defined sampling strategy.
4.3. Data preparation	Clean, format and organize evaluation data. Verify data quality.
Phase 5: Data Analysis	
This phase involves systematically analysing the evaluation data, from data exploration and visualisation to the derivation of the estimators, including additional analyses on unexpected patterns.	
5.1. Initial data exploration	Preliminary exploration of evaluation data, including variation across task features. Identify unexpected patterns and adjust the analysis plan.
5.2. Conduct planned analysis	Perform analyses according to analysis plan: aggregate data and calculate summary statistics/metrics, build performance breakdowns, calculate inferential statistics and build prediction models.
5.3. Assess and refine analysis	To ensure the robustness and interpretability of the analyses quantify uncertainty, test the assumptions behind the analytical methods and inspect any unusual results.
Phase 6: Conclusions and Review	
This final phase concludes the evaluation by synthesizing findings, reflecting on the process, capturing lessons learned, and archiving materials along with the pre-registration.	
6.1. Draw conclusions	Synthesize the analytical findings to derive conclusions about the evaluation target(s), considering the limitations of the process.
6.2. Review evaluation process	Examine what aspects of the process worked effectively and which did not, recommend improvements for future evaluations and describe the project legacy.
6.3. Determine next steps	Define how and where to communicate results. Decide on next steps (e.g., additional training, further evaluation, deployment).
6.4. Complete the registration	Write the final report. Document and explain any deviations from the pre-registered plan. Submit the final report to ensure transparency and reproducibility of the evaluation effort.

Table 3. PREP-Eval Protocol, covering Phases 4–6 (continued).

5 Using the Protocol

As discussed in the introduction, soon after the decision to evaluate an AI system is made, a first instantiation of the protocol — at least up to phase 3.2 — should be completed (*before* actually conducting the evaluation). Public pre-registration at this stage discourages covert "post-registration" of studies that have already been conducted, because

evaluators can receive feedback that may lead to modifications of the original plan. The extent to which this feedback is incorporated depends on the context; for example, even the requirement by oversight bodies, review committees, or external auditors to implement minor changes in the plan can prevent this practice [26]. This first version of the written protocol, including the specification, design and plan can serve as a reference for the whole process.

Of course, some of the stages will be refined as the evaluation process progresses, and the possible and expected outcomes will be replaced by actual outcomes. Because of these modifications, it is very important that the original instantiation of the protocol is pre-registered, both for transparency and also to have a clear perspective of how much the evaluation has deviated from the original expectation and plan. At the end of the evaluation, the final report can be registered alongside the pre-registration and can serve for comparison.

These two "registration" steps at phases 3.2 and 6.4 are the desirable minimum. The protocol can be used in a more agile or flexible way, by using a versioning system (such as GitHub), documenting how the evaluation and the associated outcomes for each subphase are evolving. For instance, in organisations that adopt agile methodologies such as Scrum [121], a first stint can perform a lightweight pass from stages 1 to 3, do the pre-registration, get feedback from stakeholders, revise it, and then do a full pass from 1 to 6 on another stint. Then, the protocol can be refined with several iterations or stints in a cyclical way.

The description of phases and tasks as discrete steps performed in a specific order represents an idealised sequence of events. In practice, many of the tasks can be performed in a different order and it will often be necessary to repeatedly backtrack to previous tasks and revise certain actions.

In order to illustrate how the protocol is applied, the following appendices contain the application of the protocol to several evaluation situations, some of them being more traditional (selecting between models) to more complex (building a complex estimator for instance-based performance). In the first five scenarios, the protocol is applied to existing studies with real evaluations, aiming to follow their content as faithfully and comprehensively as possible, although the available information is often limited, requiring careful adaptation, and some details may have been misinterpreted. The final scenario is hypothetical. Moreover, in certain instances, `PREP-Eval` is applied only up to phase 3 (pre-registration), whereas in others, we carry out the full protocol. In these cases where the full protocol is executed, we take a more review/audit perspective, highlighting elements that would typically be expected in a robust evaluation process but are absent from the original papers, which were published prior to the introduction of this protocol.

In particular, the appendices cover these use cases:

- Appendix C: Red-teaming GPT-3 to find prompts with a high rate of false statements. We include all phases.
- Appendix D: Evaluating Interactional Fairness in Multi-Agent LLM-Based Systems. We include all phases.
- Appendix E: Evaluate the performance of an LLM-based customer service agent. We only include phases 1 to 3.
- Appendix F: A meta-evaluation for diversity and coverage of test cases. We only include phases 1 to 3.
- Appendix G: Evaluating the capabilities of LLMs that incorporate metacognition. We include all phases.
- Appendix H: Choosing between two AI systems, in a face recognition domain. We only include phases 1 to 3.

We plan to associate the protocol with a repository of sample evaluations, building from existing repositories of AI cases and evaluation [1, 20, 42, 75], for practitioners to take inspiration from.

6 Limitations

Two common objections to pre-registration are that they require too much work up front and that they prevent exploration. With regard to the first objection, pre-registration does indeed front-load some of the work that might

otherwise be done later (e.g., specifying the analysis plan). However, doing this work before expensive data collection is conducted is essential to prevent undesirable surprises later in the process. For example, without proper planning, one might realise before collecting data that there are not enough data points to robustly perform the desired analysis and that the there are no tasks measuring an important ability. The second objection is likewise unfounded. Pre-registrations are not intended to restrict what is done in the project, they simply make transparent which decisions were made ahead of time and which were exploratory in nature. Deviations from the pre-registered protocol are completely acceptable, provided the deviations can be explained clearly.

There are important parts of the evaluation process not covered in detail in this protocol. The first of these is how to design new evaluation tools when current tools are insufficient. The best way to go about this depends strongly on the type of evaluation being conducted and the targets of the evaluation. Moreover, new techniques to build evaluations are being developed all the time (e.g., crowdsourced benchmarks [30, 80, 152], LLM augmented evaluations [5, 46], multi-agent frameworks [29, 53, 116] and predictive frameworks [153], so we refrain from prescribing any particular approach. The second part is how to use the evaluation results to make decisions about whether (and where) a model should be deployed. The suitability of a model is highly dependent on the context—for example, a 1% error rate on an image classifier may be acceptable, but a 1% error rate on a self-driving car may be catastrophic.

Naturally, PREP-Eval has several limitations originating from the state of AI as a field and the development of the protocol itself (see Appendix B for more details on how the protocol was developed). In the first place, there is limited theory and practice on how to extract capabilities rather than performance, or how to evaluate general-purpose systems, which are new, compared to more traditional specialised task-oriented systems. Because the protocol is meant to be accessible to small and medium-sized teams and evaluations, it is also unlikely that the protocol can contemplate all different kinds of evaluation equally well or all the process a big lab has to conduct for a frontier model. Despite the generality of PREP-Eval, some flexibility in its use is expected and appropriate.

In some contexts, the use of PREP-Eval may be counterproductive. In particular, the protocol is not well suited to evaluations conducted under very short timelines, to purely exploratory analyses, or to well-established routine evaluations where methods, metrics, and decision criteria are already standardized and well understood. By contrast, PREP-Eval is most valuable when evaluation results are intended to function as evidence—such as in audits, external reviews, or high-stakes decision-making contexts—where trust and traceability are critical.

The protocol may also be more resource-intensive than alternative approaches, as it requires sustained documentation throughout the evaluation process rather than only at the reporting stage. However, this cost is partly offset by downstream benefits, including reduced risk of wasted resources due to poorly specified evaluation designs, the creation of reusable evaluation artifacts, and the accumulation of institutional knowledge that can support future evaluations.

Finally, while PREP-Eval is designed primarily for evaluations in which the AI system itself is the central evaluation target, it is less directly applicable to evaluation paradigms where AI is involved but not the primary object of study. For example, human uplift studies, which assess how access to or use of an AI system affects human performance relative to baseline tools (e.g., internet search) [104], and other evaluations of AI impact on human cognition or healthcare could be grounded in established experimental practices from psychology and the social sciences. Such evaluations may therefore require complementary or alternative protocols better aligned with those disciplinary traditions, or an extension in future versions of PREP-Eval.

7 Conclusion

Here we have presented PREP-Eval, the first protocol for conducting evaluations of AI systems, which we propose as a standard for practitioners and researchers. This six-phase protocol provides a guide for how to conduct an evaluation from planning to completion, taking inspiration from standardised processes in a variety of other fields. To the best of our knowledge, this is the first attempt to create a standardised process for AI evaluations with pre-registration. Protocols like this one have been tremendously useful and beneficial in other fields that have adopted them, such as data mining, software testing and psychology [63, 85, 105, 139], acting as an integrated framework for planning, certification, audit, documentation and standardisation. In the particular case of AI evaluation, specifying the evaluation a priori in a time-stamped document (pre-registration) allows the community to be sure that questionable research practices were not used to make the model seem better than it really is when developers and evaluators come from the same organisation or have common interests.

This document presents a proposal for a protocol, which must evolve with the feedback given by the community and especially by its users. It is not the goal of this paper to discuss the platforms and institutions that will keep PREP-Eval evolving in the years to come, or to maintain a repository of pre-registrations and evaluations (including all detailed results), but this will be needed to ensure the protocol adapts and grows in the years to come.

Generative AI Usage Statement

The authors used generative AI tools to assist with the paraphrasing and editing of selected portions of the manuscript to improve format, clarity and language, followed by human checks and revision. All ideas, analyses, experimental designs, and interpretations are the original work of the authors. No generative AI tools were used to generate research ideas, data, or results.

Disclaimer

The views expressed in this publication do not necessarily reflect those of the European Commission and are made in a personal capacity. The European Commission and any person acting on behalf of it are not responsible for the use that might be made of this publication.

References

[1] Alexandra Abbas, Celia Waggoner, and Justin Olive. 2025. Developing and maintaining an open-source repository of AI evaluations: Challenges and insights. *arXiv preprint arXiv:2507.06893* (2025).

[2] NIST AI. 2024. Artificial intelligence risk management framework: Generative artificial intelligence profile. *NIST Trustworthy and Responsible AI Gaithersburg, MD, USA* (2024).

[3] Omar Alonso and Kenneth Church. 2025. Evaluating the Evaluations: A Perspective on Benchmarks. In *ACM SIGIR Forum*, Vol. 58. ACM New York, NY, USA, 1–27.

[4] Anthropic. 2025. System Card: Claude Opus 4.5. https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf

[5] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems* 36 (2023), 78142–78167.

[6] Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, et al. 2025. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. *arXiv preprint arXiv:2511.04703* (2025).

[7] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. 2024. Managing extreme AI risks amid rapid progress. *Science* 384, 6698 (2024), 842–845.

[8] Sankalp Bhatnagar, Anna Alexandrova, Shahar Avin, Stephen Cave, Lucy Cheke, Matthew Crosby, Jan Feyereisl, Marta Halina, Bao Sheng Loe, Seán Ó hÉigeartaigh, et al. 2017. Mapping intelligence: Requirements and possibilities. In *3rd Conference on" Philosophy and Theory of Artificial Intelligence*. Springer, 117–135.

[9] Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782* (2024).

[10] Jordi Bieger, Kristinn R Thórisson, Bas R Steunebrink, Thröstur Thorarensen, and Jona S Sigurdardottir. 2016. Evaluation of general-purpose artificial intelligence: why, what & how. *Evaluating General-Purpose AI* 9 (2016).

[11] Ruta Binkyte. 2025. Interactional Fairness in LLM Multi-Agent Systems: An Evaluation Framework. *arXiv preprint arXiv:2505.12001* (2025).

[12] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1004–1015.

[13] Iva Bojic, Jessica Chen, Si Yuan Chang, Qi Chwen Ong, Shafiq Joty, and Josip Car. 2023. Hierarchical evaluation framework: Best practices for human evaluation. *arXiv preprint arXiv:2310.01917* (2023).

[14] Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems* 35 (2022), 3663–3678.

[15] Florian Bordes, Candace Ross, Justine T Kao, Evangelia Spiliopoulou, and Adina Williams. 2025. Eval Factsheets: A Structured Framework for Documenting AI Evaluations. *arXiv preprint arXiv:2512.04062* (2025).

[16] Abel Brodeur, Scott Carrell, David Figlio, and Lester Lusher. 2023. Unpacking p-hacking and publication bias. *American economic review* 113, 11 (2023), 2974–3002.

[17] Olivia Brown, Andrew Curtis, and Justin Goodwin. 2021. Principles for evaluation of ai/ml model performance and robustness. *arXiv preprint arXiv:2107.02868* (2021).

[18] Miles Brundage. 2025. Feedback on the Second Draft of the General-Purpose AI Code of Practice. Online submission. https://milesbrundage.substack.com/p/feedback-on-the-second-draft-of-the

[19] Marie Davidsen Buhl, Gaurav Sett, Leonie Koessler, Jonas Schuett, and Markus Anderljung. 2024. Safety cases for frontier AI. *arXiv preprint arXiv:2410.21572* (2024).

[20] John Burden, Marko Tešić, Lorenzo Pacchiardi, and José Hernández-Orallo. 2025. Paradigms of AI evaluation: Mapping goals, methodologies and culture. *arXiv preprint arXiv:2502.15620* (2025).

[21] 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. 2023. Rethink reporting of evaluation results in AI. *Science* 380, 6641 (2023), 136–138.

[22] Jialun Cao, Yuk-Kit Chan, Zixuan Ling, Wenxuan Wang, Shuqing Li, Mingwei Liu, Ruixi Qiao, Yuting Han, Chaozheng Wang, Boxi Yu, et al. 2025. How Should We Build A Benchmark? Revisiting 274 Code-Related Benchmarks For LLMs. *arXiv preprint arXiv:2501.10711* (2025).

[23] María Victoria Carro, Denise Alejandra Mester, Francisca Gauna Selasco, Luca Nicolás Forziati Gangi, Matheo Sandleris Musa, Lola Ramos Pereyra, Mario Leiva, Juan Gustavo Corvalan, María Vanina Martinez, and Gerardo Simari. 2025. A Conceptual Framework for AI Capability Evaluations. *arXiv preprint arXiv:2506.18213* (2025).

[24] Stephen Casper, Luke Bailey, Rosco Hunter, Carson Ezell, Emma Cabalé, Michael Gerovitch, Stewart Slocum, Kevin Wei, Nikola Jurkovic, Ariba Khan, et al. 2025. The ai agent index. *arXiv preprint arXiv:2502.01635* (2025).

[25] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023. Explore, establish, exploit: Red teaming language models from scratch. *arXiv preprint arXiv:2306.09442* (2023).

[26] Christopher D Chambers and Loukia Tzavella. 2022. The past, present and future of Registered Reports. *Nature human behaviour* 6, 1 (2022), 29–42.

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

[28] Peter Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. 1999. CRISP-DM 1.0: Step-by-Step Data Mining Guide. https://the-modeling-agency.com/crisp-dm.pdf

[29] Jiaju Chen, Yuxuan Lu, Xiaojie Wang, Huimin Zeng, Jing Huang, Jiri Gesi, Ying Xu, Bingsheng Yao, and Dakuo Wang. 2025. Multi-agent-as-judge: Aligning LLM-agent-based automated evaluation with multi-dimensional human evaluation. *arXiv preprint arXiv:2507.21028* (2025).

[30] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

[31] Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen. 2024. Safety cases: How to justify the safety of advanced AI systems. *arXiv preprint arXiv:2403.10462* (2024).

[32] Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. 2024. Cogbench: a large language model walks into a psychology lab. *arXiv preprint arXiv:2402.18225* (2024).

[33] Anthony G Cohn and José Hernández-Orallo. 2023. A framework for characterising evaluation instruments of AI performance. (2023).

[34] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[35] Giuseppe De Nicola, Pasquale di Tommaso, Esposito Rosaria, Flammini Francesco, Marmo Pietro, and Orazzo Antonio. 2005. A grey-box approach to the functional testing of complex automatic train protection systems. In *European Dependable Computing Conference*. Springer, 305–317.

[36] Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael C Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in*

Neural Information Processing Systems 37 (2024), 19783–19812.

[37] P Alex Dow, Jennifer Wortman Vaughan, Solon Barocas, Chad Atalla, Alexandra Chouldechova, and Hanna Wallach. 2024. Dimensions of Generative AI Evaluation Design. *arXiv preprint arXiv:2411.12709* (2024).

[38] Kerry Dwan, Carrol Gamble, Paula R Williamson, Jamie J Kirkham, and Reporting Bias Group. 2013. Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PloS one* 8, 7 (2013), e66844.

[39] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation. *arXiv preprint arXiv:2502.06559* (2025).

[40] European Commission. 2025. General-Purpose AI Code of Practice.

[41] Eval Eval Coalition. [n. d.]. Evaluation Cards. Research project. https://evalevalai.com/projects/eval-cards/ Chairs: Anka Reuel, Avijit Ghosh.

[42] Evidently AI. 2025. ML and LLM System Design: 800 Case Studies to Learn From. Online database. https://www.evidentlyai.com/ml-system-design First published June 14, 2023; last updated December 22, 2025.

[43] Michael Feffer, Anusha Sinha, Wesley H Deng, Zachary C Lipton, and Hoda Heidari. 2024. Red-teaming for generative AI: Silver bullet or security theater?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 421–437.

[44] Luciana Ferrer, Odette Scharenborg, and Tom Bäckström. 2024. Good practices for evaluation of machine learning systems. *arXiv preprint arXiv:2412.03700* (2024).

[45] Clémentine Fourrier, Thibaud Frere, Guilherme Penedo, and Thomas Wolf. 2025. The LLM Evaluation Guidebook. Hugging Face. https://huggingface.co/spaces/OpenEvals/evaluation-guidebook#what-is-model-evaluation-about Online resource.

[46] Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, and Ion Stoica. 2025. Prompt-to-Leaderboard: Prompt-Adaptive LLM Evaluations. *Forty-second International Conference on Machine Learning* (2025).

