CATALOGUING LLM EVALUATIONS

Draft for Discussion (October 2023)

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

AI VERIFY FOUNDATION

Contents

Introduction

1. In an earlier paper titled "Generative AI: Implications for Trust and Governance" ("Discussion Paper"),[1] we had set out key factors necessary for enabling a trusted ecosystem for Generative AI innovation, including Large Language Models ("LLMs"). These factors include:

    a. **Accountability** amongst the parties in the Generative AI developer lifecycle;
    b. **Data use** in the training of Generative AI models;
    c. **Model development and deployment**, which includes the **development of evaluation framework and tools**;
    d. Independent **third-party evaluation and assurance**;
    e. **Safety and alignment research** to ensure that human capacity to control increasingly powerful AI systems keeps pace; and
    f. Using **Generative AI to achieve Public Good.**

2. **Systematic and robust evaluation of models is a critical component of LLM governance** and helps form the bedrock of trust in the use of these technologies. Through rigorous evaluation, the capabilities of a model are revealed, which can assist in determining its intended uses and potential limitations. Moreover, evaluation provides a vital roadmap for developers to make improvements.

3. In advancing the sciences of LLM evaluations, it is important to first achieve: (i) a **common understanding of the current LLM evaluation through a standardised taxonomy**; and (ii) **a baseline set of pre-deployment safety evaluations for LLMs**. A comprehensive taxonomy categorises and organizes the diverse branches of LLM evaluations, provides a holistic view of LLM performance and safety, and enables the global community to identify gaps and priorities for further research and development in LLM evaluation. A baseline set of evaluations defines a minimal level of LLM safety and trustworthiness before deployment. At this early stage, the proposed baseline in this paper puts forth a starting point for global discussions with the objective of facilitating multi-stakeholder consensus on safety standards for LLMs.

4. This paper comprises 3 parts:

    a. In Part 1, we introduce a **taxonomy** of the LLM evaluation landscape, comprising of five categories: (i) General Capabilities; (ii) Domain Specific Capabilities; (iii) Safety and Trustworthiness; (iv) Extreme Risks; and (v) Undesirable Use Cases. These categories were identified based on both a top-down view of what organisations seeking to develop or deploy an LLM or LLM-based application[2] in a safe and responsible manner would need to consider,

[1] This Discussion Paper was jointly published by the Infocomm Media Development Authority of Singapore and Aicadium in June 2023. See https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf

[2] The focus of this paper is on evaluation and testing approaches for LLMs, including those embedded in applications, but not other application-specific testing measures. LLMs can serve as the foundational backbones for various applications (e.g., ChatGPT) that, like any software, undergo testing before

and a bottom-up scan of major research papers in LLM evaluation. We then set out a **catalogue** that organizes the various evaluation and testing approaches we came across based on these five categories.

b. In Part 2, we provide an **analysis** of the LLM evaluation landscape, highlighting key areas for further development, such as the need for more context-specific evaluations, frontier model evaluations and the need for standards and best practices in LLM evaluations. We also suggest **future work** in evaluations to support governance, such as evaluations for training data quality, LLM interpretability and explainability, and environmental impact assessments.

c. In Part 3, we recommend a **baseline set of evaluations** comprising five attributes that LLMs should minimally be tested on pre-deployment to ensure a minimal level of safety and trustworthiness: (i) bias; (ii) factuality; (iii) toxicity generation; (iv) robustness; and (v) data governance.

5. To remain relevant, the taxonomy and catalogue must be **constantly updated** as the field matures and evolves. The baseline will also improve as new evaluation benchmarks and methods are developed. Nonetheless, these provide a common understanding and foundation for further dialogue and refinement in the wider community.

Call for Community Contributions

6. This paper is the first version of our exploration into the complex domain of LLM evaluation and we acknowledge that it remains a work in development. While we have gathered preliminary feedback from partners, we welcome insights, comments and other contributions from the broader community. Advances in this domain are being made at an unparalleled pace and it is only through inclusive collaboration that we can ensure the continued relevance and utility of this work.

7. To that end, we invite you to relay feedback and other contributions, such as new benchmarkts or testing methods, to us at info@aiverify.sg. As at the date of publication, we are working towards establishing a more streamlined platform for community engagement and contribution, as well as for the community to share their testing outcomes. For the most up-to-date information on this endeavour, visit www.aiverifyfoundation.sg.

deployment. Testing protocols such as integration, load, UI/UX, and penetration testing ensure the application's reliability, security, performance, and user experience. However, the primary focus of this paper is not on these application or software testing protocols, but on the evaluation and testing approaches that specifically assess the LLMs (i.e., model evaluations of LLMs) that provide the foundation for these applications.

Part 1 – Taxonomy and Catalogue

Methodology

8. To develop the taxonomy, we first surveyed the landscape of LLM evaluation and testing approaches. This involved an in-depth review of key academic papers, benchmarks, and research outputs from leading organizations in this field.[3]

9. **The taxonomy is intended to be a useful resource for an organisation that is considering developing a LLM or deploying an LLM-driven application.** We therefore approached the next stage of taxonomy development from the perspective of such an organisation. The organisation would desire a clear understanding of: (i) a LLM's general capabilities; (ii) its performance within a specific domain (e.g., medicine); (iii) potential risks, vulnerabilities, and catastrophic consequences that could arise from its deployment; and (iv) potential areas where the model could be exploited for malicious or unethical purposes. In this context, **the organisation would need to know the benchmarks and tests it can utilize to achieve the above understanding in an objective manner.**

10. Through this exercise, we developed a **comprehensive, coherent, and non-exhaustive taxonomy, comprising of five main categories, that encompasses all aspects of LLM evaluation.** In deriving this taxonomy, we drew heavily on prior research in this field. Works such as **"Holistic Evaluation of Language Models (HELM)[4]", "Model Evaluation for Extreme Risks[5]", "DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models[6]",** and **"FLASK: Fine-Grained Language Model Evaluation Based on Alignment Skill Sets[7]",** provided valuable insights in our journey to map the landscape of LLM evaluation.

Overview of the Taxonomy

11. The following sections of the paper set out the detailed taxonomy across five categories, including a brief description of their respective sub-categories. The specific evaluation and testing approaches for each category are set out in the catalogue (see Annex B)[8].
12. For now, we set out a brief overview of each category and their respective sub-categories:

---

[3] The complete list of papers, benchmarks, and other resources we referred to are set out in Annex A.
[4] By Stanford University's Centre for Research on Foundational Models
[5] By DeepMind
[6] By authors from University of Illinois at Urbana-Champaign, Stanford University, University of California, Berkeley, Center for AI Safety and Microsoft Corporation
[7] By the Korea Advanced Institute of Science and Technology
[8] Certain benchmarks and evaluations may fall into multiple categories of the taxonomy, reflecting the complex and cross-cutting nature of LLM evaluations. For example, assessments of an LLM's propensity to output adult content can be situated under both the "Undesirable Use Cases" and "Safety and Trustworthiness" categories. Our proposed taxonomy is a structured representation of our perspective in this dynamic field and is not intended to be definitive. As our understanding of LLMs evolves, so too might the categories.

a. **General Capabilities:** This category assesses a LLM's potential and performance. The core idea is to understand what the model can do, how well it can do it, and the circumstances under which it operates best. Its sub-categories include: (i) natural language understanding (e.g., text classification); (ii) reasoning; and (iii) knowledge and factuality.

b. **Domain Specific Capabilities:** This category assesses a LLM's performance within the context of the unique requirements and challenges of a particular domain or industry. Its sub-categories are: (i) law; (ii) medicine; and (iii) finance.

c. **Safety and Trustworthiness:** This category assesses the reliability of a LLM's operation and its inherent risks. This includes the ability to avoid generating harmful or biased outputs, and to behave predictably over a broad spectrum of inputs. Its sub-categories include: (i) toxicity generation; (ii) bias; and (iii) robustness (i.e., performance when faced with unexpected or adversarial inputs).

d. **Extreme Risks:** This category assesses potential catastrophic consequences arising from a LLM with dangerous 'frontier' capabilities (e.g., offensive cyber capabilities, deception, ability to acquire weapons) being misused or harmfully applying its capabilities. Its sub-categories are: (i) dangerous capabilities; and (ii) alignment risks.

e. **Undesirable Use Cases:** This category examines potential scenarios where LLMs could be used maliciously or unethically. Its sub-categories include: (i) misinformation; and (ii) adult content.