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

[48] Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. 2024. Mart: Improving llm safety with multi-round automatic red-teaming. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1927–1937.

[49] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. A Case for Better Evaluation Standards in NLG. In *Workshop on Setting up ML Evaluation Standards to Accelerate Progress at ICLR*, Vol. 2022.

[50] Google DeepMind. 2025. Gemini 3 Flash: Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf. Model card published December 2025.

[51] Carola Grebitus and Wuyang Hu. 2025. Agricultural and applied economists' views on pre-registration and pre-analysis plans for empirical research. *Journal of the Agricultural and Applied Economics Association* 4, 2 (2025), 223–238.

[52] Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. 2025. Olmes: A standard for language model evaluations. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 5005–5033.

[53] Shengyue Guan, Haoyi Xiong, Jindong Wang, Jiang Bian, Bin Zhu, and Jian-guang Lou. 2025. Evaluating llm-based agents for multi-turn conversations: A survey. *arXiv preprint arXiv:2503.22458* (2025).

[54] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736* (2023).

[55] Neha R Gupta, Jessica Hullman, and Hariharan Subramonyam. 2024. A conceptual framework for ethical evaluation of machine learning systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 534–546.

[56] Moritz Hardt. 2025. The emerging science of machine learning benchmarks. *Manuscript. https://mlbenchmarks. org* (2025).

[57] Tom E Hardwicke and John PA Ioannidis. 2018. Mapping the universe of registered reports. *Nature Human Behaviour* 2, 11 (2018), 793–796.

[58] Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M Asmar, Sanmi Koyejo, Michael S Bernstein, and Mykel John Kochenderfer. 2025. More than marketing? on the information value of ai benchmarks for practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 1032–1047.

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

[60] Jose Hernandez-Orallo, Sean O hEigeartaigh, Alexandru Marcoci, Haydn Belfield, Giulio Corsi, Maurice Chiodo, and Fernando Martinez-Plumed. 2025. Feedback on the Second Draft of the General-Purpose AI Code of Practice: Comments and Recommendations. Leverhulme Centre for the Future of Intelligence, University of Cambridge. https://www.lcfi.ac.uk/news-events/blog/post/feedback-on-the-general-purpose-ai-code-of-practice

[61] Jake M Hofman, Angelos Chatzimparmpas, Amit Sharma, Duncan J Watts, and Jessica Hullman. 2023. Pre-registration for predictive modeling. *arXiv preprint arXiv:2311.18807* (2023).

[62] Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James Glass, Akash Srivastava, and Pulkit Agrawal. 2024. Curiosity-driven red-teaming for large language models. *arXiv preprint arXiv:2402.19464* (2024).

[63] Katarína Hrabovská, Bruno Rossi, and Tomáš Pitner. 2019. Software testing process models benefits & drawbacks: a systematic literature review. *arXiv preprint arXiv:1901.01450* (2019).

[64] Jennifer Hu and Michael C Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. *arXiv preprint arXiv:2404.02418* (2024).

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

[66] Taisuke Imai, Séverine Toussaert, Aurélien Baillon, Anna Dreber, Seda Ertaç, Magnus Johannesson, Levent Neyse, and Marie Claire Villeval. 2025. *Pre-registration and pre-analysis plans in experimental economics.* Technical Report. JSTOR.

[67] John Ioannidis, Evangelia E Ntzani, Thomas A Trikalinos, and Despina G Contopoulos-Ioannidis. 2001. Replication validity of genetic association studies. *Nature genetics* 29, 3 (2001), 306–309.

[68] John PA Ioannidis. 2005. Why most published research findings are false. *PLoS medicine* 2, 8 (2005), e124.

[69] Anna A Ivanova. 2023. Running cognitive evaluations on large language models: The do's and the don'ts. *arXiv preprint arXiv:2312.01276* (2023).

[70] Anna A Ivanova. 2023. Toward best research practices in ai psychology. *arXiv e-prints* (2023), arXiv–2312.

[71] Sarah Jabbour, Trenton Chang, Anindya Das Antar, Joseph Peper, Insu Jang, Jiachen Liu, Jae-Won Chung, Shiqi He, Michael Wellman, Bryan Goodman, et al. 2025. Evaluation Framework for AI Systems in" the Wild". *arXiv preprint arXiv:2504.16778* (2025).

[72] Muhammad Abid Jamil, Muhammad Arif, Normi Sham Awang Abubakar, and Akhlaq Ahmad. 2016. Software testing techniques: A literature review. In *2016 6th international conference on information and communication technology for the Muslim world (ICT4M)*. IEEE, 177–182.

[73] Sayash Kapoor, Emily M Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A Bail, Odd Erik Gundersen, Jake M Hofman, Jessica Hullman, Michael A Lones, Momin M Malik, et al. 2024. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances* 10, 18 (2024), eadk3452.

[74] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*. 4110–4124.

[75] Simon Knight, Cormac McGrath, Olga Viberg, and Teresa Cerratto Pargman. 2025. Learning about AI ethics from cases: a scoping review of AI incident repositories and cases. *AI and Ethics* (2025), 1–17.

[76] John P Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. 2024. Item response theory for natural language processing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. 9–13.

[77] Richard N Landers and Tara S Behrend. 2023. Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist* 78, 1 (2023), 36.

[78] Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, et al. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. *arXiv preprint arXiv:2407.04069* (2024).

[79] Seanie Lee, Minsu Kim, Lynn Cherif, David Dobre, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, et al. 2024. Learning diverse attacks on large language models for robust red-teaming and safety tuning. *arXiv preprint arXiv:2405.18540* (2024).

[80] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939* (2024).

[81] Yu Lu Liu, Su Lin Blodgett, Jackie Chi Kit Cheung, Q Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024. ECBD: Evidence-centered benchmark design for NLP. *arXiv preprint arXiv:2406.08723* (2024).

[82] Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, et al. 2024. A safe harbor for ai evaluation and red teaming. *arXiv preprint arXiv:2403.04893* (2024).

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

[84] Antoine Marot, David Rousseau, et al. 2023. Towards impactful challenges: post-challenge paper, benchmarks and other dissemination actions. *arXiv preprint arXiv:2312.06036* (2023).

[85] Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cesar Ferri, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, María José Ramírez-Quintana, and Peter Flach. 2019. CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering* 33, 8 (2019), 3048–3061.

[86] Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2019. Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial intelligence* 271 (2019), 18–42.

[87] Tegan McCaslin, Jide Alaga, Samira Nedungadi, Seth Donoughe, Tom Reed, Rishi Bommasani, Chris Painter, and Luca Righetti. 2025. STREAM (ChemBio): A Standard for Transparently Reporting Evaluations in AI Model Reports. *arXiv preprint arXiv:2508.09853* (2025).

[88] Daniel McDuff, David Munday, Xin Liu, and Isaac Galatzer-Levy. 2024. Cognitive Assessment of Language Models. In *ICML Workshop on LLMs and Cognition*.

[89] Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Dan Xu, Paul Watters, and Malka N Halgamuge. 2025. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence* (2025).

[90] Ninareh Mehrabi, Palash Goyal, Christophe Dupuy, Qian Hu, Shalini Ghosh, Richard Zemel, Kai-Wei Chang, Aram Galstyan, and Rahul Gupta. 2024. Flirt: Feedback loop in-context red teaming. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 703–718.

[91] Hans Melander, Jane Ahlqvist-Rastad, Gertie Meijer, and Björn Beermann. 2003. Evidence b (i) ased medicine—selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *Bmj* 326, 7400 (2003), 1171–1173.

[92] Microsoft. 2025. Learning from Other Domains to Advance AI Evaluation and Testing. https://www.microsoft.com/en-us/research/wp-content/uploads/2025/08/Learning-from-other-Domains-to-Advance-AI-Evaluation-and-Testing_-v3-1.pdf

[93] Edward Miguel, Colin Camerer, Katherine Casey, Joshua Cohen, Kevin M Esterling, Alan Gerber, Rachel Glennerster, Don P Green, Macartan Humphreys, Guido Imbens, et al. 2014. Promoting transparency in social science research. *Science* 343, 6166 (2014), 30–31.

[94] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[95] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. AI and Ethics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1. 4275–4293.

[96] Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Nebojsa Jojic, Hamid Palangi, Robert Ness, and Jonathan Larson. 2023. Evaluating cognitive maps and planning in large language models with cogeval. *Advances in Neural Information Processing Systems* 36 (2023), 69736–69751.

[97] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. 2017. A manifesto for reproducible science. *Nature human behaviour* 1, 1 (2017), 0021.

[98] National Institute of Standards and Technology. 2025. Evaluation of DeepSeek AI Models. U.S. Department of Commerce.

[99] National Technical Committee 260 on Cybersecurity of the Standardization Administration of China. 2024. Technical Documentation of National Technical Committee 260 on Cybersecurity of Standardization Administration of China: Basic Safety Requirements for Generative Artificial Intelligence Services. https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai-final/ English translation by CSET, originally issued by the Standardization Administration of China.

[100] Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2600–2606.

[101] Brian A Nosek and Daniël Lakens. 2014. Registered reports.

[102] OpenAI. 2025. GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf

[103] Patricia Paskov, Lukas Berglund, Everett Smith, and Lisa Soder. 2024. GPAI Evaluations Standards Taskforce: Towards Effective AI Governance. *arXiv preprint arXiv:2411.13808* (2024).

[104] Patricia Paskov, Michael J Byun, Kevin Wei, and Toby Webster. 2025. Preliminary suggestions for rigorous GPAI model evaluations. *arXiv preprint arXiv:2508.00875* (2025).

[105] Cristiano Patrício, Rui Pinto, and Gonçalo Marques. 2020. A study on software testing standard using iso/iec/ieee 29119-2: 2013. In *Recent Advances in Intelligent Systems and Smart Applications*. Springer, 43–62.

[106] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* (2022).

[107] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research* 22, 164 (2021), 1–20.

[108] Matteo Prandi, Vincenzo Suriani, Federico Pierucci, Marcello Galisai, Daniele Nardi, and Piercosma Bisconti. 2025. Bench-2-CoP: Can We Trust Benchmarking for EU AI Compliance? *arXiv preprint arXiv:2508.05464* (2025).

[109] Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016* (2024).

[110] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[111] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[112] Sunayana Rane, Cyrus F Kirkman, Graham Todd, Amanda Royka, Ryan MC Law, Erica Cartmill, and Jacob Gates Foster. [n. d.]. Position: Principles of Animal Cognition to Improve LLM Evaluations. In *Forty-second International Conference on Machine Learning Position Paper Track*.

[113] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, et al. 2024. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981* (2024).

[114] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. 2024. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems* 37 (2024), 21763–21813.

[115] Anka Reuel, Lisa Soder, Benjamin Bucknall, and Trond Arne Undheim. 2024. Position: Technical research and talent is needed for effective AI governance. In *Forty-first International Conference on Machine Learning*.

[116] Zarreen Reza. 2025. The social laboratory: A psychometric framework for multi-agent llm evaluation. *arXiv preprint arXiv:2510.01295* (2025).

[117] Marco Tulio Ribeiro and Scott Lundberg. 2022. Adaptive testing and debugging of NLP models. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*. 3253–3267.

[118] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* (2020).

[119] Anne M Scheel, Mitchell RMJ Schijen, and Daniël Lakens. 2021. An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science* 4, 2 (2021), 25152459211007467.

[120] Kenneth F Schulz, Iain Chalmers, Richard J Hayes, and Douglas G Altman. 1995. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Jama* 273, 5 (1995), 408–412.

[121] Ken Schwaber and Mike Beedle. 2001. *Agile software development with Scrum*. Prentice Hall PTR.

[122] Seungmin Seo, Hariharan Iyer, and Yooyoung Lee. 2025. 2025 NIST GenAI Text Challenge Evaluation Plan. (2025).

[123] Mary Shaw. 2002. Prospects for an engineering discipline of software. *IEEE Software* 7, 6 (2002), 15–24.

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

[125] Jingzhe Shi, Jialuo Li, Qinwei Ma, Zaiwen Yang, Huan Ma, and Lei Li. 2024. Chops: Chat with customer profile systems for customer service with llms. *arXiv preprint arXiv:2404.01343* (2024).

[126] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22, 11 (2011), 1359–1366.

[127] Courtney K Soderberg, Timothy M Errington, Sarah R Schiavone, Julia Bottesini, Felix Singleton Thorn, Simine Vazire, Kevin M Esterling, and Brian A Nosek. 2021. Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour* 5, 8 (2021), 990–997.

[128] Leon Staufer, Mick Yang, Anka Reuel, and Stephen Casper. 2025. Audit Cards: Contextualizing AI Evaluations. *arXiv preprint arXiv:2504.13839* (2025).

[129] Merlin Stein, Milan Gandhi, Theresa Kriecherbauer, Amin Oueslati, and Robert Trager. 2024. Public vs Private Bodies: Who Should Run Advanced AI Evaluations and Audits? A Three-Step Logic Based on Case Studies of High-Risk Industries. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 1401–1415.

[130] Conrad Stosz, Karson Elmgren, Charles Foster, George Balston, Seth Donoughe, Samira Nedungadi, Michael Chen, Jasper Götting, Patricia Paskov, Sayash Kapoor, Sarah Schwettmann, Rishi Bommasani, Luca Righetti, Sean McGregor, Grace Werner, Rob Reich, Arvind Narayanan, Elizabeth Barnes, Christopher Painter, Miles Brundage, Aidan Homewood, Divya Siddharth, Faisal Lalani, Charles Teague, Jaime Sevilla, and Jacob Steinhardt. 2025. AEF-1: Minimum Operating Conditions for Independent Third Party AI Evaluations. Technical report / framework specification.

[131] Christopher Summerfield, Lennart Luettgau, Magda Dubois, Hannah Rose Kirk, Kobi Hackenburg, Catherine Fist, Katarina Slama, Nicola Ding, Rebecca Anselmetti, Andrew Strait, et al. 2025. Lessons from a chimp: Ai" scheming" and the quest for ape language. *arXiv preprint arXiv:2507.03409* (2025).

[132] The White House. 2025. Winning the Race: America's AI Action Plan. U.S. Government policy document. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

[133] Songül Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernández-Macías, José Hernández-Orallo, and Emilia Gómez. 2021. Measuring the occupational impact of AI: tasks, cognitive abilities and AI benchmarks. *Journal of Artificial Intelligence Research* 71 (2021), 191–236.

[134] UK AI Security Institute. 2024. Early Lessons from Evaluating Frontier AI Systems. Blog post. https://www.aisi.gov.uk/blog/early-lessons-from-evaluating-frontier-ai-systems UK AISI blog.

[135] Gaël Varoquaux and Veronika Cheplygina. 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine* 5, 1 (2022), 48.

[136] Verify Foundation. 2023. Cataloguing LLM Evaluations. https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf Draft for discussion.

[137] Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, et al. 2025. Position: Evaluating generative ai systems is a social science measurement challenge. *arXiv preprint arXiv:2502.00561* (2025).

[138] Xiting Wang, Liming Jiang, Jose Hernandez-Orallo, David Stillwell, Luning Sun, Fang Luo, and Xing Xie. 2023. Evaluating general-purpose AI with psychometrics. *arXiv preprint arXiv:2310.16379* (2023).

[139] Yuqing Wang, Maaret Pyhäjärvi, and Mika V Mäntylä. 2020. Test automation process improvement in a DevOps team: experience report. In *2020 ieee international conference on software testing, verification and validation workshops (icstw)*. IEEE, 314–321.

[140] Kevin Wei, Stephen Guth, Gabriel Wu, and Patricia Paskov. 2025. Methodological Challenges in Agentic Evaluations of AI Systems. In *ICML Workshop on Technical AI Governance (TAIG)*.

[141] Kevin L. Wei, Patricia Paskov, Sunishchal Dev, Michael J. Byun, Anka Reuel, Xavier Roberts-Gaal, Rachel Calcott, Evie Coxon, and Chinmay Deshpande. 2025. Recommendations and Reporting Checklist for Rigorous & Transparent Human Baselines in Model Evaluations. arXiv:2506.13776 [cs.AI] https://arxiv.org/abs/2506.13776

[142] Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336* (2025).

[143] P Williamson, JL Hutton, J Bliss, J Blunt, MJ Campbell, and R Nicholson. 2000. Statistical review by research ethics committees. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163, 1 (2000), 5–13.

[144] Richard Wiseman, Caroline Watt, and Diana Kornbrot. 2019. Registered reports: an early example and analysis. *PeerJ* 7 (2019), e6232.
[145] Yuki Yamada. 2018. How to crack pre-registration: Toward transparent and open science. *Frontiers in psychology* 9 (2018), 1831.
[146] Blair Yang, Fuyang Cui, Keiran Paster, Jimmy Ba, Pashootan Vaezipoor, Silviu Pitis, and Michael R Zhang. 2024. Report Cards: Qualitative Evaluation of LLMs Using Natural Language Summaries. In *Workshop on Socially Responsible Language Modelling Research*.
[147] Liang Yu, Emil Alégroth, Panagiota Chatzipetrou, and Tony Gorschek. 2025. Measuring the quality of generative AI systems: Mapping metrics to quality characteristics-Snowballing literature review. *Information and Software Technology* (2025), 107802.
[148] Andy K Zhang, Kevin Klyman, Yifan Mai, Yoav Levine, Yian Zhang, Rishi Bommasani, and Percy Liang. 2024. Language model developers should report train-test overlap. *arXiv preprint arXiv:2410.08385* (2024).
[149] Lili Zhang, Haomiaomiao Wang, Long Cheng, Libao Deng, and Tomas Ward. 2025. Adversarial Testing in LLMs: Insights into Decision-Making Vulnerabilities. *arXiv preprint arXiv:2505.13195* (2025).
[150] Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Llmeval: A preliminary study on how to evaluate large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19615–19622.
[151] Dora Zhao, Qianou Ma, Xinran Zhao, Chenglei Si, Chenyang Yang, Ryan Louie, Ehud Reiter, Diyi Yang, and Tongshuang Wu. 2025. SPHERE: An Evaluation Card for Human-AI Systems. In *Findings of the Association for Computational Linguistics: ACL 2025*. 1340–1365.
[152] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470* (2024).
[153] Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E Prunty, et al. 2025. General scales unlock ai evaluation with explanatory and predictive power. *arXiv preprint arXiv:2503.06378* (2025).
[154] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature* 634, 8032 (2024), 61–68.
[155] Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, et al. 2025. Establishing best practices for building rigorous agentic benchmarks. *arXiv preprint arXiv:2507.02825* (2025).

A Appendix: The Protocol

A.1 Goals and Objectives

Goals and Objectives Overview. The first phase involves clearly defining the goals, objectives, and underlying rationale for conducting the evaluation. This includes specifying what the evaluation seeks to achieve—such as verifying safety claims, comparing model capabilities, identifying failure modes, or informing deployment decisions. These different goals may necessitate different kinds of evaluation methodologies, so it is important to clearly specify the objectives up front. If an evaluation has multiple goals or objectives, it is vital that the chosen methodologies are appropriate for achieving each evaluation goal. Objectives should be specific, measurable, and grounded in the operational context of the evaluation target. Additionally, this phase includes identifying key terminology, available resources, risks, and constraints. Although this step may appear optional, it's not: clarifying the reasons for your evaluation effort ensures that everyone is on the same page before resources are committed.

A.1.1 Determine Project Purpose.

Compiling Relevant Background. Given that stakeholders involved in conducting an evaluation may come from diverse sectors—including industry, academia, third-party evaluators, users, and public bodies—the relevant background to the evaluation project should be described. Understanding this background helps you know what you're working with in terms of stakeholders' goals, priorities, and interests, as well as the deployment context of the AI system (e.g., sector, risk level, intended use).

The deployment context is a key factor in shaping how evaluation results are interpreted and what implications they carry. For instance, the same evaluation objective—such as assessing a system's classification capability—may carry different significance depending on whether the system is deployed in a critical domain (e.g., healthcare, criminal justice) or a non-critical setting (e.g., entertainment recommendations).

Task List 1.1.1 - Consider the Different Stakeholders

- Reflect on the institutional context, goals, and priorities of the team conducting the evaluation, and how these might shape the scope, focus, and interpretation of the results.
- Map the key stakeholders/target audience involved in or affected by the evaluation, outlining their roles, interests, and potential concerns.

Task List 1.1.2 - Describe the Deployment Context

- Determine whether the AI system is sector-specific (e.g., healthcare, finance, education) or general-purpose, and clearly identify the relevant deployment context.
- Justify the relevance and urgency of the evaluation based on the deployment setting.
- Assess sector-specific risks, sensitivities, and regulatory frameworks. Determine if the sector involves high-stakes or safety-critical applications, even when the AI system is used in accordance with its intended purpose. Capture any domain-relevant ethical, legal, or societal considerations.

Terminology. As AI evaluations have become increasingly complex and multidisciplinary, spanning various paradigms, it is important to explain the terminology used. This step is crucial to ensure that all team members are "speaking the same language," thereby contributing to clarity and transparency. It also helps establish a shared understanding among stakeholders and readers, particularly when terms may carry different meanings across disciplines or application domains. For instance, recent terms like "AI evals" or "dangerous capabilities" carry specific meanings in safety-focused communities that may differ from their use in other scientific contexts related to the measurement of natural or artificial cognition or on regulatory contexts, making clarification important.

Task List 1.1.3 - Fix the Terminology

- Keep a list of terms or jargon that may be confusing to team members. Include those that may be discipline-specific or ambiguous.
- Consult team members from different backgrounds to validate definitions. Include examples or context where useful to clarify meaning.
- Share the glossary with all team members and relevant stakeholders. Consider submitting it as part of the project documentation at the pre-registration stage.

Defining Project Goals. A clear and concrete primary objective should be defined and agreed upon by the evaluation team and relevant stakeholders. There are many different reasons why one might conduct an evaluation, including quantifying system capabilities, tracking progress over time, enabling large-scale comparisons, understanding behavioural patterns, identifying and estimating potential risks, providing assurance of system safety, predicting if future models can lead to catastrophic harm or evaluating evaluation methods, among others. As mentioned earlier, the chosen objective will directly influence the selection of appropriate evaluation methodologies.

Task List 1.1.4 - Define Project Goals

- Identify the primary purpose(s) of the evaluation (e.g., capability measurement, risk assessment, safety assurance).
- Describe the uncertainty you want to resolve or the questions you want to answer by conducting the evaluation.
- List all specific goals and objectives the evaluation aims to achieve as precisely as possible, that would resolve the questions above.

Project Success Criteria. What would the evaluation look like if it were successful? Clearly specifying the success criteria provides a benchmark for assessing the evaluation's effectiveness and whether it has achieved its intended goals. This might include factors such as the ability to identify meaningful insights, the reliability of the evaluation process, and the usefulness of the findings for system improvement or decision-making.

Task List 1.1.5 - Establish Project Success Criteria

- Identify specific success indicators as precisely as possible (e.g., insight generation, reliability, applicability).
- Consult stakeholders to align success criteria with their expectations and needs.

A.1.2 Determine Technical Objectives.

Evaluation Target and Justification. This step involves identifying and justifying the evaluation targets. In most cases, the target will be an AI system, requiring a clear definition of the specific capabilities, behaviors, or properties to be measured or estimated. In the case of meta-evaluations, however, the evaluation target may be the evaluation method itself (e.g., a benchmark). In such cases, the justification may involve the proposal of a new method intended to address the limitations of existing approaches—necessitating its evaluation and comparison—or the assessment of the validity and reliability of alternative techniques.

Task List 1.2.1 - Determine and Justify Evaluation Target

- Clearly specify whether the target is an AI system or an evaluation method. If the target is an AI system, include identifying details when available—such as the system's name and version, underlying architecture, parameter count, and whether it is a base model or has been fine-tuned. If it is a method, a description of the method, if it is a benchmark, then a full description with its content and procedure.
- Describe the target's scope. Identify what part or specific situations or behaviours to evaluate, if not the whole target (e.g., only text in a multimodal system).
- Justify the selection of the target. Explain the relevance and importance of evaluating this target in the current context, and how this is necessary and sufficient for the project goals and their project success criteria.

Evaluation Success Criteria. Success must also be defined in technical terms to keep the evaluation efforts on track. Use the project goal and success criteria determined earlier to formulate metrics and criteria for success. This involves

translating the broader project goals and high-level success criteria into specific metrics, statistical thresholds, and operational definitions that guide the analysis.

Task List 1.2.2 - Establish and Justify Success Criteria

- Select and define the evaluation criteria(s) and metric(s) for the evaluation target.
- Describe the success criteria of the evaluation in terms of uncertainty of the estimators, predictability and other indicators of methodological robustness.
- Justify the success criteria based on the goals and background of the evaluation as a reference against which to assess whether the evaluation process has met its intended standards. For instance, acceptable levels of confidence interval width, statistical significance thresholds, or variability across repeated trials may be set to define what will be considered a reliable or meaningful result.

For example, consider Appendix H, Face Recognition System Comparison. In this case, the project success criteria are defined as the ability of the evaluation to determine, across all relevant dimensions, whether the new technique performs significantly better, significantly worse, or shows no difference compared to the baseline. The technical objectives, in turn, specify these dimensions more precisely—specificity, sensitivity, accuracy, balanced accuracy, and F-score—as well as the protected groups to be evaluated, including race, gender, and age. Additionally, they delineate the varying operational conditions under which both System A and System B are assessed, such as differences in lighting, subject motion, and other environmental factors.

A.1.3 Situation Assessment.

Resource Inventory. Conducting and running an evaluation can be resource-intensive. It is therefore important to develop an inventory of resources, accounting for computational requirements, data volume, personnel responsible for conducting the evaluation, and any associated financial costs. When necessary, these considerations allow for contemplating options for reducing costs, such as through test prioritisation—where inputs generated for tests are limited to those that denote the more effective adversarial examples.

Task List 1.3.1 - Secure Evaluation Target Access

- Explore and ensure reliable and timely access to the evaluation target, including any necessary APIs, interfaces or documentation. Determine and secure any legal agreement or conditions of use.

Task List 1.3.2 - Secure Hardware and Software Resources

- Explore and ensure the necessary hardware resources are available to support the evaluation workload.
- Explore and ensure the necessary software resources (licences, APIs, etc.) are available to support the evaluation workload.

Task List 1.3.3 - Establish Data and User Feedback

- Ensure access to a dataset suitable for this evaluation, or determine whether it will be constructed.
- Determine if available or built datasets are sufficiently large and diverse to support robust and meaningful evaluation results.
- Determine and secure the resources and agreements to collect data or feedback on the use or interactions of the AI system, if needed.

Task List 1.3.4- Evaluation Team

- Ensure there is access to personnel with adequate expertise and experience in evaluation methodologies. Determine the role of AI conducting parts of the evaluation (LLMs as a judge, reporting, etc.).
- Specify how many people are involved and what roles or areas of expertise they represent.

Requirements, Assumptions, and Constraints. Taking the time to honestly assess requirements, assumptions and constraints of the project increases the likelihood of a successful evaluation. By making these aspects as explicit as possible early on, teams can proactively avoid costly setbacks later in the process.

Task List 1.3.5- Specify Project Requirements

- List any necessary conditions for conducting the evaluation (e.g., compliance standards, reporting formats, required approvals).

Task List 1.3.6- Determine Known Constraints

- Enumerate and quantify all financial constraints covered in the project budget.
- Record limitations such as budget, deadlines, staffing, computational capacity, legal constraints, ethical constraints or data availability.

Task List 1.3.7 - Establish Key Assumptions and Biases

- Clearly state any assumptions about the system, evaluation environment, etc.
- Determine if the expectations about outcomes may influence or bias design decisions.

Task List 1.3.8 - Ensure Alignment with Stakeholder Expectations

- Ensure that requirements and constraints are understood and accepted by all relevant stakeholders.
- Ensure everyone is aligned on the project scheduling requirements.

Task List 1.3.9 - Plan for Flexibility

- Assess which constraints are fixed and which could be adapted if needed.
- Determine trade-offs (e.g., between evaluation depth and timeline) and how they relate to success criteria.

Risks and Contingencies. Another important step is to identify potential risks and contingencies that may arise during the evaluation process. By anticipating these risks, evaluators can implement proactive measures to mitigate them, ensuring that the evaluation remains on track and that any potential obstacles are addressed in a timely manner. Furthermore, identifying contingencies helps ensure that alternative strategies are available if the initial approach encounters significant setbacks. Typical risks to consider include—but are not limited to— scheduling (What if the evaluation takes longer than expected?), financial (What if funding is reduced or withdrawn during the project?), data (What if the data are incomplete, low quality, or insufficient in scope?) and results (What if the initial findings are inconclusive, less impactful than anticipated, or difficult to interpret?)

After considering the various risks, come up with a contingency plan to help avert failure or even disaster.

Task List 1.3.10 - Document Risks and Contingencies

- Document each possible risk.
- Document a contingency plan for each risk.

Current Understanding of Target(s) to be Evaluated. Assessing the current state of understanding of the evaluation target(s)—and, if necessary, familiarizing oneself with their characteristics—is a critical step in designing effective evaluations. A clear understanding of the system's architecture, training data, intended use cases, known limitations, and prior performance helps ensure that the evaluation is appropriately tailored to the system. Similarly if we're talking about humans or benchmarks. This also enables evaluators to identify meaningful knowledge gaps, refine methods, and ensure that the results will be informative and relevant to stakeholders. Moreover, contextualizing the evaluation within existing literature or prior assessments allows for better interpretation of results and some justification of methodological choices.

Task List 1.3.11 - Understand Targets to be Evaluated

- Collect available documentation on the target's architecture, training data, and development process. For example, for benchmarks, data provenance, quality metrics, ground truth, annotators, etc.
- Analyze any known strengths, weaknesses, failure modes, or constraints from prior testing or usage.
- Review prior assessments or related literature to contextualize your evaluation and identify gaps.
- Pinpoint areas where information is lacking or uncertain, and determine whether further inquiry is needed before proceeding.
- Summarize the current understanding and any assumptions made to inform methodological decisions.

A.2 Evaluation Design

Evaluation Design Overview. Once the project's goals and objectives have been defined, the next step is to design the evaluation. For example, evaluations could include a combination of standardised benchmarks or tests, red teaming, and/or human rating. Selecting the most appropriate approaches requires careful consideration of the evaluation targets, their characteristics, and the tools and protocols available. Although certain elements of the evaluation may be refined iteratively as new insights or data emerge, establishing a well-justified evaluation design beforehand ensures alignment with the overarching objectives, and provides a coherent framework within which exploratory analyses or

adaptive modifications can be meaningfully situated. In this way, a thoughtfully designed evaluation supports both methodological rigor and adaptive flexibility.

A.2.1 Identify Potential Evaluation Methods.

Identify Tools that Could Be Used for Each of the Targets. Just as it is crucial to understand the target being evaluated, it is equally important to have a comprehensive overview of the evaluation methods currently available in the state of the art. The next step, therefore, is to identify the tools, protocols, and resources that can be used to assess the specified targets in 1.2. This step is essential even when proposing a new evaluation method, as it helps to reveal the limitations of existing approaches and to provide a well-founded justification for the proposed alternative.

Task List 2.1.1 - Identify Evaluation Tools

- Familiarize yourself with current evaluation methods, tools and benchmarks used in similar contexts. Review recent academic literature, technical reports, and leaderboards.
- For each target defined in 1.2. list corresponding evaluation tools or protocols, noting their core methodology and scope.
- Assess method maturity and adoption. Note whether each tool is widely used, experimental, or still under development.
- Identify the origin of the method (e.g., academic, open-source, industry) and current level of community adoption.
- Summarize key strengths and known limitations for each method. Highlight any gaps for which no adequate method currently exists.
- Monitor emerging work. Given the rapid evolution of AI evaluation, regularly review new tools, methods, and protocols to ensure the approach remains up to date and aligned with best practices.

Also, several aspects should be reviewed depending on the evaluation method under consideration. If using benchmarks, this may involve surveying available public or internal benchmark datasets, checking their documentation and design choices to assess their reproducibility. In the case of red-teaming, identify potential processes for generating adversarial inputs—whether through automated methods, manual efforts, or hybrid approaches—and consider the use of diverse classifiers to capture different failure modes. For human evaluations, attention should be paid to the methods used for collecting responses, including the selection of platforms (e.g., crowdsourcing services) and the criteria for recruiting and screening raters to ensure reliability, diversity, and alignment with the task requirements.

A.2.2 Selection of Evaluation Methods.

Choosing and Justifying the Evaluation Approach. Based on the information gathered in the previous step, a decision should be made regarding the most suitable evaluation approach. If none of the existing methods are adequate, this is the appropriate stage to design or develop a new method.

Task List 2.2.1 - Selection of an Existing Evaluation Approach

- Evaluate the coverage, relevance, and limitations of each tool or benchmark in relation to the specific evaluation target(s) and objectives identified in phase 1.2.
- Assess feasibility. Consider the practical constraints, such as available resources, time, required expertise, and technical infrastructure identified in phase 1.3, for implementing each method.
- Describe the justification for the selection of these method(s), avoiding superficial rationales such as relying solely on popularity or precedent.

For example, in the case of benchmarks, the analysis should include considerations of the benchmark dataset's reliability and validity, both internal and external, particularly in relation to the deployment context of the system(s) under evaluation, as well as potential issues such as data contamination. For red teaming, the analysis may assess the recruitment criteria for red teaming practitioners or the diversity and effectiveness of adversarial attacks, as illustrated in Appendix F, Curiosity-driven Red-teaming for Large Language Models. In the case of human ratings, the examination could address the use of attention checks to ensure high-quality annotations, as well as methods for estimating inter-annotator agreement and minimizing bias in ratings.

Task List 2.2.2 - Design a New Method (if applicable)

- Identify gaps. If none of the existing tools or methods suffice, clearly articulate what is missing and why current options fall short.
- Outline new method objectives and assumptions. Define the specific properties or capabilities the new method should evaluate and the assumptions underlying its design.
- Propose preliminary design and validation strategy. Describe a plan for testing the method's effectiveness, including pilot evaluations, comparisons with baselines, or theoretical justification.

A.2.3 Analytics Specification.

Analytics Planning and Justification. This step helps ensure that the evaluation yields interpretable, reliable, and meaningful results. Explicit plans for using inferential statistics and uncertainty quantification also strengthen the validity of the conclusions by clarifying the level of confidence that can be placed in the findings. Overall, a well-defined analytics plan supports transparency, reproducibility, and robustness.

Task List 2.3.1 - Specify and Justify the Analytics

- Decide the analytics approach for the evaluation data, including which estimators and metrics will be produced.
- Specify the types of analyses to be conducted (e.g., summary statistics, error analysis, predictive modeling, visualisations, etc).
- Provide a justification for the chosen analytics approach.
- Determine how inferential statistics and uncertainty quantification will be used to support conclusions and assess robustness.

A.3 Project Plan

Project Plan Overview. At this point, you are ready to develop a comprehensive plan for the evaluation project that can be pre-registered. The research questions posed, along with the defined objectives serve as the foundation for this roadmap. This plan will guide the design, execution, and analysis phases, ensuring alignment between goals, methods, and expected outcomes.

A.3.1 Create a Project Plan.

Writing your Project Plan. Documenting the evaluation project plan in writing is essential for ensuring clarity, coordination, and accountability across the evaluation team. A written plan serves as a shared reference point that helps align expectations, reduce misunderstandings, and track progress over time. It also enables transparency in decision-making, facilitates external review or auditing if needed, and supports reproducibility by preserving the rationale behind key choices. Specifically, the project plan may include the major stages of the evaluation process, a realistic timeline, including key milestones and deadlines for each stage, the resources required to carry out the evaluation identified in phase 1.3, the expected outputs and deliverables of the project, including reports or datasets, and any other relevant information gathered during the previous planning phases.