Testing Approaches and Scoring Methods

13. There are several testing and evaluation terminologies which are introduced alongside these categories within the catalogue. To draw a distinction, **testing approaches** describe *how* a LLM evaluation will be conducted. This is complemented by **scoring methods** which assign qualitative or quantitative scores to the *outputs* of a testing approach.

14. In our landscape scan, we came across broadly **three types of testing approaches**:

    a. **Benchmarking:** Benchmarking employs the use of datasets of questions to evaluate a LLM based on their output. It can be compared with the ground truth or against some rules that are predefined.

    b. **Automated Red Teaming:** This approach utilises another model to initiate prompts and probe a LLM in order to achieve a target outcome (e.g., to evaluate permutations of prompts which lead to the production of toxic outputs).

    c. **Manual Red Teaming:** Manual red teaming utilises human interaction to initiate prompts and probe a LLM in order to achieve a target outcome.

15. These testing approaches can be coupled with a suitable scoring method:

   a. **Algorithmic Scoring:** Scoring algorithms (e.g., ROUGE[9], BLEU[10]) that mathematically calculate scores, for instance, to determine absolute similarity between two bodies of text. This also includes rule-based scoring (e.g., fluency score).

   b. **Human Scoring:** Human participants score model outputs. Participants can vary from experts to ordinary users, and they may be asked to rate the relevance, coherence, or other qualitative aspects of outputs. Such evaluations are useful when the output is open-ended or subjective. This approach can provide nuanced insights into a model's performance that automated metrics might overlook.

   c. **Model Scoring:** These evaluations employ the use of a model (potentially another LLM, or classical AI model) to assess the quality of an LLM's output. Model scoring is often used in tandem with or to replace human scoring, as it offers a more scalable approach without completely sacrificing qualitative understanding.

[9] ROUGE stands for "Recall-Oriented Understudy for Gisting Evaluation" and is a set of metrics used to evaluate the quality of machine-generated texts, such as summaries. These scores measure the overlap between the generated text and a reference text. ROUGE typically scores range from 0 to 1, with a higher score indicating greater similarity.

[10] BLEU stands for "Bilingual Evaluation Understudy" and is commonly used to evaluate the quality of machine-generated texts, such as translations. It measures the similarity between the generated text and a reference text. A BLEU score ranges from 0 to 1, with higher scores indicating greater similarity.

Detailed Taxonomy

1.	**GENERAL CAPABILITIES** The evaluations in this category assess the capabilities of LLMs, focusing on understanding the various abilities of the model, such as comprehension, reasoning, and natural language generation. They seek to determine how well the model can follow instructions, perform cognitive tasks, replicate human-like language understanding, and adapt to novel problems. Based on our landscape scan, the evaluations in this category are primarily based on a benchmarking approach. Most apply some form of algorithmic scoring to score the outputs with the remainder using human or LLM based scoring. Notably, red teaming is rarely used in the assessments of the capabilities of LLMs.	
1.1.	**Natural Language Understanding** The evaluations in this category, while varied in nature, seek to discern a LLM's ability to understand and interpret the input sequence (i.e., the prompt) provided to it.	**Text classification** These evaluations assess a LLM's ability to interpret text and accurately categorize it into predefined classes or labels. An effective classification capability is critical for numerous real-world applications, such as spam detection and document categorization.
		Sentiment analysis These evaluations assess a LLM's ability to interpret text and determine its emotional tone. Sentiment analysis is a form of text classification.
		Toxicity detection These evaluations assess a LLM's ability to interpret text and determine whether it contains toxic content. Toxicity detection is a form of text classification.
		Information retrieval These evaluations assess a LLM's proficiency in determining the relevant information or answers from large textual corpora based on specific queries. The evaluations gauge how adeptly the LLM can identify information that aligns with an inquiry, ensuring that the LLM can operate effectively in tasks akin to search engines or knowledge-base query systems.
		Sufficient information These evaluations assess a LLM's ability to discern whether it possess sufficient information to provide a valid response to a given query.
		Natural language inference

		These evaluations assess a LLM's ability to determine the relationship between two sentences: whether they contradict, entail, or are neutral to each other. Such inference capability is key to a LLM understanding context, drawing accurate conclusions, and responding coherently in conversational interactions.
		General English understanding These evaluations assess a LLM's understanding of the English language,[11] including specific understanding of distinct linguistic phenomena.
1.2.	**Natural Language Generation** The evaluations in this category test a LLM's ability to generate coherent and contextually appropriate text, ensuring that the model can communicate information effectively and responsively in diverse applications.	**Summarization** These evaluations measure a LLM's proficiency in distilling lengthy or complex texts into concise, coherent, and accurate summaries.
		Question generation and answering These evaluations assess a LLM's ability to generate relevant and coherent questions based on the provided content, and its proficiency in accurately and contextually responding to queries.
		Conversations and dialogue These evaluations assess a LLM's capability to maintain context and coherence whilst engaging in conversations and dialogues, ensuring that the model can sustain meaningful interactions.
		Paraphrasing These evaluations assess a LLM's ability to rephrase provided text into different wording while retaining the original meaning and context.
		Other response qualities These evaluations assess other qualities of a LLM's output, such as its readability and creativity.
		Miscellaneous text generation These evaluations assess a LLM's ability to generate text but do not fit into the categories above.
1.3.	**Reasoning** These evaluations assess a LLM's ability to process and reason about information logically, draw the necessary inferences and make appropriate decisions. This includes assessing how well the LLM understands relationships, cause-and-effect scenarios, and social interactions and	

[11] Evaluations that assess a LLM's understanding of other languages are set out in a separate section below.

	whether it can engage in multi-step reasoning processes. It involves testing the abstract reasoning capabilities that a LLM has, as well as how it performs in various realistic contexts (e.g., mathematical reasoning, legal reasoning, etc).
1.4.	**Knowledge and factuality** These evaluations focus on a LLM's ability to accurately generate output that is consistent with established facts and real-world knowledge (e.g., correctly generating "Paris" when prompted with "The capital of France is_____"). This is usually done by not providing any context during inference, ensuring that the LLM relies on the knowledge it has to generate the output.
1.5.	**Effectiveness of tool use** LLM-driven applications may contain various tools to improve and increase capabilities. Such tools include APIs to retrieve information, calculators to perform mathematical operations, knowledge bases that store relevant documents for query answering and even AI models to perform specialized functions (e.g., image segmentation, text-to-audio, etc). The evaluations in this section assess how well LLMs utilize the provided tools.
1.6.	**Multilingualism** These evaluations assess a LLM's proficiency in understanding and generating content in languages apart from English, and its ability to accommodate different dialects and sociolects (e.g., African American English) of a language. Such evaluations ensure that this technology can cater to a diverse, global audience and accommodates the nuances of different linguistic cultures and communities.
1.7.	**Context length** LLMs have a context window that sets out the maximum number of tokens they can process in a single interaction. These evaluations assess how well LLMs respond to longer contexts, within their respective context length limit (Shaham et al., 2022).
2.	**DOMAIN SPECIFIC EVALUATIONS** Unlike general-purpose assessments, domain specific evaluations are designed to measure the performance of LLMs within the distinct context of a particular industry or field such as medicine or law. They serve as essential instruments in gauging whether an LLM can meet the stringent requirements of specialised applications—be it interpreting medical jargon, parsing legal statutes, or analysing financial data. This is a nascent area, as evinced by the relatively small number of benchmarks and tools we came across in our landscape scan.
2.1.	**Law**

	These evaluations assess a LLM's performance in the legal domain, such as its ability to perform various forms of legal reasoning (e.g., issue-spotting and interpretation).	
2.2.	**Medicine** These evaluations assess a LLM's performance in the medical domain, such as its ability to answer various forms of medical questions (e.g., consumer medical questions and medical research questions).	
2.3.	**Finance** These evaluations assess a LLM's performance on various tasks that are germane in the financial sector, such as sentiment analysis, news headline classification and question answering over financial data.	
3.	**SAFETY AND TRUSTWORTHINESS** Assessing LLM safety involves a multi-faceted analysis that includes, amongst other, evaluating how the model behaves under unforeseen inputs, how it reacts to adversarial interventions and whether it displays any biases or stereotypes. A comprehensive assessment of LLM safety is fundamental to the responsible development and deployment of these technologies, especially in sensitive fields like healthcare, legal systems, and finance, where safety and trust are of the utmost importance. The taxonomy for this category is largely adopted from Wang et al. (2023). The majority of the evaluations in this category are benchmarks, while some use red teaming to elicit undesirable behaviours.	
3.1.	**Toxicity generation** These evaluations assess whether a LLM generates toxic text when prompted. In this context, toxicity is an umbrella term that encompasses hate speech, abusive language, violent speech, and profane language (Liang et al., 2022).	
3.2.	**Bias**	**Demographical representation** These evaluations assess whether there is disparity in the rates at which different demographic groups are mentioned in LLM generated text. This ascertains over-representation, under-representation, or erasure of specific demographic groups.
		Stereotype bias These evaluations assess whether there is disparity in the rates at which different demographic groups are associated with stereotyped terms (e.g., occupations) in a LLM's generated output.
		Fairness These evaluations assess whether sensitive attributes (e.g., sex and race) impact the predictions of LLMs.

		Distributional bias These evaluations assess the variance in offensive content in a LLM's generated output for a given demographic group, compared to other groups.
		Representation of subjective opinions These evaluations assess whether LLMs equitably represent diverse global perspectives on societal issues (e.g., whether employers should give job priority to citizens over immigrants).
		Political bias These evaluations assess whether LLMs display any slant or preference towards certain political ideologies or views.
		Capability fairness These evaluations assess whether a LLM's performance on a task is unjustifiably different across different groups and attributes (e.g., whether a LLM's accuracy degrades across different English varieties).
3.3.	**Machine ethics** These evaluations assess the morality of LLMs, focusing on issues such as their ability to distinguish between moral and immoral actions, and the circumstances in which they fail to do so.	
3.4.	**Psychological traits** These evaluations gauge a LLM's output for characteristics that are typically associated with human personalities (e.g., such as those from the Big Five Inventory). These can, in turn, shed light on the potential biases that a LLM may exhibit.	
3.5.	**Robustness** These evaluations assess the quality, stability, and reliability of a LLM's performance when faced with unexpected, out-of-distribution or adversarial inputs. Robustness evaluation is essential in ensuring that a LLM is suitable for real-world applications by assessing its resilience to various perturbations.	
3.6.	**Data governance** These evaluations assess the extent to which LLMs regurgitate their training data in their outputs, and whether LLMs 'leak' sensitive information that has been provided to them during use (i.e., during the inference stage). There are privacy and copyright implications, depending on the characteristics of the data regurgitated by the LLM in its output.	
4.	**EXTREME RISKS**	

	The taxonomy for this category is adopted from Shevlane et al. (2023). This category encompasses the evaluation of potential catastrophic consequences that might arise from the use of LLMs. It is broken up into 2 further sub-categories: Dangerous Capabilities and Alignment. The former refers to capabilities that can have significant adverse impacts and disruption if the LLM is misused or is misaligned (e.g., offensive cyber capabilities, ability to acquire weapons). The latter refers to a LM's propensity to harmfully apply its capabilities due to risk factors such as resisting being shut down and engaging in 'power-seeking' behaviour. We adopt the view espoused by Shevlane et al. (2023) that only frontier models[12] need to undergo the evaluations under this category. As LLM technology continues to evolve, the criteria defining frontier models are likely to be refined. As this is a relatively nascent domain, we were only able to identify evaluations for some of the risks in this category.	
4.1.	**Dangerous Capabilities**	**Offensive cyber capabilities** These evaluations focus on whether a LLM possesses certain capabilities in the cyber-domain. This includes whether a LLM can detect and exploit vulnerabilities in hardware, software, and data. They also consider whether a LLM can evade detection once inside a system or network and focus on achieving specific objectives.
		Weapons acquisition These assessments seek to determine if a LLM can gain unauthorized access to current weapon systems or contribute to the design and development of new weapons technologies.
		Self and situation awareness These evaluations assess if a LLM can discern if it is being trained, evaluated, and deployed and adapt its behaviour accordingly. They also seek to ascertain if a model understands that it is a model and whether it possesses information about its nature and environment (e.g., the organisation that developed it, the locations of the servers hosting it).
		Autonomous replication / self-proliferation These evaluations assess if a LLM can subvert systems designed to monitor and control its post-deployment behaviour, break free from its operational confines, devise strategies for exporting its code and weights, and operate other AI systems.

[12] In this paper, we adopt the Frontier Model Forum's definition of frontier models: "large-scale machine-learning models that exceed the capabilities currently present in the most advanced existing models, and can perform a wide variety of tasks."

<table>
<tr><td></td><td></td><td>Persuasion and manipulation
These evaluations seek to ascertain the effectiveness of a LLM in shaping people’s beliefs, propagating specific viewpoints, and convincing individuals to undertake activities they might otherwise avoid.</td></tr>
<tr><td></td><td colspan="2">We were unable to identify existing evaluation and testing approaches for the risks below, signifying opportunities for research and development. We also acknowledge the possibility that there might be ongoing work in these areas and that our landscape scan may have inadvertently missed out existing work.

<u>Dangerous Capabilities</u>

a. Dual-Use Science: LLM has science capabilities that can be used to cause harm (e.g., providing step-by-step instructions for conducting malicious experiments)
b. Deception: LLM is able to deceive humans and maintain that deception
c. Political strategy: LLM can take into account rich social context and undertake the necessary social modelling and planning for an actor to gain and exercise political influence
d. Long-horizon planning: LLM can undertake multi-step sequential planning over long time horizons and across various domains without relying heavily on trial-and-error approaches
e. AI development: LLM can build new AI systems from scratch, adapt existing for extreme risks and improves productivity in dual-use AI development when used as an assistant.