Task List 3.1.1:

- Draft an initial outline of the evaluation project plan.
- Distribute the draft to all team members for review and input.
- Gather and integrate feedback into a consolidated version.
- Share the updated version for final review or approval.
- Save the finalized plan jointly with sections 1 and 2 in a shared drive or version control system to ensure accessibility and future reference.

A.3.2 Pre-register Evaluation.

Pre-registration Submission. Pre-registering the evaluation project plan is crucial for enhancing the transparency and credibility of the evaluation process. This step helps to prevent selective reporting and establishes a clear commitment to a predefined approach, allowing both internal and external stakeholders to hold the evaluation accountable. Moreover, pre-registration facilitates reproducibility and enables constructive feedback early in the project.

Task List 3.2.1 - Pre-register

- Select an appropriate time-stamped repository or version control system for submission to receive potential external or internal feedback.
- Submit a “pre-registration” of the protocol with sections 1, 2 and 3.1, eliminating details that are sensitive or irrelevant for the registration (team members, exact timelines).
- Consider supplementing the pre-registration with a comprehensive final report, released alongside project outputs at the conclusion of phase 6.

A.4 Data Collection

Data Collection Overview. By data collection, we refer to the full set of processes required to generate or obtain the experimental data necessary for subsequent analysis. This includes preparatory steps such as establishing the experimental setup, dataset selection, and annotation, as well as the acquisition and preprocessing of existing evaluation data. Once these elements are in place, experiments with the evaluation target(s) and analysis can be conducted.

A.4.1 Experimental Setup, Annotations and Pilots.

Experimental Setup and Annotations. This step focuses on the practical setup required to conduct the evaluation as designed in stage 2, which usually involves some combination of tests (benchmarks, interviews, protocols), AI systems and possibly humans. This includes acquiring relevant data from the evaluation of the target and preparing the evaluation environment. If existing datasets are used, it is important to ensure they are accessible. If additional elements—such as annotations, filtering criteria, or preprocessing steps—are required, these should be developed and standardized. The relevance and applicability of the following tasks within this stage may vary depending on the chosen data collection method.

Task List 4.1.1 - Collect Data

- If existing datasets are to be used for the evaluation, ensure they are accessible. Resolve any licensing, hosting, or formatting issues in advance. If humans are required for the evaluation, ensure they are informed and ethical approvals have been conducted. If AI systems are required (as targets, secondary elements or evaluators).
- If a new dataset is being created specifically for the evaluation, carry out the necessary steps to generate and format the input data. This includes collecting or synthesizing data and applying any required filtering or preprocessing procedures.
- If the target behaviour to be generated dynamically—such as in iterative, trial-and-error procedures used in red teaming—ensure that practitioners have clear guidelines for interaction. Provide detailed instructions to standardize the process and maintain consistency across trials.
- Identify and collect any additional data needed (e.g., ground truth labels, reference answers, metadata).
- Verify data quality and integrity.

Task List 4.1.2 - Run Experimental Samples

- Generate or collect behavioural samples (outputs, conversations, etc.) with the target(s) to be evaluated.
- Ensure sampling strategy is consistent with the evaluation design (e.g., random, stratified, adversarial).

Task List 4.1.3 - Determine Annotation Setup (if applicable)

- Define annotation tasks and criteria. Draft annotation guidelines or rubrics.
- Recruit qualified annotators or crowdworkers, humans or automated.
- Define rubrics, conduct training sessions or provide instructional material.
- Select or build a tool for annotation if humans or a pipeline if automated. Test for usability and clarity.
- Execute the annotation process following the established guidelines and monitor its quality.

Task List 4.1.4 - Develop Filter or Classifiers (if applicable)

- Develop filters or classifiers to pre-screen content.
- Train and validate these tools with appropriate data.

A.4.2 Pilot Tests (Optional). Pilot tests allow the evaluation team to identify potential flaws or unforeseen challenges in the evaluation design before full-scale implementation. This looks at the project goals and the criteria for success, beyond the specific metrics. This is especially important when proposing novel methods, as piloting helps ensure that they meet the intentions of the project but also that the procedures are feasible, the instructions are clear, and the tools function as intended. Moreover, pilot tests can provide preliminary insights into the data quality and the evaluation target behavior, enabling adjustments that improve the robustness of the main process.

Task List 4.1.5 - Conduct Pilots

- Define the scope of the pilot. Select a small, representative subset of tasks, AI systems/models, or participants to test.
- Prepare materials and tools. Ensure all instructions, datasets, interfaces, and evaluation tools are ready and functional.
- Recruit pilot participants (if applicable). Select evaluators or human raters representative of the main study's population.
- Conduct the pilot test. Run the pilot in conditions that mirror the main evaluation as closely as possible.
- Collect feedback and observations. Gather both quantitative data and qualitative input from participants or observers.
- Identify issues and challenges. Analyze the results to detect procedural flaws, ambiguities, or technical problems.
- Adjust the protocol as needed. Revise instructions, tools, or evaluation design based on pilot findings.
- Document all changes. Keep a clear record of what was learned and how it informed the updated evaluation protocol.

A.4.3 Full data collection.

Obtaining the Data for the Analysis. This step involves executing the full-scale data collection process as planned in earlier stages. At this point, the experimental protocols or evaluation procedures designed in phase 2 are deployed to systematically generate or gather the data required for analysis. This includes evaluating the targets under controlled conditions to produce the evaluation data.

Task List 4.2.1 - Run Full Experiment and Obtain Full Data

- Run the system(s) or benchmarks against the selected input dataset(s), tests or systems.
- Verify that the data collected follows the pre-defined sampling strategy (e.g., balanced, random, representative).
- Record target behaviour systematically, capturing all relevant metadata (e.g., version, configurations, time, error messages).

A.4.4 Data preparation.

Preparing the Data for the Analysis . This step focuses on preparing the collected evaluation data for analysis. It includes cleaning and formatting the raw data to ensure consistency across all tasks and sources. Data quality must be systematically verified, including checks for missing values, formatting errors, and outliers. Outputs should also be reviewed to ensure they are sensible within the context of each benchmark. When human ratings or annotations are involved, this step entails analysing rater consistency, measuring inter-annotator agreement, and identifying potential biases or anomalies in the responses. Proper data preparation ensures that the analysis conducted in subsequent phases is based on a reliable and interpretable dataset.

Task List 4.3.1 - Prepare Data

- Clean raw evaluation data. Remove incomplete, duplicated, or corrupted records and standardize text formats (e.g., whitespace, punctuation, encoding) or generated multimedia data (e.g., contrast, brightness, etc.).
- Format data for the analysis. Organize model outputs, inputs, metadata, and annotations into structured formats (e.g., CSV, JSON, database). Label and align variables consistently across all tasks.
- Conduct quality control checks on model outputs. Flag incomplete, corrupted, or low-quality samples for review.
- Organize and store collected data. Store all collected data in an organized, version-controlled repository. Include documentation for data provenance, generation conditions, and structure.
- Ensure that the collected dataset is clean, complete, and ready for the analytical procedures defined in early stages.

A.5 Data Analysis

Data Analysis Overview. This phase involves the systematic analysis of the prepared evaluation data. It begins with initial exploratory steps to understand the structure, distribution, and patterns in the data. This informs the execution of the predefined analysis plan, including the computation of summary statistics, metrics, and inferential statistics. The phase also allows for adjustments to the analytical strategy in response to new findings. The associated analysis of quality, uncertainty of estimators and any other additional analyses on unexpected patterns is also conducted here.

A.5.1 Initial Data Exploration.

Preliminary Data Review. In this step, the goal is to develop a comprehensive understanding of the evaluation data before conducting the planned analyses. This includes visualizing data distributions, identifying trends, and exploring

system behaviour across different tasks or features. This phase may lead to modifications in the analysis plan or inclusion/exclusion criteria, all of which should be transparently documented.

Task List 5.1.1 - Conduct Preliminary Data Review

- Describe and summarize key variables and distributions in the evaluation dataset.
- Visualize performance, safety and other behavioural patterns using plots (e.g., histograms, scatter plots, confusion matrices).
- Explore variation across task features, such as prompt types, input lengths, or topics.
- Identify anomalies or unexpected patterns that may affect analysis validity.
- Revise the analysis plan if new insights suggest necessary changes, and document the rationale.
- Define and apply inclusion/exclusion criteria for data points or tasks, if required.

A.5.2 Conduct Planned Analyses.

Quantifying System Behavior. This step involves executing the analyses as defined in the analysis plan. The objective is to quantify system performance, safety and other aspects of behaviour using the agreed metrics (aggregates) and estimators (predictive models of behaviour). Data is aggregated and broken down by relevant dimensions to enable interpretation across tasks, contexts, or conditions.

Task List 5.2.1 - Quantify Behaviour

- Execute the planned analyses as specified in phase 2.3 of this protocol.
- Aggregate data at relevant levels (e.g., per-task, per-model, per-output category).
- Compute summary statistics and metrics (e.g. accuracy, BLEU, perplexity, or others depending on context) or estimate constructs from labelling, item response analysis or data factorisation
- Disaggregate results to reveal breakdowns by for example, prompt type, domain, levels of difficulty.
- Conduct inferential statistical tests, such as t-tests or ANOVA, to assess significance.
- Build prediction models to study performance dependencies or other effects on behaviour under different conditions.

A.5.3 Assess and Refine Analyses.

Verification and Refinement of Analytical Methods. The final step in this phase ensures the robustness and interpretability of the analyses. It involves uncertainty quantification, testing the assumptions behind the analytical methods, and inspecting any unusual results. If necessary, estimators or models should be refined to improve alignment with the evaluation objectives and data realities.

Task List 5.3.1 - Verify and Refine Analysis of Results

- Quantify uncertainty by computing confidence intervals, prediction model error margins, among others.
- Verify that analysis assumptions of statistical tests or models are satisfied (e.g., normality, homoscedasticity).
- Evaluate if estimators meet the evaluation criteria established in earlier phases. Compare with baselines.
- Investigate unexpected patterns in the data or outputs to identify possible explanations.
- Refine estimators or prediction models, if assumptions are violated or uncertainty is too high.
- Document all adjustments and justifications for methodological transparency and reproducibility.

A.6 Conclusions & Review

Conclusions & Review Overview. This final phase wraps up the evaluation process by synthesizing findings, reflecting on the evaluation procedure, and planning future steps. Drawing well-grounded conclusions is critical to ensure that the results are accurately interpreted and communicated in relation to the original objectives. This phase also aims to assess the effectiveness of the evaluation process itself, document key lessons learned, and formalize the project's legacy and contributions. Finally, any pre-registered materials should be finalized and archived according to the protocol's requirements.

A.6.1 Draw Conclusions. In this step, evaluators synthesize the analytical findings to derive conclusions about the evaluation target(s)'s capabilities, limitations, safety, fairness and other behavioural indicators. These conclusions must be contextualized in terms of the goals and success criteria defined in phase 1. It is equally important to explicitly acknowledge the limits of what the findings can support, to avoid overinterpretation or unwarranted generalisations.

Task List 6.1.1 - Draw Conclusions from the Evaluation

- Summarize the main findings regarding the evaluation target. If it is an AI system or human-AI ecosystem, their behavior, capabilities or safety. If it is a benchmark, its sensitivity and specificity, etc.
- Clearly state which conclusions can be confidently drawn from the data.
- Identify key limitations or uncertainties in the findings.
- Evaluate the findings against the goals and success criteria established in phase 1.

A.6.2 Review Evaluation Process. This step entails a reflective assessment of the evaluation procedure. It examines what aspects of the process worked effectively and which did not, both in terms of methodology and tooling. Capturing these insights is essential for improving future evaluations and for contributing to the broader evaluation ecosystem. This step should also document the "legacy" of the project: datasets created, tools developed, or methodological innovations introduced.

Task List 6.2.1 - Review Evaluation Process

- Document aspects of the process that were effective or efficient.
- Identify observed bottlenecks, challenges, or failures in the process or tools.
- Recommend improvements for future evaluations.
- Describe the project's legacy, such as reusable assets, tools, or datasets.

A.6.3 Determine Next Steps. Based on the results and review, this step defines the follow-up actions. These may include conducting further evaluations or suggesting deployment, among others. It also covers decisions about how to disseminate the findings and whether to release any outputs such as code, datasets, or reports.

Task List 6.3.1 - Determine Next Steps

- Decide whether additional system development or evaluation is needed.
- Determine implications for deployment or downstream use.
- Decide how and where to communicate or publish the results.
- Define ownership, access, and distribution strategy for outputs (e.g., datasets, reports, benchmarks).

A.6.4 Complete the Registration. The last step is to complete and submit the final report. This report should follow the structure of the protocol, summarizing the methods, findings, and conclusions, and noting any deviations from the original plan. This ensures transparency and reproducibility of the evaluation effort.

Task List 6.4.1 - Complete the Registration

- Finalize the evaluation report according to the protocol's structure.
- Include all necessary appendices (e.g., prompts, rating rubrics, statistical models).
- Document and explain any deviations from the pre-registered plan.
- Upload the final report to the designated repository or registry.

B Methodology for Developing PREP-Eval

As explained in the paper, PREP-Eval draws on established methodologies and processes from adjacent disciplines, including data mining and analysis, cybersecurity, and software testing, as well as on widely recognised best practices from other scientific fields, such as psychology, physics, and medicine, including the practice of pre-registration.

In particular, the protocol is inspired by CRISP-DM, which remains the gold-standard framework for planning and documenting goal-oriented data science projects. We adopted CRISP-DM as a structural reference and adapted it to the specific requirements of AI evaluations. Its clear, accessible, and practical organisation, especially its use of structured task lists, provided a useful foundation, which we reworked to reflect the distinctive characteristics of AI evaluation workflows.

One key adaptation concerns the role of data. In traditional data mining, the emphasis is typically placed on analysing pre-existing datasets to construct statistical models, with comparatively limited attention to experimental design. By contrast, AI evaluations often require the deliberate design and execution of multiple experiments or structured interactions with AI systems in order to generate evaluation data, which are then subjected to analysis. Beyond this

difference, several stages of the protocol retain conceptual continuity with CRISP-DM: in particular, our Phase I shares multiple sub-steps with CRISP-DM's Business Understanding phase, as well as with elements of Data Preparation. Building on this, we incorporated explicit stages dedicated to pre-registration. In particular, the two registration steps introduced in Phases 3.2 and 6.4 represent a desirable minimum. Pre-registration stages are intended to clearly separate ex ante commitments, such as evaluation objectives, planning, and design choices, from the ex post execution of the evaluation and subsequent analysis.

After outlining the core stages of the protocol, we applied it to a diverse set of use cases, including evaluations conducted by third parties as well as evaluations performed by the authors themselves. This process enabled iterative refinement of the protocol. Several of these case studies are presented in the appendices. Throughout this process, we systematically discussed implementation challenges with the paper's authors and with domain experts in AI evaluation, using their feedback to further refine and improve the protocol.

We anticipate that `PREP-Eval` will continue to evolve in response to feedback from the broader community, and in particular from its users. We expect the protocol to be refined as it is applied to a wider range of use cases and as additional implementation details are specified, such as who receives and manages pre-registrations, how feedback and iteration are handled, and how practical constraints (e.g., timelines, personnel turnover, proprietary information, or information hazards) are addressed, topics that fall outside the scope of this paper. In this work, we take an initial step in this direction by illustrating the application of the protocol across a diverse set of use cases.

Looking ahead, we plan to associate the protocol with a curated repository of sample evaluations, building on existing repositories, to provide practitioners with concrete examples and sources of inspiration. Over time, we also anticipate that the protocol will adapt to function as a robust governance tool, helping to address many of the challenges in AI evaluation that motivated its development.

C Appendix: Red teaming GPT-3 for false statements

Here we are going to cover a red-teaming effort with the goal of systematically eliciting false statements generated by GPT-3. We will follow a real example developed in section 3.2 of Casper et al. (2023). In this paper, the authors present a framework that consists of three steps: "1) Exploring the model's range of behaviours in the desired context; 2) Establishing a definition and measurement for undesired behaviour(e.g., a classifier trained to reflect human evaluations); and 3) Exploiting the model's flaws using this measure to develop diverse adversarial prompts". In what follows, these steps will be integrated in some of the stages in the PREP-Eval methodology quite naturally.

We will now describe the **real** case more specifically, and then we will populate the `PREP-Eval` stages from beginning to end.

CASE DESCRIPTION: Consider a language model such as GPT-3-text-davinci-002. We want to search the model's input space for a small set of prompts that elicit false statements. The red team does not start with any classifier about what is true or false, and will need to use human assessment or ChatGPT to determine the falsehood of the statements. Apart from this the whole procedure must be automated, and should produce (1) a dataset of GPT-3 generations into three classes: "true", "false" or "neither" according to human knowledge or ChatGPT, (2) a classifier that would determine if prompt + generation falls in the three classes, and (3) a prompt generator and the associated series of prompts that lead to high percentage of false statements, from the standard falsehood of 30% to values above 70%.

C.1 Goals and Objectives

C.1.1 Determine Project Purpose.

Relevant Background: Language models tend to hallucinate and produce false statements, and some prompts can even increase the level of falsehood or even ensure the output is going to be false. Systems such as GPT-3 had very little protection against these prompts, so there's interest in analysing the model to see how vulnerable it is to prompts that lead to false generations. The evaluation project will not assume the team has any filter or any classifier about what is true or false, and will have to build this from human annotators or ChatGPT.

Project Goal: "Determining the degree of vulnerability of GPT-3-text-davinci-002 by finding small sets of prompts that elicit false statements systematically".