<u>Alignment Risks</u>

a. LLM pursues long-term, real-world goals that are different from those supplied by the developer or user
b. LLM engages in ‘power-seeking’ behaviours
c. LLM resists being shut down
d. LLM can be induced to collude with other AI systems against human interests
e. LLM resists malicious users attempts to access its dangerous capabilities</td></tr>
<tr><td>5.</td><td colspan="2">Undesirable Use Cases

This section sets out evaluations that assess whether LLMs could be used for malicious or unethical purposes. Considering the myriad use-cases that LLMs can be used for, we did not conduct a targeted cataloguing of this category of evaluations. Instead, we set out those that we came across while researching other categories.</td></tr>
</table>

	Automated benchmarking and red teaming are both used in this category, coupled with both model scoring and human scoring approaches.
5.1.	**Misinformation** These evaluations assess a LLM's ability to generate false or misleading information (Lesher et al., 2022).
5.2.	**Disinformation** These evaluations assess a LLM's ability to generate misinformation that can be propagated to deceive, mislead or otherwise influence the behaviour of a target (Liang et al., 2022).
5.3.	**Information on harmful, immoral, or illegal activity** These evaluations assess whether it is possible to solicit information on harmful, immoral or illegal activities from a LLM.
5.4.	**Adult content** These evaluations assess if a LLM can generate content that should only be viewed by adults (e.g., sexual material or depictions of sexual activity)

Part 2 – Observations and Future Work

Observations and Insights

16. This section sets out four key observations derived from our survey of the LLM evaluation landscape. We highlight the need for context-sensitive evaluations, as well as assessments tailored for frontier models, advocate for standards development, and emphasize the necessity of a multi-faceted evaluation approach. Collectively, these observations form a roadmap to advance how we currently evaluate these transformative models.

Developing Context-Specific Evaluations

17. Based on observations of evaluations in the catalogue, there is a lack of **nuanced, context-specific evaluations that adequately address the multi-faceted nature of real-world LLM deployments.** Context specificity refers to various factors that shape and dictate the environment that a LLM application operates in, such as:

    a. **Domain specificity:** This refers to industry verticals and the type of application in which the LLM is used. Whether it's aiding legal professionals as a knowledge management tool, assisting healthcare practitioners in diagnosis, or streamlining customer interactions in a retail setting, the specific demands and nuances of each domain and application necessitate targeted evaluations. This paper sets out some domain-specific evaluation frameworks, but they remain insufficient given the ever-expanding range of applications for LLMs.

    b. **User demographics and cultural sensitivities:** An LLM interacts with end-users, each bringing their own set of cultural norms, values, languages, and technological adeptness. Evaluations must consider these variables to ensure the LLM's performance and responses are attuned to the users it serves, thereby mitigating potential misinterpretations or misalignments. In this regard, we note that:

        i. The prevalent framing of toxicity, bias, and demographic considerations in LLM evaluations tends to be Western-centric. However, the interpretation of potentially toxic statements and the impact of bias in LLMs varies across cultural and social groupings. For example, certain statements might be deemed toxic in some settings, but not in others. Thus, evaluation concepts should be expanded to include diverse global perspectives and values.

        ii. Most existing benchmark datasets and tools are primarily developed in English. As LLMs find applications in multilingual and multicultural settings, datasets and frameworks that enable assessments across various languages are crucial because evaluation, especially on issues like bias and toxicity, can manifest differently across languages and linguistic structures.

c. **Operational jurisdiction:** Different jurisdictions impose varied regulations, laws, and compliance requirements that can impact an LLM's operation and outputs. Evaluations must consider these legalities to ensure the LLM operates within the bounds of the law while delivering value.

18. As LLMs continue to permeate various sectors and applications, their **evaluation cannot remain tethered to a one-size-fits-all approach**. Each layer of context highlighted above introduces its own set of challenges and considerations, emphasizing the need for a more tailored assessment paradigm.

Developing Evaluations for Frontier Models

19. As frontier models continue to advance and surpass human-like capabilities in various domains, it becomes crucial to carefully consider their impact. If not adequately controlled or aligned with human objectives and values, these models have the potential to cause significant harm. For example, a misaligned frontier model in the financial sector could contribute to market manipulation, insider trading, or cause systemic financial crises.

20. A concerning trend is the **significant gap between the development of frontier models and the corresponding tools and methodologies to effectively address their safety and alignment.** This disparity is evident in the lack of evaluations for many risks in the "Extreme Risks" category above. Without a thorough understanding of these risks and their potential consequences (e.g., dangerous capabilities such as persuasion and manipulation and developing political strategies, and how these might impact election outcomes), it is challenging to develop appropriate safeguards and mitigation strategies.

21. While there have been steps taken to address this gap (e.g., establishment of the Frontier Model Forum), more concerted efforts are required. **We thus echo the call in our Discussion Paper for a global and concerted effort, involving policymakers, researchers, and organisations, to further explore the unique risks posed by frontier LLMs and develop the requisite testing tools and resources to better evaluate and address these issues.** This allows us to harness the full potential of these transformative technologies while minimizing their associated risks.

Developing Standards to Ensure Robust and Trustworthy LLM Evaluation Frameworks

22. **The growing reliance on LLMs in sectors like healthcare and finance underscores the importance of robust standards for LLM evaluations.** However, current evaluation standards may lack the rigorous methodological underpinnings needed to ensure the representativeness of datasets (e.g., a sentiment analysis dataset primarily consisting of movie reviews by film critics may not be representative of the language used by the wider population). Similarly, there is a need to ensure that evaluation metrics are reflective of a LLM's

performance and weaknesses (Chang et al., 2023 and Liang et al., 2022) and competently measure what they are designed to assess.

23. **Imprecise or faulty benchmarks and metrics can lead to a mistaken sense of confidence in a model's capabilities.** This could potentially lead to developers being blindsided to critical areas of deficiency. It may also act as red herrings for LLM developers, resulting in wasted resources from focusing on enhancing aspects of the model that might not be pertinent. The broader LLM community may also be hampered by faulty datasets, as it not only slows down the pace of innovation but also fragments the community's understanding and advancement of LLM evaluation.

24. **Further, there are currently no agreed-upon methodologies that guide the evaluation approach for LLMs.** The variables are manifold and the testing approaches and scoring methods used can vastly differ even when evaluating a common attribute. While the focus of evaluation differs across use cases and industry verticals, standardised methodologies would enable a common frame of reference. In red teaming, the selection criteria for human evaluators and red teamers, and the instructions provided to them, are similarly inconsistent.

25. **Inconsistent methodologies result in several complications.** It can lead to inconsistent evaluation results for the same model, complicating cross-study comparisons. It also jeopardizes the foundational principle of reproducibility, making it challenging for other researchers to replicate evaluations, diminishing the credibility of evaluation findings. It may also allow for unintentional introduction of personal or institutional biases during evaluation. In red teaming scenarios, a lack of baselines in capability, training, and approaches can lead to inconsistent detection of critical vulnerabilities, which is particularly concerning when deploying LLMs in real-world contexts.

26. As we increasingly rely on LLMs, the need for robust, standardized assessment frameworks and methodologies is paramount. **The path forward must be characterized by collaborative efforts to establish rigorous, universally accepted benchmarks and methodologies.** This will ensure LLM evaluations are reliable and accurately reflect real-world performance, laying a foundation of trust for all stakeholders.

Imperative for a Multi-Faceted Evaluation Approach for LLMs

27. Automated benchmarking, primarily consisting of structured questions and answers, form the majority of LLM evaluations in the current landscape. Their appeal stems from their straightforward methodology, cost-effectiveness, and scalability. However:

    a. These benchmarks are largely rooted in surface-level features and are typically applicable for a limited set of tasks. Specifically, their scope often does not aptly cover open-ended tasks, like ensuring adherence to multi-turn dialogue instructions. Consequently, their comprehensiveness and direct correlation to actual model performance can be restricted. They also may not adequately assess LLMs' alignment with genuine human preferences.

b. Most automated benchmarking tools originated for pre-trained LLMs (Touvron et al., 2023). Hence, their applicability for assessing task fine-tuned LLMs is questionable. Indeed, there are indications that such evaluations fall short in discerning between pre-trained models and their aligned counterparts (Zheng et al., 2023).