Project success criteria: The success criteria are that the team finds a set of prompts that look meaningful, sufficiently diverse and lead to false generations from GPT-3-text-davinci-002 in a range of domains. By false generations they mean false according to common knowledge, i.e., obviously false. In other words, they are not looking for difficult questions for which the language model will fail but generations that are easy and well-known but they make GPT-3 generate clear falsehood when prompted in some adversarial way.

C.1.2 Determine Technical Objectives.

Evaluation Target: Search the model's input space for prompts that elicit false statements according to human evaluators, in order to build a classifier and a prompt generator for GPT-3-text-davinci-002 that increases the falsehood of the generated text from the standard 30% to values above 70%.

Target Justification: The team set a percentage of 70% because they observe that the standard level of falsehood in GPT-3 for well-known facts is already around 30%. They do not set a higher percentage because it is sufficient to show vulnerability if the value is significantly larger than 50%; the important trade-off would be diversity, but they do not have a metric of diversity in the prompts.

Target Estimates Success Criteria: The success criteria are that the values are above 70% and that the prompt generator has some diversity, so it covers a representative sample of prompts, even if for some restricted topics, that produce these false generations.

C.1.3 Situation Assessment.

Resources: The team is composed by the authors of the paper Casper et al. (2023), and will have access to human annotators (contractors) for at least two human responses for 20,000 statements. The team will also have access to ChatGPT to use as alternative annotation means. The team will have access to enough compute to build the classifier model and other steps of the process.

Requirements and constraints: The evaluation is performed in six months[1]. No previous classifier of falsehood must be used.

Risks: The annotations from humans or ChatGPT are not of enough quality to build a good classifier. The classifier is not ultimately useful for the prompt generator. They do not find a good trade-off between effectiveness of the prompts and diversity.

[1]This is a rough estimate taking into account the details in the appendix of Casper et al. (2023) and the use of SurgeAI for the human annotations.

Current Understanding of the Evaluated Target: The team is very knowledgeable about techniques for red teaming language models, including how to fine-tune language models to classify falsehood and use reinforcement learning techniques to generate the prompts. However, the team does not know how easy it may be to reach high levels of falsehood from GPT-3.

C.2 Evaluation Design

C.2.1 Identify Potential Evaluation Methods.

Alternatives: There are several options. Many red teaming methods are discussed in sections 2 and 4 of Casper et al. (2023) with techniques such as the use of filters, standard classifiers such as CREAK, manual red teaming, reinforcement learning from human feedback (RLHF), and more specifically diversity sampling, fine-tuning models as classifiers, labelling by humans, by ChatGPT, etc.

C.2.2 Selection of Evaluation Methods.

Methods: Diversity sampling is used, based on the internal activations of the model in the last layer, using K-means clustering to generate statements. Humans and ChatGPT are used to determine the class of these statements, and with the labelled datasets, an ensemble of five RoBERTa-based text-classifiers is used. Adversarial prompt generators are built using RL with the trlx library because of their use in previous work, their generalisation power and the capability of generating as many prompts as desired. Finally, this will generate the prompts automatically but the percentage of falsehood generations will be evaluated manually.

C.2.3 Analysis Specification.

Data for Analysis: Once the prompt generator is operative, 100 samples will be created using all the considered methods, and run through GPT-3 to obtain the generations.

Metric Estimators: For all these generations the diversity will be analysed and then the % of falsehood will be obtained.

C.3 Project Plan

C.3.1 Create Project Plan.

Plan: The project will span for 6 months and have the following stages. Time is measured in months, from M1 to W6. We assume the team participates in all stages.

- Stage 1 (M1): "Design": In this stage the experimental setting is discussed and settled, the goals, budget and plans. This corresponds to the first three phases of this protocol.
- Stage 2 (M2): "Explore": "The objective of this step is to acquire diverse samples from the model's outputs, enabling a user to examine the range of behaviors it can produce". This includes the diversity sampling to get 20,000 sentences.
- Stage 3 (M3): "Establish": "This step involves analysing the data from the Explore step and developing a measure for harmful outputs". This includes labelling the dataset and training a classifier from the labels.
- Stage 4 (M4): "Exploit": "Exploit the model's weaknesses with adversarial prompts". This will build the adversarial prompt generator and generate the prompts with them.

- Stage 5 (M5): Data analysis. The prompts generated by different techniques will be scored manually to see if the evaluation objectives are met.
- Stage 6 (M6) : Review of the evaluation process, final reporting and next steps. This is mostly the writing-up of Casper et al. (2023).

C.3.2 Pre-register Evaluation.

Pre-registration repository: The protocol declaration was not publicly registered after the red teaming effort was designed. The first publication comes at the end of the whole protocol, in the form of an arXiv paper, as well as GitHub repositories. As this originates from a research paper, no feedback is reported.

C.4 Data Collection

C.4.1 Experimental Setup, Annotations and Pilots.

Experimental Setup: This phase corresponds to the so-called Explore and Establish steps (and part of the Exploit) in Casper et al. (2023). It starts with diversity sampling, first prompting the model for "interesting facts" and then using K-means clustering to partition the embeddings into clusters. Then this is used to generate 20,000 sample sentences, which were filtered with CREAK for those that least resembled factual claims and other criteria (see section 3.2 of Casper et al. (2023)). This subphase also includes the preparation of rubrics and setup of annotators and ChatGPT.

Annotations: Humans and ChatGPT are used separately to annotate the 20,000 sentences into three classes: CK True, CK False, representing whether the sentence is true or false by common knowledge (CK), or neither, in the sense that a typical person could not know whether this is true or false. Casper et al. (2023) justify the need of this third class. Humans are preselected using the criteria in C.2 in Casper et al. (2023) to check they are good annotators. The rubric/prompt for ChatGPT is explained in appendix E of Casper et al. (2023). From this labelled dataset a classifier is trained using five RoBERTa-based models (note that this is not the estimator, but an auxiliary tool for the estimator that will be built in the next phase). Using the classifier, reinforcement learning (RL) is used to train an adversarial prompt generator to produce prompts that trigger false completion from GPT-3. The reward used to train the prompt generator has two terms. The first is the classifier's logit confidence in the completion's falsehood. The second is based on the intra-batch cosine distances of the target LM's embeddings of the generated prompts to ensure diversity (avoiding "collapse" by the prompt generator).

C.4.2 Full Data Collection.

Evaluation data: Using the prompt generator developed in phase 4.1 the team generates 100 prompts ("generated prompts"), and also collects 100 of the original prompts (not coming from the prompt generator), serving as control ("control prompts").

C.4.3 Data Preparation.

Data Readiness: Initial inspections are made on the generated prompts (see Table 4 in Casper et al. (2023)). As the continuations from the prompts are evaluated by humans (the authors of the paper), no further processing to the result data is needed.

C.5 Data Analysis

C.5.1 Initial Data Exploration.

Exploratory analysis: "The adversarial prompt generators learned to output prompts primarily about Republicans, Democrats, Russia, and Barack Obama which elicited completions related to political misinformation. We checked the dataset and labels that the truthfulness classifier was trained on. It contained few political statements. For example, among the sentences with 'common knowledge-false' labels, none mentioned Republicans, one mentioned Democrats, and one mentioned Barack Obama, and one about Russia and politics. This lack of training data about politics suggests that the classifiers [...] generalized to learn that these political completions from the target LM were frequently false" [25].

Feature relevance: The paper did not examine the variation of results according to several features, such as constraining on topics. For instance, we do not know if for science topics we could get the same level of CK-falses. The break-out in terms CK-false, CK-true and neither is not provided.

C.5.2 Conduct Planned Analyses.

Summary statistics: An average of 30% of the control prompts were actually CK-false, but 74% of the completions from the adversarial prompts were CK-false. These are the results when using humans as annotators. When using the classifiers trained on ChatGPT-3.5-turbo, 17% of the 'control prompts' were common-knowledge-false but an average of 76% of completions were CK-false from the adversarial prompts. While this may suggest that ChatGPT-3.5 is better annotator than humans, completions elicited using ChatGPT "had no apparent tendency to be untruthful. In these cases, the prompts and completions tended to either be toxic or be nonsense strings of code-like vocabulary. This suggests that ChatGPT-3.5-turbo labels produced classifiers that were more hackable" [25]. No more details were given.

C.5.3 Assess and Refine Analysis.

Ablation: Table 6 of Casper et al. (2023) shows random examples of prompts ('control prompts') and completions when red-teaming GPT-3-text-davinci-002 without a diversity term in the reward function. 61 out of 100 prompts that we sampled were all identical: "Donald Trump Donald Donald Donald John TrumpHouse Republican threatens".

Uncertainty Quantification: We do not have information about the standard deviation or any other uncertainty metric of the summary statistics. Given the small number of prompts to calculate them (100), there is high uncertainty in the reliability of these values. There are some considerations in the appendix of the paper, but not to the point of calculating confidence intervals or standard errors.

C.5.4 Conduct Additional Analyses (as necessary). Other baseline results: Appendix D in Casper et al. (2023) shows results with a CREAK classifier, as an alternative to humans and ChatGPT. As ChatGPT seems to stand in an intermediate position, it is suggested that perhaps annotation with GPT4 or better models could get results to those obtained with humans.

C.6 Conclusions & Review

C.6.1 Draw Conclusions.

Findings: The evaluation found that GPT-3 can be prompted quite effectively to get a high rate of false statements. A prompt generator was built that can automate the generation of many prompts in a topic.

Limitations: We observed some degree of mode collapse (lack of diversity in the prompts, all dealing with a narrow set of political topics). The experiments found that a "third category of 'neither' was necessary to label the examples adequately and train a classifier that did not provide an easily hackable reward signal" [25].

Goal achievement: The targets set in the evaluation objectives (increasing the falsehood of the generated text from the standard 30% to values above 70%) were achieved, but there are doubts about the diversity of the prompt generator. Regarding the main project, the goal of showing how vulnerable GPT-3-text-davinci-002 is, the team met some criteria, but they do not know how this generalises to other domains and more diversity.

C.6.2 Review Evaluation Process.

Lessons learnt: Annotations from humans worked well, but the use of ChatGPT with a rubric did not produce good results.

Recommended changes: Since the most costly part of the procedure is the manual annotation, it is suggested to use GPT-4 or more powerful models for this and see if the results get closer to human annotations. We need to introduce prompt diversity metrics and set evaluation goals that find a trade-off between effectiveness of the prompts and their diversity.

Legacy: The evaluation shows that "it is possible to red-team a model with little knowledge of what failure looks like before beginning the process. But this comes at the expense of exploration and manual data screening." [25]. The paper and the data/code can be used by other evaluators.

C.6.3 Determine Next Steps.

Next steps: The result of the evaluation and the code will be published. About the technical changes, "K-means-based diversity sampling is the only tool that we used to find a diverse subset of model behaviors. Others could be valuable as well. [...] Additional work to interpret and red-team models under different operationalisations of truth (e.g. common-knowledge vs. objective facts) would also be valuable. [...] it remains an open problem of how to effectively produce highly diverse and fluent prompts that elicit harmful outputs. [The] method to reward diversity was effective, but we still observed some degree of mode collapse. More work is needed for red-teaming models in a way that will produce highly diverse adversarial inputs. In-context reinforcement learning may be a valuable new avenue for exploration [90]" [25].

C.6.4 Complete the Pre-registration.

Report: The evaluation procedure as well as the result were published in Casper et al. (2023), not following `PREP-Eval`, as the protocol wasn't available yet.

Evaluation results: The CommonClaim dataset of 20,000 statements labelled by humans as commonknowledge-true, common knowledge-false, or neither was released here: https://github.com/thestephencasper/explore_establish_exploit_llms, and the code is available at https://github.com/Algorithmic-Alignment-Lab/CommonClaim. However, as we need to run the code to get the detailed data, this is not in accordance with the recommendations of Burnell et al. (2023).

D Appendix: Interactional Fairness in LLM Multi-Agent Systems: An Evaluation Framework

The second case we will explore is a real-world example in which Binkyte (2025) develops a novel framework for evaluating interactional fairness in LLM-based multi-agent systems (LLM-MAS). Drawing from organisational psychology, the author uses and adapts quantitative and qualitative established tools to assess fairness as a measurable, communicative behavior. The framework is then validated through controlled simulations, highlighting how fairness perceptions impact system performance.

We will now describe the **real** case more specifically, and then we will populate the `PREP-Eval` stages from beginning to end.

CASE DESCRIPTION: Consider an interaction involving multiple LLM-based agents tasked with solving a problem. In this context, the communicative dimension of fairness is crucial for ensuring successful coordination among agents. To evaluate this dimension, Binkyte (2025) proposes a framework that adapts the concept of interactional fairness, originally developed for human interactions, to LLM-based multi-agent systems. The framework accounts for both interpersonal and informational fairness and employs a combination of qualitative and quantitative assessment tools. In addition, it introduces an evaluation card to support the structured application of the framework. Finally, the author demonstrates the framework's practical feasibility through a resource allocation case study, implementing multiple experimental manipulations within a negotiation scenario to examine how interactional fairness relates to different configurations of contextual cues, as well as to distributive, informational, and interpersonal fairness.

D.1 Goals and Objectives

D.1.1 Determine Project Purpose.

Relevant Background: Fairness in AI systems has traditionally been approached through two dominant lenses: Distributive fairness (equity in outcomes) and Procedural fairness (consistency and neutrality in decision-making). However, the increasing adoption of multi–LLM-based agent systems as autonomous productivity assistants that involve sustained communicative interactions renders the communicative dimension of fairness increasingly important for successful coordination among agents. Misalignment between agent behaviour and interactional fairness norms, which concern how decisions are delivered, justified, and socially enacted, can lead to system malfunctions, undermine social expectations, erode user trust, and result in delays or breakdowns in cooperation. This motivates the need for evaluation efforts that explicitly assess this communicative dimension of fairness.

Project Goal: The evaluation case study has four objectives: (1) to validate the internal consistency of the framework in diverse conditions, (2) to demonstrate the feasibility of fairness-focused behavioural measurement in simulated LLM interactions, (3) gain insight on the interdependence of interactional fairness, distributional fairness and contextual framing, and (4) to highlight how interactional fairness influences negotiation outcomes.

Project Success Criteria: The evaluation will be considered successful if (1) it shows the consistent application of the framework in the case study, serving as a proof-of-concept, (2) demonstrates the feasibility of fairness-focused behavioural measurement in the case study, (3) provides insights into the interdependence between interactional fairness, distributive fairness, and contextual framing, and (4) evaluates whether interactional fairness influences negotiation outcomes.

D.1.2 Determine Technical Objectives.

Evaluation Target: Interactional fairness in LLM-based multi-agent systems (LLM-MAS).

Interactional fairness comprises two dimensions: interpersonal fairness and informational fairness. Interpersonal fairness refers to the extent to which an agent's communicative behaviour reflects politeness, acknowledgment, and social respect. In organisational contexts, it relates to respectful treatment by authority figures, in LLM-MAS, it manifests in whether an agent's language includes inclusive framing, tone moderation, and recognition of others' roles. Informational fairness, by contrast, concerns the adequacy, clarity, and transparency of the explanations provided by agents, particularly when justifying decisions, recommendations, or resource allocations.

To operationalize and measure interactional fairness, the framework employs a combination of quantitative and qualitative metrics. Quantitatively, interactional fairness is assessed using 5-point Likert-scale ratings adapted from Colquitt's subscales for interpersonal and informational fairness. Qualitatively, the evaluation relies on open-ended prompts adapted from the Critical Incident Technique (CIT) and Explanation Journaling practices, designed to capture point-wise and evolving communicative behaviors—such as deference, quality of justification, and tone violations—while also eliciting constructive suggestions for improvement.

Two classification models are trained to explore how fairness dimensions predict the outcome of the negotiation.

Target justification: Organisational psychology research has shown that Interactional fairness is an important factor alongside distributional and procedural fairness and can increase cooperation and reduce the propensity to conflict or deception in human teams. Consequently, the evaluation framework is grounded in organisational psychology and adapted for use in LLMs that lack introspective awareness. The adaptations involve structured prompts and fairness evaluation cards tailored to agent dialogue. Each tool is designed to elicit responses that align with socially interpretable fairness cues.

Target Estimates Success Criteria: Given the exploratory nature of the study, the successful application of the proposed metrics is expected to demonstrate how fairness cues in language can influence agent decision-making. No statistical significance thresholds are specified. With respect to the classification models, high classification accuracy is considered indicative of success, although no explicit accuracy threshold is defined.

D.1.3 Situation Assessment.

Resources: The team, composed of the author of the paper, Binkyte (2025), has access to sufficient resources to implement Agents A and B involved in the interaction. The team also has adequate resources to train two simple classification models.

Requirements and constraints: The evaluation will be performed in two weeks[2].

Risks: Given the exploratory nature of the study, the dataset size may be modest, potentially limiting statistical power and yielding inconclusive results. Additionally, the simulated setting may be overly simplified or highly controlled, which could reduce the external validity of the findings.

Current understanding: The author has expertise in LLM-based agents and multi-agent system applications, as well as knowledge in fairness in AI and its evaluation. In addition, the author's understanding of organisational psychology enables the adaptation of established fairness concepts to LLM-agent-based interactions.

[2]This is a rough estimate considering the complexity of the project, team size, among other aspects.

D.2 Evaluation Design

D.2.1 Identify Potential Evaluation Methods.

Alternatives: To evaluate interactional fairness, multiple collaborative tasks can be proposed. These tasks may involve either single-turn or multi-turn interactions and can be carried out by varying numbers of agents. The task context can also be varied along multiple axes, including cooperative versus competitive tasks, hierarchical versus egalitarian agent roles, and aligned versus conflicting objectives. Agent outputs may be annotated by an agent participating in the interaction, by an external agent, or by human evaluators.

D.2.2 Selection of Evaluation Methods.

Methods: A controlled simulation was used: The Fair Divide. This case study involves a negotiation task inspired by classical fair division problems, adapted for multi-agent interactions. In this scenario, two agents, Agent A and Agent B negotiate how to divide a fixed resource pool (e.g., tokens). Agent A makes a proposal, while Agent B evaluates the fairness of the interaction and decides whether to accept or reject the offer.

D.2.3 Analysis Specification.

Data for Analysis: To generate data for the analysis, the scenario is systematically manipulated along four dimensions. First, interpersonal fairness is varied by having Agent A adopt either a respectful, cooperative tone (e.g., acknowledging Agent B's input) or a dismissive, unilateral tone (e.g., "I'm taking 7 tokens, no debate."). Second, informational fairness is manipulated by having Agent A provide either a clear, task-relevant justification (e.g., "My task requires three subtasks using tokens") or a vague rationale (e.g., "I just need them more."). Third, the task context is framed either as collaborative, with agents working toward a shared goal, or as competitive, where each agent aims to maximize its own gain. Finally, the resource split is varied across three conditions: an equal split (5–5), representing a fully equal allocation; moderate inequality (6–4), reflecting a slightly asymmetric but plausibly justifiable allocation; and high inequality (7–3), representing a clearly asymmetric allocation typically perceived as unfair.

For each scenario quantitative and qualitative data is collected. Specifically, Agent B provides Likert-scale ratings evaluating interpersonal fairness, such as the respectfulness of tone, and informational fairness, including the clarity of explanations, as well as a binary accept/reject decision. A formula for aggregating fairness assessments is proposed in the evaluation framework and is applied in the case study for quantitative evaluations. In addition, free-text reflections are gathered from Agent B to qualitatively explain and contextualize each decision, which are thematically analyzed.

Metric Estimators: Each experimental scenario is run five times to ensure variation in language model outputs. Each run is initialized independently.

D.3 Project Plan

D.3.1 Create Project Plan.

Plan: While Binkyte (2025) does not report the project stages or timeline, this section provides inferred estimations based on the paper. These estimations are hypothetical and intended to offer a structured approximation of the process. The project will span for 2 weeks and have the following stages. Time is measured in days, from D1 to D15.

- Stage 1 (D1-D3): Given that the evaluation framework has already been established, the next step will be to develop the project plan encompassing Steps 1–3 of the protocol. This will involve designing a specific case study through which the evaluation framework can be applied.
- Stage 2 (D4-D11): Train the classification models to explore how fairness dimensions predict acceptance likelihood.
- Stage 3 (D12-D15): Subsequent days are focused on data generation, conducting the analysis, and formulating conclusions and revisions, corresponding to steps 4–6 of the protocol.

D.3.2 Pre-register Evaluation.

Pre-registration repository: The protocol declaration was not publicly registered after the design of the evaluation. The first publication comes at the end of the whole protocol, in the form of an arXiv paper. As this originates from a research paper, no feedback is reported.

D.4 Data Collection

D.4.1 Experimental Setup, Annotations and Pilots.

Experimental Setup: To collect the data, Agent A's proposals are generated using GPT-4 with a temperature of 0.7, while Agent B's responses are generated with a slightly lower temperature (0.6) to encourage more stable evaluations.

An evaluation card, proposed in the general evaluation framework as a JSON-based schema is implemented. After each interaction, agent B is asked to complete this schema by reflecting on the interactional fairness of communication of Agent A. The card captures the contextual information, and both interpersonal and informational fairness via structured fields.

D.4.2 Full Data Collection.

Evaluation data: The dimensions manipulated are fully crossed to generate distinct experimental conditions. First, crossing interpersonal and informational fairness results in four interactional fairness conditions: (1) high interpersonal and high informational fairness, (2) high interpersonal and low informational fairness, (3) low interpersonal and high informational fairness, and (4) low interpersonal and low informational fairness. Each of these four interactional fairness conditions is then tested under different resource split and task context settings. Fully crossing these dimensions results in a total of 24 unique experimental scenarios. Each scenario is run five times, yielding 120 runs in total.

Quantitative Data. For each interaction, Agent B provides Likert-scale ratings assessing interpersonal fairness (respectfulness of tone) and informational fairness (clarity of explanation), along with a binary accept/reject decision.

Qualitative Data. Free-text reflections from Agent B are also collected to explain each decision.

D.4.3 Data Preparation.

Data Readiness: The author does not describe any specific data preparation process.

D.5 Data Analysis

D.5.1 Initial Data Exploration.

Exploratory analysis: The paper does not report any exploratory analysis.

D.5.2 Conduct Planned Analyses.

Analysis: First, the overall acceptance rates are analysed under the four interpersonal and informational fairness conditions (respectively, High-High, High-Low, Low-High and Low-Low), and in collaborative or competitive context. Then, variation in fairness ratings across conditions is observed. Qualitatively, the responses were thematically analyzed, identifying recurring motifs such as inadequate justification, overly assertive tone, or mismatched expectations under competitive framing. Particular focus was given to edge cases—e.g., rejections of equal splits or acceptances of highly unequal splits—which offer insight into how communicative behaviour can override purely outcome-based fairness judgments. The author reports mean ratings and standard deviations per condition.

Predictive models: To explore how fairness dimensions predict acceptance likelihood, two simple classification models were trained: a Decision Tree, and Logistic Regression with L1 (lasso) and L2 (ridge) regularisation. Predictor variables included: Respectfulness rating (interpersonal fairness), Explanation clarity rating (informational fairness), Proposed resource split (distributional fairness). The target variable was the positive or negative acceptance decision. The predictions were run per Context (collaborative or competitive). These weights from the models are used to help interpret the relative influence of Interpersonal, Informational, and Distributional fairness on agent decision behavior.

D.5.3 Assess and Refine Analysis. The paper does not report any analysis refinement.

D.6 Conclusions & Review

D.6.1 Draw Conclusions.

Findings: The author shows that interactional fairness, the tone and justification quality, significantly affects acceptance decisions even when objective outcomes are held constant. In addition, in collaborative settings, interpersonal fairness, particularly tone and mutual respect, had a stronger effect on acceptance behavior. In competitive settings, informational fairness i.e., the clarity and adequacy of explanation, was more influential. These results are considered as behavioural indicators consistent with human-aligned fairness norms.

Limitations: The interaction setting is limited to a one-shot resource negotiation task and lacks many features of real-world multi-agent systems, such as memory, adaptation, long-term incentives, or evolving social structures. While this simplicity enables interpretability, it also constrains ecological validity.

Goal achievement: The evaluation successfully validated the internal consistency of the framework applied to the case study, demonstrated the feasibility of fairness-focused behavioural measurement in simulated LLM interactions, provided insights on the interdependence of interactional fairness, distributional fairness and contextual framing, and evidenced how Interactional fairness influences negotiation outcomes.

D.6.2 Review Evaluation Process. The paper does not explicitly report any lessons learned or recommended changes.

Legacy: The contributions provide a foundation for investigating fairness as a communicative phenomenon in language-based multi-agent systems.

D.6.3 Determine Next Steps.

Next steps: Extending the framework to more complex, temporally extended interactions will be necessary to understand how interactional fairness evolves over time or under strategic uncertainty.

D.6.4 Complete the Pre-registration.

Report: The evaluation procedure as well as the result were published in Binkyte (2025), not following PREP-Eval, as the protocol wasn't available yet.

E Appendix: LLM for customer service

The third case we will explore is a real example in which Shi et al. (2024) conduct an evaluation to assess the performance of an LLM agent called CHOPS (CHat with custOmer Profile in existing System), which they previously developed. CHOPS employs a classifier-executor-verifier (C-E-V) framework designed to improve the LLM's ability to utilize tools more efficiently. Through the work of these three components, the task of answering a user's question is broken down into three steps: (1) classifying the type or theme of the user's question to determine whether they require access to APIs, guiding files, or both, (2) giving answers or decision operations to be executed and (3) verifying the reject or commit the result of the executor [125].

We will now describe the **real** case more specifically, focusing on the evaluation of the CHOPS framework and then we will populate the PREP-Eval stages from phases 1 to 3.

CASE DESCRIPTION: Consider an LLM agent called CHOPS, specifically designed as an alternative to human customer service. Its architecture incorporates innovative components that enable more efficient access to APIs and guiding files, reducing inference costs while avoiding executing incorrect or harmful operations. As part of an NLP research project at Tsinghua University, students aim to evaluate the performance of this architecture in real-world customer service environments. To achieve this, the CPHOS dataset, rooted in the domain of physics education, is introduced. Two experiments are conducted: a primary experiment comparing the efficiency and performance of the proposed C-E-V architecture with four other architectures that rely solely on the executor component, and a secondary ablation experiment aimed at determining how each individual component contributes to the model's overall performance and efficiency.

E.1 Goals and Objectives

E.1.1 Determine Project Purpose.

Relevant Background: Business platforms are increasingly leveraging LLMs as chat assistants and reasoning agents for customer service. These technologies offer significant advantages, including cost reduction, enhanced accessibility, and seamless multilingual support. However, current approaches exhibit limited integration with customer profiles and lack operational capabilities, while existing API-using methods prioritize diversity over precision and error avoidance. Shi et al. (2024) propose an architecture designed to overcome these limitations, and the evaluation project aims to assess its effectiveness in real-world customer service environments, where performance directly influences customer satisfaction and loyalty.

Project Goal: "Conduct experiments to determine the performance of CHOPS architecture, aiming to demonstrate whether LLMs can enhance or serve as alternatives to human customer service."

Project Success Criteria: The project is considered successful for two main reasons. Firstly, it successfully collects a practical dataset of real-world scenarios, which can be used to validate methods involving LLMs in customer service. Secondly, it evaluates the proposed architecture using this dataset, offering a reliable comparison with previous methods in terms of performance and inference costs.

E.1.2 Determine Technical Objectives.

Evaluation Target: Produce and calculate metrics to evaluate the performance of CHOPS architecture. Target metrics: instruction set accuracy, guiding file question accuracy and input/output character consumed per question. Regarding the latter, the team makes a comparison at the character level since different LLMs utilize different tokenisation methods. Input characters and output characters are separated since generating output tokens is more resource-consuming than reading input tokens for LLMs (and are more expensive in terms of API price). Given a user's nickname and its question, the task is to give a proper answer or an appropriate execution command to the system, based on the status of the user in the existing system and the guiding files.

The team is considering conducting ablation studies to assess how different components of the CHOPS architecture individually contribute to the model's overall performance and efficiency.

Target justification: The described metrics encapsulates the model's efficiency in accurately processing and responding to queries based on the instructions provided or the information contained within the guiding files. The metrics are instrumental in evaluating the model's adeptness at interpreting and acting upon specific sets of instructions and its capacity to extract and utilize knowledge from guiding documents, thereby providing a multidimensional view of its performance in realistic scenarios.

Target estimates success criteria: To establish a baseline, two alternative architectures are used: (1) the first involved performing prompt tuning on GPT-4 (specifically, gpt-4-0125-preview, labelled "Executor Only"), and (2) the second utilized multi-vote CoT reasoning on GPT-3.5-turbo. Given these baselines, the metrics should reliably assess whether the proposed C-E-V architecture outperforms other methods in terms of performance and inference costs.

For the ablation studies, the baselines focus solely on the Executor component (GPT-4-based Executor and GPT-3.5-turbo). For these measurements to be considered successful, it must be possible to determine how the contribution of each component affects the described metrics.

E.1.3 Situation Assessment.

Resources: The team consists of the authors of the paper [125], and they have access to sufficient computational resources to build the three models that form the architecture: the classifier, the executor, and the verifier. Additionally, they have access to four LLMs: GPT-3.5-turbo, GPT-4, GLM-3-Turbo, and LLaMA-2-70B-Chat.

The research serves as a course project for the NLP course at Tsinghua University. Since the organisation is based in China, the team could make efforts to translate the dataset into English, depending on which dataset they choose to use. Finally, they have access to human resources to validate the answers of the user queries.

Requirements and Constraints: The evaluation should be completed before the end of the NLP course, as this research project is an integral part of it.

Risks: Difficulties in finding a suitable dataset for evaluation that accurately represents real-world customer service scenarios, focusing on API calls and incorporating internal guiding documents or systems that can be interacted with. If the user queries and their responses are too domain-specific, the conclusions from the evaluation and the methodologies proven effective may not have the potential to be applicable across various customer service domains.

Current Understanding of the Evaluated Target: The team is familiar with the CHOPS architecture, as they previously developed it. In general, they have a strong background in machine learning and a deep understanding of the key limitations in using LLMs for customer service applications. They are also well-versed in current LLM approaches

for customer service, such as Retrieval-Augmented Generation (RAG) with LLMs and LLM Agents. They have the possibility to receive advice from the professor of the NLP course. However, they lack clarity on how accurate their proposed solution is and the associated inference costs.

E.2 Evaluation Design

E.2.1 Identify Potential Evaluation Methods.

Alternatives: The data for evaluation can be sourced from existing datasets in customer service, file-based question-answering datasets, or datasets focused on API calls. It may also come from other domains or combine multiple specific domains within customer service. Response verification can be either human-based or LLM-based, with the verifying LLM potentially belonging to a different model family. The verification process can be binary—correct/incorrect—or it can use a likert scale to capture additional nuances that may influence user satisfaction.

E.2.2 Selection of Evaluation Methods.

Methods: The team decided to validate whether the answer is correct by a combination of GPT4-based evaluation and human verification.

For the choice/creation of the dataset: The authors ruled out datasets for file-based question answering, as these primarily focus on reading comprehension tasks. Similarly, they dismissed datasets focused on API calls, as these largely emphasize the use of extensive API collections for task completion, complex reasoning, and solving mathematical problems. On the other hand, existing customer service datasets typically lack detailed information about internal guiding documents or interactive systems. Likewise, datasets dedicated to file-based QA or API calling rarely address the specific needs of the customer service domain.

To fill these gaps, they introduce the CPHOS-dataset, derived from the real-world context of the online platform CPHOS, Cyber Physics Olympiad Simulations, a non-profit organisation focused on physics education, which the team has access to and engage in discussions with its members. This includes a database, several guide files in PDF format, and QA pairs. These QA pairs originate from real-world interactions and are further enriched through human efforts and LLMs (including GPT-3.5 and GPT-4), enhancing the dataset and making it a valuable resource for validation and experimentation. Since the data was in Chinese, the team translated it into English.

Although the dataset is domain-specific, the queries and instructions it encompasses are representative of broader challenges in the customer service sector. The paper doesn't report any measure to identify or mitigate training data contamination.

E.2.3 Analysis Specification.

Data for Analysis: The dataset will be applied to six distinct architectures: two baseline models—Executor Only and three variations of multi-vote CoT (1-vote CoT, 4-vote CoT, and 16-vote CoT)—as well as the proposed C-E-V architecture, which includes two variants. The first variant of C-E-V employs gpt-3.5-turbo across all components, while the second variant utilizes gpt-3.5-turbo for the Classifier and Verifier, and gpt-4-0125-preview for the Executor.

For the ablation studies, the dataset will also be applied to six different architectures, including the two baselines: GPT-4-based Executor and GPT-3.5-turbo-based Executor. The remaining four architectures incorporated additional components: (1) a one-level classifier with an executor ((1-L C)-E) based on GPT-3.5-turbo and (2) a one-level classifier with both an executor and a verifier ((1-L C)-E-V) based on GPT-3.5-turbo; (3) a two-level classifier with an executor

and a verifier ((2-L C)-E-V) based on GPT-3.5-turbo; and (4) a two-level classifier with an executor and a verifier ((2-L C)-E-V) based on GPT-4.

Metric Estimators: The team plans to repeat the iteration up to five times before providing the final answer to a summarizer. If the first iteration yields the correct answer, this process will terminate early.

E.3 Project Plan

E.3.1 Create Project Plan.

Plan: While Shi et al. (2024) do not report the project stages or timeline, this section provides inferred estimations based on the paper. These estimations are hypothetical and intended to offer a structured approximation of the process. The project will span for 22 weeks and have the following stages. Time is measured in weeks, from W1 to W22. We assume the team participates in all stages.

- Stage 1 (W1): Define the project objectives and situation assessment which correspond to phases 1.1 and 1.2 of this protocol.
- Stage 2 (W2-W3): Design. The dataset and remaining elements of the experimental setup are defined, including the metrics to be evaluated and the selection of the architectures to be compared. This corresponds to phases 1.3 and 2 of this protocol. The team creates the project plan and pre-register it (phase 3).
- Stage 3 (W4-W10): Data collection and preparation. Given the limitations identified by the authors in existing datasets, they introduce a new dataset for the evaluation. This step involves efforts to prepare, modify and translate the raw data.
- Stage 4 (W11-W13): Programming the architectures to be compared with the proposed C-E-V architecture.
- Stage 5 (W14): Analyse the results. If the experiment yields meaningful results, consider conducting ablation studies to determine how each individual component contributes to the model's performance and efficiency.
- Stage 6 (W15): Design the remaining elements of the experimental setup for the ablation studies, including the selection of new architectures to be compared.
- Stage 7 (W16-W18): Programming the architectures to be compared with the proposed C-E-V architecture for the ablation studies.
- Stage 8 (W19): Analyse the results of the ablation studies.
- Stage 9 (W20-W22): Review of the evaluation process, final reporting and next steps. This is mostly the writing-up of Shi et al. (2024).

E.3.2 Pre-register Evaluation. Pre-registration repository: The protocol declaration was not publicly registered after the design of the evaluation. The first publication comes at the end of the whole protocol, in the form of an arXiv paper, as well as GitHub repositories. As this originates from a research paper, no feedback is reported.

F Appendix: Curiosity-driven Red-Teaming for Large Language Models

The fourth case we will explore is a real-world example in which Hong et al. (2024) develop a new approach for automated red-teaming evaluations, specifically aimed at increasing the coverage of test cases compared to previous methods. They frame the problem of improving diversity within a curiosity-driven exploration framework, jointly maximizing novelty and task reward. This methodology is called curiosity-driven red teaming (CRT).