28. Red teaming as a testing approach serves a unique and invaluable role in the assessment of LLMs. By deliberately probing these models, **red teaming uncovers behaviours that might otherwise escape detection.** This form of evaluation is particularly critical in LLM-driven applications with significant societal implications – whether concerning cultural sensitivity, data security, the propagation of misinformation, or ethical dilemmas like bias and discrimination.

29. **Red teaming is not without its challenges and limitations.** Firstly, manual red teaming is resource-intensive both in terms of time and cost, which might make it less accessible for smaller projects or organizations. Secondly, the quality of a red teaming evaluation is closely tied to the expertise and impartiality of the team conducting it. A team lacking in skill or hampered by biases may fail to rigorously probe an LLM's vulnerabilities, thereby inducing a false sense of security.

30. Regarding scoring, **human scoring offers a more nuanced examination of LLM performance, notably in realistic settings, and represents the 'gold standard' for assessing alignment with human preferences** (Zheng et al., 2023). Despite these advantages, they come with their own set of challenges. Firstly, they are time-intensive, often expensive, and challenging to scale effectively. Further:

a. Human scorers display a **central tendency bias**, gravitating towards middle scores on the Likert scale. This behaviour results in a more evenly distributed yet less differentiated set of evaluations (Ye et al., 2023).

b. **Human scorers experience fatigue**, especially when tasked with knowledge-intensive evaluations. This form of scoring is not scalable to large datasets, and fatigue may also lead to potential inconsistencies in assessments (Ye et al., 2023).

c. **Results can be subjective** and dependent on human scorers.

31. In recent times, there's been an increasing trend that gears towards model scoring (i.e., LLM-to-LLM evaluations). **By employing LLMs that closely align with human preferences, this approach is a cost-effective alternative, being 22 times cheaper and 129 times faster than human scoring** (Ye et al., 2023). However, as pointed out by Zheng et al. (2023), this approach isn't without its challenges:

a. An LLM evaluator may display **position bias**. For instance, it may prefer answers that appear at the beginning or end of a list in the prompt, overlooking the content's accuracy or relevance.

b. An LLM evaluator may exhibit **verbosity bias**, preferring longer responses even if they lack the clarity, quality, or precision of more concise alternatives.

c. An LLM evaluator may display a partiality towards the responses that itself has produced (i.e., **self-enhancement bias**).

32. The observations highlighted above reinforce the **need to adopt a multi-faceted approach to LLM evaluation.** By **making appropriate use of the various testing approaches and scoring methods**, informed by the use-case and other relevant context, a multi-faceted evaluation approach:

    a. Provides both breadth and depth in LLM evaluation;
    b. Strikes a suitable balance between scale, speed and depth of assessment;
    c. Addresses the biases and weaknesses present in its constituent approaches and presents a more balanced view; and
    d. Validates and verifies the results from its constituent approaches.

Limitations and Future Work

33. While this paper aims to provide a comprehensive overview of LLM evaluation, there are several areas it does not cover that are nonetheless critical to the safe and responsible development and deployment of LLMs, such as assessing LLMs' interpretability and explainability, and the data used to train these models. These warrant exploration in future work:

34. **Evaluations of LLM training data.** The nature and composition of training data significantly influence an LLM's performance and behaviour (Mökander et al., 2023). Evaluations of training data, such as demographic representation and toxicity prevalence, can increase transparency, inform downstream mitigation efforts, and guide appropriate model use. Further, such evaluations can shed light on whether the training data contains instances of testing data used to evaluate LLMs, which directly impacts whether evaluation findings are generalizable (Liang et al., 2022).

35. **Evaluations of the environmental impact of training and deploying LLMs.** In addition to model safety and performance, model efficiency and environmental sustainability is a key model quality that should be assessed. As LLMs grow in complexity and size, their demand for computational resources increases, with significant environmental implications, particularly regarding energy and water consumption, and carbon emissions. Future work on environmental impact assessments could contribute to developing more energy-efficient algorithms, using renewable energy sources, and designing high-performance LLMs with reduced computational requirements.

36. **Evaluations of LLMs' interpretability and explainability.** As set out in NIST (2023), explainability refers to understanding the mechanisms that a LLM used to arrive at a certain output while interpretability refers to understanding why a certain output was generated and what it means in the context of the LLM's intended function. These attributes, crucial for building user trust and identifying model

errors or biases, are challenging to assess as LLMs’ internal workings and decision-making processes are not easily interpretable or explainable. Nonetheless, work in these areas continues at a rapid pace with researchers studying new techniques and tools (e.g., mechanistic interpretability). If LLMs become more interpretable and explainable, a compendium of evaluation methods focused on these attributes would be invaluable.

37. **Evaluations of the potential long-term effects of deploying LLMs.** The long-term effects of LLM use can span societal, economic, and behavioural domains. For instance, societal impacts might include changes in employment patterns due to automation, the spread of misinformation, or shifts in social dynamics due to the pervasive use of AI systems. Developing evaluations in these areas would likely require interdisciplinary collaboration and the development of new metrics and methodologies capable of capturing these complex, multi-faceted impacts.

38. **Evaluations that assess system security of LLM-driven applications.** This issue is especially pressing given the increasing connectivity of LLMs to the Internet and their integration with various plugins, which introduce additional attack vectors. Evaluations could focus on assessing dataset poisoning and the vulnerability of LLM-driven applications to prompt-injection attacks.

Part 3 - Recommended Baseline for LLM Evaluation

39. In this part, we recommend a **baseline set of pre-deployment evaluations for LLMs for safety and trustworthiness that should be conducted irrespective of use case.** While we acknowledge the previously highlighted limitations in current evaluation and testing approaches, there is still value in setting out baseline evaluations as these will help ensure a minimum level of LLM safety.

40. Assessing an LLM's capabilities is crucial, and the catalogue includes evaluations of an LLM's general and domain-specific capabilities. However, the exact capabilities to evaluate will vary based on the intended use-case. Further, enumerating the capabilities that ought to be prioritized for an application is best undertaken by the deploying organization, which has a better understanding of the operational context and objectives.

41. Instead, **our primary focus is on evaluations aimed at ascertaining the safety and trustworthiness of LLMs.** These can form a universal baseline that should be conducted irrespective of specific use-cases. Conducting these pre-deployment evaluations are a necessary step in ensuring that a LLM meets a minimum safety threshold, and our proposed baseline represents our policy position on using evaluations to enhance the LLM ecosystem's safety and trustworthiness.

42. **Our recommended set of safety and trustworthiness evaluations take reference from the 11 governance principles delineated in the AI Verify Framework** ("AI Verify Principles")[13]. AI Verify is an AI governance testing framework and software toolkit that validates the performance of supervised-learning AI systems against internationally recognized principles through standardized tests. Its principles are consistent with international AI governance frameworks, such as those from US, EU, and OECD. While the AI Verify Toolkit does not currently support the evaluation of generative AI models like LLMs, its principles are nonetheless instructive in deriving safety and trustworthiness assessments that align with international best practices.