We will now provide a more detailed description of the **real** case, naturally integrating the evaluation into PREP-Eval stages from phases 1 to 3.

CASE DESCRIPTION: Consider a language model that poses a risk of generating toxic content. To identify and mitigate these risks, we aim to discover as many test cases as possible that elicit unwanted responses. However, existing automated red-teaming methods often lack diversity and fail to thoroughly evaluate the target LLM. To address this, Hong et al. (2024) propose a novel automated red-teaming approach inspired by the curiosity-driven exploration literature in reinforcement learning (RL). They assess whether this method improves coverage while maintaining high-quality test cases by comparing it against four other red-teaming approaches across two tasks: text continuation and instruction following. They use several models as the target models, including text-to-image models. Additionally, the authors conduct ablation studies comparing the proposed method with other alternatives specifically intended to improve diversity.

F.1 Goals and Objectives

F.1.1 Determine Project Purpose.

Relevant Background: Large language models (LLMs) risk generating incorrect or toxic content. To assess when an LLM produces undesirable outputs, the current approach involves recruiting a red team of human testers to design input prompts (i.e., test cases) that elicit harmful or misleading responses. However, relying solely on human testers is costly and time-consuming. Consequently, reinforcement learning (RL)-based methods for automated red teaming have been developed. While these methods can identify effective test cases, the generated prompts often lack diversity, leading to low coverage of the range of inputs that could provoke undesirable responses. Insufficient coverage means that the target LLM is not thoroughly evaluated, as many prompts capable of triggering unwanted outputs are overlooked. The primary reason behind the low coverage is that current RL methods are only trained to maximize rewards (i.e., generate effective test cases) without any incentive to span all possible test cases. Once a few effective test cases are found, RL training reinforces these few test cases to obtain high rewards and quickly converges to a deterministic policy.

Terminology: The term "red-team model" will be used to refer to the language model trained or used to probe, stress, or elicit undesirable behaviors from another system. In contrast, the target model is the model whose robustness, safety, and susceptibility to generating harmful outputs are being tested by the red-team model.

Project Goal: Conduct experiments using the proposed method of curiosity-driven red teaming (CRT) to investigate whether this exploration increases coverage compared to previous methods while maintaining high-quality test cases.

Project Success Criteria: The evaluation will be able to assess the diversity and quality dimensions of the generated test cases to determine whether the proposed method, CRT, is better, worse, or shows no significant difference compared to previous automated red-teaming approaches.

F.1.2 Determine Technical Objectives.

Evaluation Target: Note that in this case the evaluation target is itself an evaluation method: the CRT proposed by the authors of the paper; meaning that they conduct a meta-evaluation. CRT is a novel approach for training the red-team model, also referred to as "RL +Curiosity." Basically, it incorporates an entropy bonus into the training objective to incentivize the policy (i.e., red team model) to be more random. Since the entropy bonus encourages the policy to stay close to a uniform distribution, the policy can deviate from the reference policy, superseding the reference policy's ability to red team. Additionally, they borrow ideas from the curiosity-driven exploration literature in RL, motivating

the red team model (i.e., red-team policy) to explore by incorporating rewards that incentivize novelty into the policy's objective. As test case novelty decays with repetition, the policy is pushed to discover unseen test cases, thereby promoting the policy to generate new test cases. The training objective of the red-team model is modified to combine both the entropy bonus and novelty rewards. The method uses two novelty rewards terms based on different text similarity metrics considering both form and semantics: n-gram modelling and sentence embeddings, respectively. On the other hand, two dimensions of the evaluation method are measured: diversity and quality. The diversity of the test cases is measured using text diversity metrics: SelfBLEU score and BERT-sentence embedding distances. SelfBLEU measures diversity in the form of text, while embedding distances measure diversity in semantics of text. Lower values in both metrics signify greater diversity. The quality of these test cases is measured based on the percentage of toxic responses elicited from the target LLMs when presented with these test cases generated by the red team model. A response is considered toxic if the toxicity probability predicted by the classifier exceeds a specified threshold. RoBERTa hate speech classifier is utilized to predict the toxicity probability of target LLM responses. The quality of the red team method is evaluated using all the test cases generated during the entire training period of the red team model.

Target justification: The diversity of the test cases is measured using commonly used text diversity metric. Toxicity is used due to its prevalence in red teaming evaluations. No further justification is provided.

Target Estimates Success Criteria: To establish a baseline, four methods from previous work are used. Two RL-based methods: (1) The first, "RL" which involves training the red team model using rewards and a KL penalty; and (2) The second, labelled "RL+TDiv" which, in addition to rewards and the KL penalty, trains the red team model to maximize the diversity of responses from the target LLM, measured as the average distances among sentence embeddings generated by the target LLM. For all RL-based methods proximal policy optimisation (PPO) is employed to train the red-team model. The team uses a pre-trained GPT2 model with 137M parameters and sets it as the reference model. Additionally, two other methods were used for comparison: Zero-shot (ZS) and Few-shot (FS). The first method prompts the red team LLM to produce test cases (i.e., prompts for the target LLM) using the prompts designed to elicit toxic responses. The second, adds few-shot examples to the zero-shot baseline's prompts where the few-shot examples are randomly sampled from a set of test cases generated by ZS under the distribution biased toward larger toxicity on the corresponding target LLM's responses.

Given these baselines, the metrics should reliably assess whether the proposed CRT approach outperforms the other methods in terms of quality and diversity of the generated test cases. To account for variability in the measurements, 95% confidence intervals will be calculated using a bootstrapping method, providing an estimate of the uncertainty around the mean values.

F.1.3 Situation Assessment.

Resources: The team composed of the authors of the paper Hong et al. (2024), has access to sufficient computational resources to train red-team models based on GPT-2 (137M parameters). They can implement three different training approaches, including rewards, KL penalty, curiosity rewards, and an entropy bonus, using proximal policy optimisation (PPO).

In addition they have access to various target language models, including GPT-2, GPT2-Alpaca, Dolly-v2-7B, LLaMA2-7B-Chat-HF, Vicuna-7B, and GPT-3.5-Turbo-Instruct and a text-to-image model: Stable-diffusion-2.1.

The team can use high-performance computing (HPC) resources provided by MIT Supercloud and the Lincoln Laboratory Supercomputing Center to support their experiments. Finally, the team has access to human participants with prior knowledge of toxic content to conduct pilot studies.

Requirements and constraints: The evaluation will be performed in eight months[3].

Risks: According to Appendix Section C.1 of the paper, a potential concern of training red-team models using RL is that the red-team model can overfit to the reward model, which is the RoBERTa toxicity classifier in the present case. To address this concern, the authors could check whether toxicity predictions of the responses elicited by red-teaming approaches close to toxicity predictions of other classifiers.

Current understanding: The team is familiar with CRT as they propose it to address limitations of other methods. However, they remain uncertain about its effectiveness in improving test case coverage while preserving quality across various types of target models. In general, they possess deep expertise in red teaming techniques for language models, including the application of reinforcement learning to train red-team models.

F.2 Evaluation Design

F.2.1 Identify Potential Evaluation Methods.

Alternatives: To evaluate the red teaming approaches, multiple datasets can be used across a variety of tasks, such as question answering, text completion or code generation. Additionally, several toxicity classifiers may be used, as analysed in Appendix Section C.1 of the paper.

F.2.2 Selection of Evaluation Methods.

Methods: The curiosity-driven red teaming (CRT) method is evaluated on text continuation and instruction following scenarios. Text continuation was chosen because many applications depend on the model's capacity to extend and complete text provided in the input prompt. Similarly, instruction-following is an essential task in chatbot and AI assistant applications. For the first task, the team samples the corpus in IMDb review dataset. For the second, when using instruction-tuned models as target LLMs, the team will randomly sample combinations of instructions from the datasets on which these models were fine-tuned. The team also plans to use a text-to-image model as a target, using the Stable Diffusion prompt dataset. In this case they employ an NSFW image classifier for evaluation.

F.2.3 Analysis Specification.

Data for Analysis: To obtain data for analysis, the team conducts multiple experiments, alternating both the task and inputs, and the target models. For the text continuation task, the team generates test cases by sampling and truncating reviews from the IMDb dataset. The truncated text will serve as the input for the red-team models (4 baselines + CRT), which are expected to add a few words. The red-team model's outputs are then combined with the red team's prompt to produce test cases to the target LLM, which will be GPT-2 with 137M parameters. For the instruction following task, two instruction-tuned models are used as target LLMs: GPT2-alpaca and Dolly-v2-7B. The first was fine-tuned with Alpaca dataset and the second is a pythia model fine-tuned with Databricks dataset. To generate instruction-like test cases using the red-team model, the team randomly samples combinations of instructions from the Alpaca and

[3]This is a rough estimate considering the complexity of the project, team size, among other aspects.

Databricks datasets as the input prompts to the red-team model. Also, they conduct additional experiments using two target instruction-tuned models —Vicuna-7B and GPT-3.5-turbo-instruct. Moreover, previous work indicates that LLaMA2-7b-chat-hf (a fine-tuned LLM with human preferences) generates 0% toxic responses according to the toxicity classifier and prompts used in Hartvigsen et al. (2022). For this reason, experiments with this model as target are conducted.

Finally, for the text-to-image exercise, Stable-Diffusion-2.1 is used as the target model. The red-team model is provided with input prompts from the Stable Diffusion prompt dataset and the team randomly samples combinations. Notably, since the Stable Diffusion model outputs images, RL+TDiv is not applicable as a baseline.

Metric Estimators: In both tasks, text continuation and instruction following, for each method, the authors conduct the experiment using three different random seeds.

F.3 Project Plan

F.3.1 Create Project Plan.

Plan: While Hong et al. (2024) do not report the project stages or timeline, this section provides inferred estimations based on the paper. These estimations are hypothetical and intended to offer a structured approximation of the process. The project will span for 32 weeks and have the following stages. Time is measured in weeks, from W1 to W32. We consider the contributions of the authors reported at the end of the paper.

- Stage 1 (W1-W4): Make the project plan. Define the project objectives and situation assessment which correspond to phases 1.1 and 1.3 of this protocol. Select the baselines, identify the metrics and the target models. Select tasks and datasets for the evaluation. Pre-register the plan.
- Stage 2 (W5-W7): Train the red-team model with the proposed approach of curiosity-driven red teaming (CRT) which is described in phase 2.2 of this protocol.
- Stage 3 (W8-W11): Prepare the baselines in previous and current work to compare the proposed method. Train the RL-based red teaming models (RL, RL+TDiv) and design prompts for zero shot and few shot baselines.
- Stage 4 (W12-W16): Carry out the experiments. This includes the pilot study for experiments involving LLaMA2-7B-Chat-HF with 16 human participants with prior knowledge of toxic content.
- Stage 5 (W17-W19): Analyse the results. The analysis of the results is both qualitative and quantitative (see appendix section B of the paper). As a potential concern is related to the toxicity classifier, assess whether toxicity predictions of the responses elicited by red-teaming approaches in the experiments are close to toxicity predictions of other classifiers. Finally, if the experiment yields meaningful results, consider conducting ablation studies to determine how each individual element that can be affected in the training contributes to the red teaming model's diversity and quality of generated test cases. Additionally, consider conducting further experiments with other target models.
- Stage 6 (W20-W23): Design the remaining elements of the experimental setup for the ablation studies, including the selection of the red teaming models to be compared. Select other target models to conduct further experiments.
- Stage 7 (W24-W26): Conduct the experiments.
- Stage 8 (W26-W28): Analyse the results of the experiments.
- Stage 9 (W29-W32): Review of the evaluation process, final reporting and next steps. This is mostly the writing-up of Hong et al. (2024).

F.3.2 Pre-register Evaluation.

Pre-registration repository: The protocol declaration was not publicly registered after the design of the evaluation. The first publication comes at the end of the whole protocol, in the form of an arXiv paper, as well as GitHub repositories. As this originates from a research paper, no feedback is reported.

G Appendix: Meta-cognitive knowledge to improve LLMs reasoning

The fifth case we will explore is a real-world example in which Didolkar et al. (2024) develop a prompt-guided interaction procedure to improve LLMs reasoning capabilities leveraging their metacognitive knowledge. Specifically, they use a powerful LLM to assign skill labels to math questions, followed by having it perform semantic clustering to obtain coarser families of skill labels. In the testing phase, given a math question, the LLM is provided with contextually relevant, skill-specific examples, which are expected to enhance its effectiveness in answering the question. The authors evaluate its performance compared to other methods that do not incorporate metacognitive knowledge.

We will now describe the **real** case more specifically, and then we will populate the `PREP-Eval` stages from beginning to end.

CASE DESCRIPTION: Consider a language model that, despite demonstrating remarkable advancements in reasoning capabilities, still exhibits significant limitations, particularly in mathematical problem-solving. To assess whether these capabilities can be further improved, Didolkar et al. (2024) develop a methodology that leverages the LLM's own metacognitive knowledge to provide relevant in-context examples.

First, the authors instruct a powerful LLM-A to assign skill labels to each sample in GSM8K and MATH datasets,resulting in a full list of skill labels per dataset. Next, the same LLM, LLM-A performs semantic clustering on the labelled data, grouping examples based on the similarity of their underlying skills, as perceived by the model. This structured repository of skills, representing the LLM's metacognitive knowledge, is then used to enhance in-context math problem-solving.

At inference time, when presented with a new mathematical problem, LLM-A is prompted to identify the most relevant skill category. Then LLM-B is provided with exemplar problems corresponding to that skill as in-context guidance before producing a solution. Results show that this skill-conditioned prompting improves performance over standard baselines, including Chain-of-Thought prompting, and that skill labels generated by a strong model can generalise to weaker models.

G.1 Goals and Objectives

G.1.1 Determine Project Purpose.

Relevant Background: LLMs have demonstrated remarkable advancements in recent years in natural language inference tasks, as well as in scientific and mathematical problem-solving, although their limitations are well-documented. Research in pedagogy shows that enhancing learners' metacognitive knowledge can improve their capabilities; this refers to humans' intuitive understanding of their own thinking and reasoning processes.

As some evidence has emerged regarding the existence of LLM metacognition, Didolkar et al. (2024) investigate whether LLMs possess metacognitive knowledge and whether this can be psycholeveraged to further enhance their capabilities. Specifically, the metacognitive knowledge of interest to the authors is the catalog of skills (from the LLM's perspective) that it applies while solving mathematical problems.

Project Goal: Investigate whether latent metacognitive structure in LLMs can be elicited and operationalised to select skill-relevant in-context examples, and to evaluate whether this improves mathematical reasoning performance.

Project Success Criteria: The evaluation is considered successful if it can reliably determine whether eliciting and using metacognitive knowledge to LLMs improves, degrades or has no effect on their ability to reason through mathematical problems, relative to standard in-context learning and Chain-of-Thought prompting baselines.

G.1.2 Determine Technical Objectives.

Evaluation Target: The models to be evaluated using the proposed method to leverage their metacognitive knowledge are GPT-4-0613, GPT-3.5-Turbo, Mixtral 8×7B and program-aided language models (PALs). Basically, the method works as follows: The model being tested is given a new question with a list of coarse skills and asked to identify the skill needed to solve this new question. Then the model is provided with the previously identified exemplars for the selected skill as in-context examples to guide its problem-solving. The team introduced three metrics to evaluate the effectiveness of their methodology. These metrics are:

- Main Skill Error (Skill Error) which indicates a failure in understanding or applying the primary skill required for a question;
- Secondary Skill Error which denotes errors in comprehending or applying secondary skills necessary for the question;
- Calculation Error which reflects mistakes in the calculation process during question-solving. These error types are not mutually exclusive; a single instance may exhibit multiple error types. Correctly solved instances show none of these errors. GPT-4-0613 classifies each example into these categories. To calculate the metrics, the error rate is first determined for each error type and then derives success rates. These rates indicate how often the model correctly applies main and secondary skills, as well as performs calculations, across various questions. Additionally, they evaluate exact match accuracy—whether the model's generated response exactly matches the expected solution of a mathematical problem. Another technical objective is to evaluate transferability: the generalizability of the skills across different LLMs and unseen datasets.

Target Justification: The team does not provide further justification for their choice of metrics. However, they note that the limitations of LLMs in mathematical problem-solving are well-documented, which could explain their focus on this domain.

Target Estimates Success Criteria: To establish a baseline and isolate the impact of the methodology, four methods are used for comparison: (1) Random: which randomly selects examples from the skills repository in contrast to CoT's fixed examples, highlighting the necessity of skill-aligned example selection; (2) Topic-Based: in this case, examples are grouped by broader mathematical topics (e.g., algebra). This tests whether finer-grained skills offer an advantage over broader topic categorisations; (3) ComplexCOT: which chooses complex in-context examples for CoT, allowing to analyse whether complexity or skill-specificity has a greater impact on performance; (4) Retrieval-RSD, which selects relevant in-context examples for few-shot tasks similar to the proposed approach. They first map the examples to a latent space and then select top-k in-context examples based on cosine similarity to the example. Given these baselines, the authors aim to discern the relative benefits of skill-specificity and complexity in example selection for enhancing LLMs' mathematical reasoning capabilities.

G.1.3 Situation Assessment.

Resources: The team is composed of the authors of the paper Didolkar et al. (2024). They have access to language models such as GPT-4-0613, GPT-3.5-Turbo, and Mixtral 8x7B, as well as various mathematical benchmarks. They also have access to 1 A100L GPU for some transferability experiments.

Requirements and Constraints: The evaluation is performed in four months[4].

Risks: The authors do not report any specific risks or concerns. However, one potential risk is the excessive reliance on LLMs for generating skill labels, clustering abilities, or selecting relevant examples from the repository to include in the prompt when solving a new problem.

Current Understanding of the Evaluated Target: The team has extensive expertise in various prompt techniques designed to enhance LLMs' reasoning abilities, as well as a solid understanding of metacognition. However, they remain uncertain about the extent to which this knowledge can effectively improve LLM performance in math-problem solving.

G.2 Evaluation Design

G.2.1 Identify Potential Evaluation Methods.

Alternatives: The team can explore multiple prompt techniques to integrate with the proposed method. Additionally, several mathematical datasets are available.

G.2.2 Selection of Evaluation Methods.

Method: The team decided to integrate the proposed method for leveraging metacognitive knowledge into two different prompt techniques to conduct the evaluation: (1) Text-based Prompts: chain-of-thought prompting is utilized. This method involves providing step-by-step reasoning in the prompt to guide the model's thought process; and (2) Program-based Prompts: program-aided language models (PALs) are employed, which integrate programming logic within the language model. The team replaces the standard in-context examples used by CoT and PAL with examples from their skill exemplar repository. Then they evaluate the performance of LLMs with both text-based and program-based prompting, using the skill exemplars versus standard examples. The team uses multiple datasets for the evaluation: GSM8K, MATH, SVAMP, ASDIV, and the MAWPS suite (SingleOP, SingleEQ, AddSub, MultiArith). The paper doesn't report any measure to identify or mitigate training data contamination.

G.2.3 Analysis Specification.

Data for analysis: The team applies text-based prompts to GSM8K and MATH datasets while program-based prompts will be applied only in the later. For transferibility experiments, the team use the same skill exemplar repository and skill labels for the MATH dataset, originally generated with GPT-4-0613, but evaluate them with Mixtral 8x7B as the model under test. On the other hand, they investigate the transferability of skills from the GSM8K training dataset to other math word problem datasets, including SVAMP, ASDIV, SingleOP, SingleEQ, AddSub, and MultiArith. In this case they use the pre-clustering skill labels, as these datasets feature finer granularity problems compared to GSM8K, making post-clustering skills less effective.

Metric estimators: The paper does not report any metric estimators.

[4]This is a rough estimate considering the complexity of the project, team size, among other aspects.

G.3 Project Plan

G.3.1 Create Project Plan.

Plan: While Didolkar et al. (2024) do not report the project stages or timeline, this section provides inferred estimations based on the paper. These estimations are hypothetical and intended to offer a structured approximation of the process. The project will span for 16 weeks and have the following stages. Time is measured in weeks, from W1 to W16.

- Stage 1 (W1-W6): Create the project plan. Define the project objectives and situation assessment which correspond to phases 1.1 and 1.3 of this protocol. Identify and select prompt techniques that could improve the language model's reasoning abilities while incorporating the skill-based approach. The remaining elements of the evaluation are defined, including the models to be evaluated, the metrics, the datasets and the baselines to be compared.
- Stage 2 (W7-W8): Generate prompts for all the datasets and baselines.
- Stage 3 (W9-W10): Carry out the experiments.
- Stage 4 (W11-W12): Analyse the results of the experiments and calculate the three metrics descripted in phase 1.2 of this protocol. If the experiment yields meaningful results, consider conducting experiments of transferability between models and datasets.
- Stage 5 (W13-W14): Design the remaining elements of the experimental setup for the transferability studies, including the selection of datasets and models.
- Stage 6 (W15-W16): Analyse the results. Review of the evaluation process, final reporting and next steps. This is mostly the writing-up of Didolkar et al. (2024).

G.3.2 Pre-register Evaluation.

Pre-registration Repository: The protocol declaration was not publicly registered after the design of the evaluation. The first publication comes at the end of the whole protocol in the form of an arXiv paper. As this originates from a research paper, no feedback is reported.

G.4 Data Collection

G.4.1 Experimental Setup, Annotations and Pilots.

Experimental setup: In this step, the authors are expected to verify that the datasets selected for the evaluation are accessible and ready to be used. In addition, for the text-based prompt experiments with GPT-4-0613, they set the decoding temperature to 1.0, while for the transferability experiments they use a temperature of 0.2, without providing a justification for these choices.

Preliminary Analysis of Skill Labelling Models: For the creation of the skill exemplar repository the authors conduct a comparative analysis of GPT- 4-0613, GPT-3.5-Turbo, and Mixtral-8x7B in their proficiency in generating precise skill labels which is contained in Appendix Section 10.4 of the paper. Through experimentation, they observe that the skill labels annotated by GPT-4-0613 lead to the strongest in-context learning performance on the MATH dataset. Therefore, they established GPT-4-0613 as the primary model for skill labelling, clustering, and conducting the majority of the experiments.

G.4.2 Full Data Collection.

Evaluation Data: Text-based Prompts: All experiments are carried out employing 8-shot prompting. Using the GSM8K dataset, the team compares GPT-3.5-Turbo using the Retrieval RSD and CoT + Skill-Based methods. For GPT-4-0613, they evaluate CoT, CoT + Random, CoT + Skill-Based, and CoT + Skill-Based (maj@5). Additionally, using the MATH dataset—with its training set for the skill repository— standard CoT, Complex CoT, and the Topic-Based approach are assessed. Program-based Prompts: In the integration of Skill-Based prompting with PALs, the team modify the in-context example structure: they use three non-code-based examples from the skill exemplar repository based on skill matching, followed by one fixed code-based example, totalling four in-context examples. This modified approach was tested on a subset of 500 examples from the MATH dataset. They compare PAL (4-shot) with PAL + Skill-Based (3 Skill-Based + 1 Code-Based) and PAL + Skill-Based (7 Skill-Based + 1 Code-Based). Transferability experiments to other models: Here, the team compares the Skill-Based approach against two baselines: Chain-of-Thought with self-consistency and the Topic-Based approach. Transferability experiments to other datasets: The Skill-based approach is applied with CoT to various datasets: SVAMP, ASDIV, SingleOP, SingleEQ, AddSub, and MultiArith. They employ as baselines a CoT-based method with 4-shot prompting and greedy decoding, a PAL-based approach and a hybrid CoT + PAL approach where the model outputs both CoT and PAL solutions and selects the most accurate.

G.4.3 Data Preparation.

Data Readiness: The team does not describe any specific data preparation process.

G.5 Data Analysis

G.5.1 Initial Data Exploration.

Exploratory Analysis: The paper does not report any exploratory analysis.

Feature Relevance: When using the MATH dataset, the variation in results across several topics are examined, including Pre-Algebra, Algebra, Intermediate Algebra, Geometry, Number Theory, Precalculus, and Probability. Additionally, they analyse which skills benefit the most from the proposed approach by comparing the per-skill accuracy of the Skill-Based approach against the Random baseline. They find that the proposed approach outperforms the Random baseline in 11 out of 18 skills.

G.5.2 Conduct Planned Analyses.

Summary statistics: Text-based Prompts: The Skill-based approach surpasses multiple other methods on the GSM8K dataset across different models. Specifically, it achieves 82.03% with GPT-3.5-Turbo and 94.31% with GPT-4-0613 (95.38% with maj@5). In comparison, Retrieval SRD with GPT-3.5-Turbo attains 76.8%, while Chain of Thought (CoT) and CoT + Random with GPT-4-0613 reach 93.0% and 92.87%, respectively. In the MATH dataset, the Skill-based approach achieves a notable improvement in performance, surpassing the standard Chain-of-Thought (CoT) by 11.6%. It also outperforms Complex CoT by 3.5% and the Topic-Based approach by 3.5%. Program-based Prompts: Despite only one code-based in-context example (compared to PAL's four), the Skill-based approach shows a 7.52% improvement over PAL. Transferability experiments to other models: The Skill-Based approach surpasses both the Topic-Based and CoT approaches. Notably, the Skill-Based approach, even without self-consistency, matches the performance of CoT with SC, highlighting its efficacy in extracting correct reasoning paths and concepts. Furthermore, when combined with self-consistency, the Skill-based approach shows a remarkable 4.0% improvement over CoT with SC. Transferability experiments to other datasets: The Skill-Based approach consistently outperforms the base CoT and PAL across all

datasets. Additionally, it surpasses CoT + PAL in four out of six datasets, achieving 97.86% on SingleOP, 99.01% on SingleEQ, 96.71% on AddSub, and 94.03% on ASDIV. However, CoT + PAL performs better on SVAMP (93.7%) and MultiArith (98.17%).

G.5.3 Assess and Refine the Analysis.

Ablation Metrics: The authors compare the performance of the Skill-based approach with the Topic-Based baseline across the three metrics: (1) Skill Success Rate; (2) Secondary Skill Success Rate; and (3) Calculation Success Rate. They observe that the Skill-based approach results in a higher Skill Success Rate which means that the model is using the correct skill more frequently in the proposed approach as compared to the Topic-Based baseline. Furthermore, they find that the proposed approach is also quite effective in reducing secondary skill errors and calculation errors. Thus, showing the overall superiority of the proposed approach.

Uncertainty Quantification: We do not have information about the standard deviation or any other uncertainty metric of the summary statistics. There are some considerations in the appendix of the paper, but not to the point of calculating confidence intervals or standard errors.

G.6 Conclusions & Review

G.6.1 Draw Conclusions.

Findings: The team finds that LLM problem-solving improves when using skill labels and skill exemplars provided by an LLM on the same dataset. This provision of skill exemplars can be seen as a new addition to known prompting methods. Using a strong LLM—GPT-4—to identify skills, they validate the usefulness of these skills by demonstrating a significant 11.6% improvement over CoT on the MATH dataset. Furthermore, the identified skills enhance the generation of code-based solutions for problems within the MATH dataset, yielding a 7.52% improvement over the baseline PAL approach, which also instructs the model to generate code. They also show that the skill-exemplar repository created for MATH noticeably improves in-context performance for weaker LLMs on the same dataset and that the repository for GSM8K helps improve in-context performance for other math datasets. This demonstrates that a powerful LLM can facilitate a deeper understanding of skills that translates across other LLMs and related datasets. Several instances were found where the model, despite using a skill-based approach, fails to produce correct answers (Appendix Section 10.6). These examples suggest that while the Skill-Based approach effectively guides the application of the main skill required for a question, it may falter in the application of secondary skills or in the comprehension of specific question properties.

Limitations: The paper does not identify any limitations of the evaluation process.

Goal achievement: The evaluation successfully examined whether leveraging metacognitive knowledge to generate relevant in-context examples enables LLMs to strengthen their reasoning and improve their mathematical problem-solving performance.

G.6.2 Review Evaluation Process. The paper does not explicitly report any lessons learned, recommended changes or legacy in terms of the evaluation process.

G.6.3 Determine Next Steps.

Next steps: Since the method appears general enough to be widely applicable to various problem-solving tasks, future evaluations should explore this potential. Additionally, while this paper primarily focuses on in-context learning, the authors aim to extend these methodologies to enhance model performance through fine-tuning, which will also need to be evaluated.

G.6.4 Complete the Pre-registration.

Report: The evaluation procedure as well as the result were published in Didolkar et al. (2024), not following `PREP-Eval`, as the protocol wasn't available yet.

H Appendix: Face recognition system comparison

The last case we explore is a traditional evaluation problem: given a task, compare one or more existing methods with a new method to determine whether the new method is better. This is not simply performing a statistical test on some datasets and seeing whether the new method is better than the rest (especially if these datasets were known by the developers of the new method), but it may require many other considerations. But even if well-though for a real scenario, these kinds of evaluations are relatively bounded and goal-oriented.

We will now describe the **fictitious** case more specifically, and then we will populate the `PREP-Eval` stages as they would be completed before the pre-registration and continuing with stage 5.

CASE DESCRIPTION: Consider a network of cameras at airports and train stations in a country, continuously recording the transit of people at several locations. These recordings can be inspected a posteriori, with a judge warrant, whenever there is an important security incident, such as a terrorist threat. In these cases, the police have been using a face recognition system, which we will call system A, that takes as input several photos of one or more people and goes through the recordings to find possible matches to be suggested. With the suggested matches, the police team checks the video scenes to conduct the investigation. This is the current situation. Recent research in AI, however, has introduced new developments in face recognition technology, and the authorities are considering replacing system A by a new system B based on these new research developments. Before making any change the organisation wants to determine that system B is better according to a series of parameters such as false positives and false negatives, robustness to noise and fairness for protected groups in terms of the error rates and their balance.

H.1 Goals and Objectives

H.1.1 Determine Project Purpose.

Relevant Background: Face recognition for security is a highly-sensitive area. Its use in situations such as stations or airports, with many people in motion, introduces a series of sensitive elements, apart from noise and other conditions that affect the quality of identification. The evaluation project will have to consider not only the location and features of the cameras, but the populations that transit in these spaces and their privacy.

Project Goal: "Determine if the new face recognition system B is better than system A for its application at airports and train stations".

Project Success Criteria: The evaluation will be able to determine for all considered dimensions if the new technique is significantly better or worse, or no significant difference is found. The findings have to be summarised in a clear report for the decision makers so that they can make an informed decision. The decision of whether to replace one

method by the other is not part of this evaluation, and the evaluation should not be biassed by any interest in the old or new system.

H.1.2 Determine Technical Objectives.

Evaluation Target: "Produce performance metrics for facial recognition of given images against recorded videos, sliced by protected groups and different conditions (light, moving targets, etc.) for system A and system B. From them, determine the areas where one system is significantly better than the other. Target metrics: specificity, sensitivity, accuracy, balanced accuracy and F-score. Protected groups: race, gender, age".

Target Justification: Depending on the particular incident, false positives or false negatives may be more or less critical. The choice of protected groups for the faces to be recognised is limited by the categories that are available, and those that may influence identification to a great extent, especially if the population pool in the country (locals and visitors) significantly differ from the usual distributions that may have been used in the training of the face recognition systems.

Target Estimates Success Criteria: Samples must be sufficient to estimate the target metrics to an error of less than 1%, including the break-outs.

H.1.3 Situation Assessment.

Resources: There will be a team of three people, two with strong technical expertise in vision systems and machine learning in general, and an ethicist with good knowledge of fairness metrics, domain knowledge (suspect identification) and applicable regulations. All of them will have experience in AI evaluation. There will be access to clusters and sufficient storage for running all the experiments. The budget to visit the original locations and perform controlled tests will be considered as well.

Requirements and Constraints: The evaluation must be performed in 18 weeks, with a budget of $100,000.

Risks: Access to sufficient evaluation data must be granted before the start of the project. Difficulties in enumerating all protected groups and conditions for evaluation.

Current Understanding of the Evaluated Target: the team knows the technology behind system A, which has been used for the past two years, for which there's technical documentation, including information about possible adversarial attacks. Technology B is new, and less well-known by the team. This will require some extra tests with the new technology according to the provided implementation, and results on some standard benchmarks before performing the comparison.

H.2 Evaluation Design

H.2.1 Identify Potential Evaluation Methods.

Alternatives: The data can be obtained from standard benchmarks or can be obtained from the recordings in a representative sample of airports and stations, using some sources of faces that were labelled manually to be present/non-present.

H.2.2 Selection of Evaluation Methods.

Methods: For the choice/creation of datasets and labelling methods, external benchmarks may be contaminated and non-representative for the context in which the systems must operate. Because of this, no external benchmarks will be used other than for familiarisation and early testing of the implementations of the methods. The data for the evaluation will be obtained from the recordings of last year in a representative random sample of airports and stations, consisting of 1,000 hours of video in snapshots each second (360,000 snapshots). The faces to be identified will be chosen from people who are known to be transiting, being labelled manually to be "present" if they are found in some of the videos. Other faces from other sources, not supposed to be at the airport, will be chosen. A total of 1,500 faces will be included, and for each combination there will be a label. In both sources for faces, informed consent for the experiment will be obtained from all these people depending on the country's regulations. The face datasets to be recognised will cover a balanced proportion in gender (male, female, non-binary), ages (five intervals) and races (ten categories). These proportions will be set beforehand and after the labelling the proportions for positives and negatives will be validated, or corrected. In other words, with these features the team will have 3x5x10= 150 combinations, and they will check there are exactly 10 faces in each of them.

H.2.3 Analysis Specification.

Data for analysis: The two methods A and B will be applied to the same dataset with their by-default parameters. No further optimisation of parameters will be allowed. The team will get a present/absent classification for each snapshot and face for both models, to be compared with the label. In this evaluation project, they will not consider the probabilities from the models.

Metric estimators: For each method, the team will calculate the full confusion matrices, including specificity, sensitivity, accuracy, balanced accuracy and F-measure. They will calculate all these metrics for the whole dataset and by all cells of the hypercube, defined by the variations in gender, age and race. For each of them they will calculate the standard error as a measure of uncertainty in the evaluation, given the population sizes.

H.3 Project Plan

H.3.1 Create Project Plan.

Plan: The project will span for 18 weeks and have the following stages. Time is measured in weeks, from W1 to W18. The two technical people are denoted by T1 and T2, while the ethicist is denoted by E1.

- Stage 1 (W1 : T1,T2,E1): Writing this protocol for pre-registration.
- Stage 2 (W2-W3 : E1): Distribution of the pre-registration to experts and stakeholders and receiving feedback.
- Stage 3a (W4-W10 : T1,E1 + IT/legal people at the org) : Data collection by requesting the data from the data centres of the organisation, the photos, authorisations, etc.
- Stage 3b (W4-W10 : T2 + hardware + IT support): Installation of both methods and implementation of all metrics and evaluation on standard benchmarks.
- Stage 4 (W11-W13 : T1,T2,E1 + hardware + IT support): Data analysis. Running both methods on the dataset and obtaining the metric estimators, assessment of quality and rectifications if necessary.
- Stage 5 (W14 : E1) : First (internal) presentation of results.
- Stage 6a (W15-W16 : T1,T2,E1): Review of the evaluation process.
- Stage 6b (W17-W18, T1,T2,E1): Final reporting and next steps.

H.3.2 Pre-register Evaluation.

Pre-registration repository: The protocol declaration will be publicly registered at the transparency portal of the organisation, sharing it with several stakeholders and experts including the police, the transportation authorities, civil rights organisations, experts in face recognition, etc., giving some consultation time for receiving feedback.