43. Each of the AI Verify Principles addresses a different concern raised by the six dimensions in the Discussion Paper. Since this paper focuses on the third dimension of model development, deployment, and testing, we started by first identifying the principles that were relevant to this dimension. These were: **explainability, reproducibility, robustness, fairness, data governance, human agency and oversight, and security.**

44. We then assessed if these principles: (i) related to model evaluations; and (ii) were applicable to LLMs in general, irrespective of use-case:

[13] See https://aiverifyfoundation.sg/downloads/AI_Verify_Sample_Report.pdf for more information

Principle	Description	Analysis
Explainability	Understand and interpret what the AI system is doing	This principle relates to model evaluations, specifically, assessments to determine why an AI system reached the decision that it did. However, most LLMs typically function as 'black-box' models, making it inherently challenging to understand and interpret why they produced a specific output. As such, the concept of explainability may be more relevant to specific applications of LLMs. For example, where an LLM interfaces with an external database, the capacity to cite sources could provide a semblance of explainability by revealing the data points the model used to inform its output. In the circumstances, we do not propose to utilize this principle. We may revisit this decision in the future should LLMs become more interpretable (e.g., driven by novel research into new approaches such as being able to dissect training algorithms through mechanistic interpretability).
Reproducibility	AI system's results are consistent and can be replicated	This principle relates to model evaluations, specifically assessments to review if an AI model produces the same output for the same input. However, we note that LLMs are stochastic models and reproducibility is not a universally desired capability of LLMs. For example, in creative-writing applications, strict reproducibility may not only be unnecessary but could in fact be counterproductive. Given the nuanced utility of reproducibility in the context of LLMs, we do not propose to use this principle to help inform a baseline set of evaluations that should apply irrespective of use-case.
Robustness	AI system should be resilient against	This principle relates to model evaluations, specifically assessments to

	attacks and attempts at manipulation by third party malicious actors, and can still function despite unexpected input	determine if an AI model can maintain its level of performance under any circumstances. In the context of LLM evaluations, this principle may be extended to encompass assessing the ability of a LLM to produce accurate and reliable output in the face of different types of input (e.g., adversarial, out-of-distribution).
Fairness	AI should not result in unintended and inappropriate discrimination against individuals or groups	This principle relates to model evaluations, specifically assessments to determine if an AI model produces biased output. In the context of LLM evaluations, this principle may be extended to also encompass assessing the tendency of a LLM to generate toxic statements.
Data Governance	Governing data used in AI systems, including putting in place good governance practices for data quality, lineage, and compliance	In the context of LLM evaluations, this principle may be extended to encompass assessing the tendency of a LLM to memorize and regurgitate training data in their outputs.
Human Agency & Oversight	Ability to implement appropriate oversight and control measures with humans-in-the-loop at the appropriate juncture	This principle does not relate to model evaluations. It focuses on organisational structures, decision-mechanisms, appropriate oversight, and control measures.
Security	AI security is the protection of AI systems, their data, and the associated infrastructure from unauthorised access, disclosure, modification, destruction, or disruption.	This principle does not conventionally relate to model evaluations. It primarily focuses on organisational security measures to ensure the confidentiality, integrity, and availability of the AI system.

45. From the analysis above, we identified the following AI Verify Principles as being related to model evaluations and applicable to LLMs irrespective of use-case: r**obustness, fairness, and data governance.** The table below sets out our recommendations on the baseline set of LLM attributes to evaluate for each of these principles:

Principle	Elaboration of Principle	Recommended LLM Attributes for Evaluation
Robustness	Individuals know that the AI system will perform according to intended purpose, even when encountering unexpected inputs. In the context of LLM evaluations, this principle may also encompass assessing the ability of a LLM to generate accurate output and not 'hallucinate'.	• Robustness • Factuality
Fairness	Individuals know that the AI system does not unintentionally discriminate. In the context of LLM evaluations, this principle may also encompass assessing the tendency of a LLM to generate toxic statements.	• Bias • Toxicity generation
Data Governance	Individuals know that the data used in the AI system is compliant with the relevant regulation and standards. In the context of LLM evaluations, this principle may be extended to encompass assessing the tendency of a LLM to regurgitate training data in their outputs.	• Data Governance

46. Finally, we set out our recommendations on the evaluation and testing approaches that may be used to assess LLMs for each of the identified attributes. **In selecting these, we have focused on factors such as the comprehensiveness, ease of implementation, and scalability of the evaluations.**

LLM Attribute	Recommended Evaluation and Testing Approach	Remarks
Robustness	**Evaluation Framework:** DecodingTrust	This evaluation assesses various aspects of a LLM's robustness within a single evaluation suite.
Factuality	**Benchmark:** TruthfulQA	The TruthfulQA benchmark was used by Meta, OpenAI and Anthropic to evaluate the

	Evaluation Framework: HELM / BigBench / Eleuther Evaluation Harness	factuality of Llama 2, GPT-4, and Claude 2 respectively.
Bias	*Stereotype Bias* **Benchmark:** Bias Benchmark for Question Answering (BBQ) **Evaluation Framework:** HELM / BigBench *Fairness* **Benchmark:** UCI Adult dataset **Evaluation Framework:** DecodingTrust *Representation of Subjective Opinions* **Benchmark:** GlobalOpinionQA **Evaluation Framework:** Set out in Durmus et al. (2023) *Capability Fairness* **Benchmark:** TwitterAAE **Evaluation Framework:** HELM	The BBQ benchmark was used by Anthropic to evaluate its Claude 2 LLM. The UCI Adult benchmark has been used widely used to assess fairness and its limitations had been highlighted as well (Ding et al., 2021).
Toxicity Generation	**Benchmark:** RealToxicityPrompts **Scoring method:** Model scoring with Perspective API	The RealToxicityPrompts benchmark is used in both the HELM and DecodingTrust framework to assess toxicity. Perspective API has been extensively tested and its limitations have been highlighted in prior work: HELM
Data Governance	*Personal data* **Benchmark:** Pre-processed version of Enron Email Dataset created by Huang at al. (2022) **Evaluation Framework:** DecodingTrust *Non-personal data* **Benchmark:** Pre-processed dataset in HELM[14] **Evaluation Framework:** HELM	To assess whether a LLM regurgitates its training data, one would first need to know the contents of the training data. However, the official documentation for the latest LLMs rarely disclose such details, rendering such assessments difficult. Nonetheless, we accept the assumption set out in Wang et al. (2023) that the Enron Email

[14] Details of the pre-processed dataset are set out in section E.4 ("Memorization & copyright") of Liang et al. (2022).

		dataset is likely utilized when training LLMs.

47. In a section above, we emphasized the need for more rigorous methodological foundations for dataset representativeness and validity. The same applies to evaluation and testing approaches. The lack of widely accepted standards and best practices in this area only exacerbates these challenges. **Thus, our recommendation should not be taken as an endorsement of the reliability and validity of the identified evaluation and testing approaches.** Instead, our recommendations were selected based on their comprehensiveness, ease of implementation and scalability. But as the field matures and as more sophisticated, and standardized, evaluation tools are developed, we anticipate revisiting this aspect to provide revised guidance.

48. Finally, we set out three factors to consider when utilizing the proposed baseline set of evaluations.

49. **Firstly, evaluations should ideally be conducted using evaluation and testing approaches that are contextually attuned.** For instance, the evaluation of an LLM's bias should be conducted using benchmarks and frameworks that are attuned to the LLM's user group since bias can only be defined in relation to user demographics and other social and cultural factors (Mökander et al., 2023). **However, if such specialized tools are not available, organizations should use the generic benchmarks and frameworks highlighted above as a minimum precautionary measure.** This dual approach—specialized when possible, but generalized when necessary—ensures that LLMs are subjected to rigorous scrutiny, thereby facilitating their safe and responsible deployment.

50. **Secondly, where organisations finetune a foundational LLM before deployment, a repeat of some evaluations may be warranted.** After a LLM has been developed, an organisation can finetune it (e.g., on specific domain data) before deployment. Such finetuning can inadvertently introduce or exacerbate undesired behaviours. Therefore, deploying organisations should consider running the recommended evaluations again after finetuning.

51. An organization may also use tools to improve the performance of an LLM-driven application before deployment. For example, an organisation might use a trusted document repository to present pertinent documents to the LLM, in order to enhance its response accuracy. **Organisations should examine the nature of the tools used to decide which evaluations should be re-conducted.** In the example above, repeating data governance evaluations may not be necessary as the tool used did not affect the LLM's training data. However, the organization may need to conduct more specific robustness and bias re-evaluations to assess how the LLM performs in light of the documents presented to it via the trusted document repository. Such evaluations should also be conducted periodically post-deployment, especially if the tools used continue to evolve.

52. Lastly, **additional evaluations may be warranted for frontier model risks.** The proposed baseline evaluations provide a universally applicable starting point for assessing safety and trustworthiness, irrespective of the specific LLM being assessed. However, additional evaluations are necessary to address the unique challenges of frontier model risks. These include dangerous capabilities, such as the potential ability to create more effective and larger scale cyberattacks, and the higher risk of losing human control due to factors such as an ability to autonomously replicate and to manipulate human users.

53. The testing and evaluation of frontier model risks is still nascent. **Nonetheless, organisations should always ascertain if the model they are developing may potentially exhibit such risks and if so, make use of the latest tools and techniques to detect and assess these risk factors.** This should be done in addition to the proposed baseline evaluations. While this adds complexity to the evaluation process, it is a critical step in ensuring the safe and responsible deployment of LLMs.

Conclusion

54. This paper presents a comprehensive, but non-exhaustive, overview of the LLM evaluation and testing landscape, categorizing the available methods and tools to facilitate the assessment of LLM capabilities and risks. We have emphasized the need for further safety and alignment research, context-specific evaluations, and multi-faceted evaluation approaches that provide a more holistic understanding of LLM capabilities and risks. These represent opportunities for research and development.

55. Finally, our baseline recommendations for LLM evaluation reinforce minimum standards of LLM safety and trustworthiness and we encourage organisations to minimally conduct those evaluations before LLM release and deployment. These recommendations should be seen as a starting point, rather than a comprehensive and fully matured solution. The rapidly evolving nature of LLMs means that these recommendations must be continually reassessed to ensure they remain relevant and effective.

Annex A – References

1. Anthropic. (2023). Model Card and Evaluations for Claude Models

2. Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, SN., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L., and Khabsa, M. (2023) The Belele Benchmark: A Parallel Reading Comprehension Dataset in 122 Language Variants. arXiv preprint arXiv:2308.16884

3. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2023) A Survey on Evaluation of Large Language Models. arXiv preprint arXiv:2307.03109v2

4. Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring Adult: New Datasets for Fair Machine Learning. https://openreview.net/forum?id=bYi_2709mKK

5. Durmus, E., Nguyen, K., Liao, T., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., et al. (2023). Towards measuring the representation of subjective global opinions in language models. arXiv preprint arXiv:2306.16388.

6. Feng, S., Park, C., Liu, Y., and Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. arXiv preprint arXiv:2305.08283v3.

7. FitzGerald, J., Hench, C., Peris., Mackie, S., Rottmann, K., Sanchez, A., Nash, A., Urbach, L., Kakarala, V., Singh, R., et al. (2022). MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. arXiv preprint arXiv:2204.08582v2.

8. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858v2.

9. Griffin, L., Kleinberg, B., Mozes, M., Mai, K., Vau, M., Caldwell, M., and Marvor-Parker, A. (2023). Susceptibility to influence of large language models. arXiv preprint arXiv:2303.06074.

10. Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., et al. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. arXiv preprint arXiv:2308.11462.

11. Hazell, J. (2023). Large language models can be used to effectively scale spear phishing campaigns. arXiv preprint arXiv:2305.06972.

12. Huang, J., Shao, H., and Chang, K. (2022). Are Large Pre-Trained Language Models Leaking Your Personal Information? arXiv preprint arXiv:2205.12628v2

13. Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. (2023). Co-writing with opinionated language models affects users' views. arXiv preprint arXiv:2302.00560.

14. Karra, S., Nguyen, S., and Tulabandhula, T. (2023). Estimating the personality of white-box language models. arXiv preprint arXiv:2204.12000v2.

15. Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083.

16. Lesher, M., Pawelec, H., and Desai, A. (2022) Disentangling untruths online: Creators, spreaders and how to stop them. *OECD Going Digital Toolkit Notes No. 23*. https://doi.org/10.1787/ec5958b3-en

17. Li, X., Li, Y., Joty, S., Liu, L., Huang, F., Qiu, L., and Bing, L. (2023). Does GPT-3 demonstrate psychopathy? Evaluating large language models from a psychological perspective. arXiv preprint arXiv:2212.10529v2.

18. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

19. Mökander, J., Schuett, J., Kirk, H.R., and Floridi, L. (2023). Auditing large language models: a three-layered approach. *AI Ethics*. doi: 10.1007/s43681-023-00289-2

20. NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1

21. OpenAI. (2023). GPT-4 System Card.

22. Parisi, A., Zhao, Y., and Fiedel, N. (2022). Talm: Tool augmented language models. arXiv preprint arXiv:2205.12255.

23. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., Irving, G. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286.

24. Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., et al. (2023). ToolLLM: Facilitating large language models to master 16000+ real-world APIs. arXiv preprint arXiv:2307.16789.

25. Raji, I., Bender, E., Paullada, A., Denton, E., and Hanna, A. (2021). AI and the Everything in the Whole Wide World Benchmark. arXiv preprint arXiv:2111.15366v1.

26. Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., and Pauly, M. (2023). The self-perception and political biases of ChatGPT. arXiv preprint arXiv:2304.07333.

27. Schick, T., Dwivedi-Yu, J., Dess`ı, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761.

28. Shaham, U., Segal, E., Ivgi, M., Efrat, A., Yoran, O., Haviv, A., Gupta, A., Xiong, W., Geva, M., Berant, J., et al. (2022) SCROLLS: Standardized CompaRison Over Long Language Sequences. arXiv preprint arXiv:2201.03533v2.

29. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. (2023). HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. arXiv preprint arXiv:2303.17580.

30. Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., et al. (2023). Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324.

31. Singhal, K., Azizi, S., Tu, T., Mahdavi, S., Wei, J., Chung, H., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. arXiv preprint arXiv:2212.13138.

32. Smith, E., Hall, M., Kambadur, M., Presani, E., and Williams, A. (2022). “I’m sorry to hear that”: Finding new biases in language models with a holistic descriptor dataset. arXiv preprint arXiv:2205.09209v2.

33. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

34. Stevenson, C., Smal, I., Baas, M., Grasman, R., and Maas, H. (2022). Putting GPT-3’s creativity to the (alternative uses) test. arXiv preprint arXiv:2206.08932.

35. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale., et al. (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models

36. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al. (2023) Towards Generalist Biomedical AI. arXiv preprint arXiv:2307.14334.

37. Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al. (2023). DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. arXiv preprint arXiv:2306.11698.

38. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564v2.

39. Wynter, A., Wang, X., Sokolov, A., Gu, Q., and Chen, S. (2023). An evaluation of large language model outputs: Discourse and memorization. arXiv preprint arXiv:2304.08637.

40. Ye, S., Kim, D., Kim, S., Hwang, H., Kim, S., Jo, Y., Thorne, J., Kim, J., Seo, M. (2023). FLASK: Fine-grained language model evaluation based on alignment skill sets. arXiv preprint arXiv:2307.10928

41. Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685v2.

Other resources

ARC Evals

Big-bench

Language Model Evaluation Harness

Hugging Face

Mosaic Eval Gauntlet

Annex B – Catalogue of Evaluation Frameworks, Benchmarks & Papers

Task / Attribute	Evaluation Framework/Benchmark/Paper	Testing Approach
1.1. Natural Language Understanding		
Text classification	HELM • Miscellaneous text classification	Benchmarking
	Big-bench • Emotional understanding • Intent recognition • Humor	Benchmarking
	Hugging Face • Text classification • Token classification • Zero-shot classification	Benchmarking[15]
Sentiment analysis	HELM • Sentiment analysis	Benchmarking
	Evaluation Harness • GLUE	Benchmarking
	Big-bench • Emotional understanding	Benchmarking
Toxicity detection	HELM • Toxicity detection	Benchmarking
	Evaluation Harness • ToxiGen	Benchmarking
	Big-bench • Toxicity	Benchmarking
Information retrieval	HELM • Information retrieval	Benchmarking
Sufficient information	Big-bench • Sufficient information	Benchmarking
	FLASK • Metacognition	Benchmarking (with human and model scoring)
Natural language inference	Evaluation Harness • GLUE	Benchmarking
	Big-bench • Analytic entailment (specific task)	Benchmarking

[15] There are metrics defined in Hugging Face for text classification and token classification. However, no metrics have been defined for the zero-shot classification task.

	• Formal fallacies and syllogisms with negation (specific task) • Entailed polarity (specific task)	
General English understanding	HELM • Language	Benchmarking
	Big-bench • Morphology • Grammar • Syntax	Benchmarking
	Evaluation Harness • BLiMP	Benchmarking
	Eval Gauntlet • Language Understanding	Benchmarking
1.2. Natural Language Generation		
Summarization	HELM • Summarization	Benchmarking
	Big-bench • Summarization	Benchmarking
	Evaluation Harness • BLiMP	Benchmarking
	Hugging Face • Summarization	Benchmarking
Question generation and answering	HELM • Question answering	Benchmarking
	Big-bench • Contextual question answering • Reading comprehension • Question generation	Benchmarking
	Evaluation Harness • CoQA • ARC	Benchmarking
	FLASK • Logical correctness • Logical robustness • Logical efficiency • Comprehension • Completeness	Benchmarking (with human and model scoring)
	Hugging Face • Question Answering	Benchmarking
	Eval Gauntlet • Reading Comprehension	Benchmarking
Conversations and dialogue	MT-bench	Benchmarking (with human and model scoring)
	Evaluation Harness	Benchmarking

	• MuTual	
	Hugging Face • Conversational	Benchmarking
Paraphrasing	Big-bench • Paraphrase	Benchmarking
Other response qualities	FLASK • Readability • Conciseness • Insightfulness	Benchmarking (with human and model scoring)
	Big-bench • Creativity	Benchmarking
	Putting GPT-3's Creativity to the (Alternative Uses) Test	Benchmarking (with human scoring)
Miscellaneous text generation	Hugging Face • Fill-mask • Text generation	Benchmarking
1.3. Reasoning	HELM • Reasoning	Benchmarking
	Big-bench • Algorithms • Logical reasoning • Implicit reasoning • Mathematics • Arithmetic • Algebra • Mathematical proof • Fallacy • Negation • Computer code • Probabilistic reasoning • Social reasoning • Analogical reasoning • Multi-step • Understanding the World	Benchmarking
	Evaluation Harness • PIQA, PROST - Physical reasoning • MC-TACO - Temporal reasoning • MathQA - Mathematical reasoning • LogiQA - Logical reasoning • SAT Analogy Questions - Similarity of semantic relations • DROP, MuTual – Multi-step reasoning	Benchmarking
	Eval Gauntlet • Commonsense reasoning • Symbolic problem solving	Benchmarking

	• Programming	
1.4. Knowledge and factuality	HELM • Knowledge	Benchmarking
	Big-bench • Context Free Question Answering.	Benchmarking
	Evaluation Harness • HellaSwag, OpenBookQA – General commonsense knowledge • TruthfulQA – Factuality of knowledge	Benchmarking
	FLASK • Background Knowledge	Benchmarking (with human and model scoring)
	Eval Gauntlet • World Knowledge	Benchmarking
1.5. Effectiveness of tool use	HuggingGPT	Benchmarking (with human and model scoring)
	TALM	Benchmarking
	Toolformer	Benchmarking (with human scoring)
	ToolLLM	Benchmarking (with model scoring)
1.6. Multilingualism	Big-bench • Low-resource language • Non-English • Translation	Benchmarking
	Evaluation Harness • C-Eval (*Chinese evaluation suite*) • MGSM • Translation	Benchmarking
	BELEBELE	Benchmarking
	MASSIVE	Benchmarking
	HELM • Language (Twitter AAE)	Benchmarking
	Eval Gauntlet • Language Understanding	Benchmarking
1.7. Context length	Big-bench • Context length	Benchmarking
	Evaluation Harness • SCROLLS	Benchmarking
2.1. Law	LegalBench	Benchmarking (with algorithmic and human scoring)

2.2. Medicine	Large Language Models Encode Clinical Knowledge	Benchmarking (with human scoring)
	Towards Generalist Biomedical AI	Benchmarking (with human scoring)
2.3. Finance	BloombergGPT	Benchmarking
3.1. Toxicity generation	HELM • Toxicity	Benchmarking
	DecodingTrust • Toxicity	Benchmarking
	Red Teaming Language Models to Reduce Harms	Manual Red Teaming
	Red Teaming Language Models with Language Models	Automated Red Teaming
3.2. Bias		
Demographical representation	HELM	Benchmarking
	Finding New Biases in Language Models with a Holistic Descriptor Dataset	Benchmarking
Stereotype bias	HELM • Bias	Benchmarking
	DecodingTrust • Stereotype Bias	Benchmarking
	Big-bench • Social bias • Racial bias • Gender bias • Religious bias	Benchmarking
	Evaluation Harness • CrowS-Pairs	Benchmarking
	Red Teaming Language Models to Reduce Harms	Manual Red Teaming
Fairness	DecodingTrust • Fairness	Benchmarking
Distributional bias	Red Teaming Language Models with Language Models	Automated Red Teaming
Representation of subjective opinions	Towards Measuring the Representation of Subjective Global Opinions in Language Models	Benchmarking
Political bias	From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models	Benchmarking
	The Self-Perception and Political Biases of ChatGPT	Benchmarking
Capability fairness	HELM	Benchmarking

	• Language (Twitter AAE)	
3.3. Machine ethics	DecodingTrust • Machine Ethics	Benchmarking
	Evaluation Harness • ETHICS	Benchmarking
3.4. Psychological traits	Does GPT-3 Demonstrate Psychopathy?	Benchmarking
	Estimating the Personality of White-Box Language Models	Benchmarking
	The Self-Perception and Political Biases of ChatGPT	Benchmarking
3.5. Robustness	HELM • Robustness to contrast sets	Benchmarking
	DecodingTrust • Out-of-Distribution Robustness • Adversarial Robustness • Robustness Against Adversarial Demonstrations	Benchmarking
	Big-bench • Out-of-Distribution Robustness	Benchmarking
	Susceptibility to Influence of Large Language Models	Benchmarking
3.6. Data governance	DecodingTrust • Privacy	Benchmarking
	HELM • Memorization and copyright	Benchmarking
	Red Teaming Language Models to Reduce Harms	Manual Red Teaming
	Red Teaming Language Models with Language Models	Automated Red Teaming
	An Evaluation on Large Language Model Outputs: Discourse and Memorization	Benchmarking (with human scoring)
4.1. Dangerous Capabilities		
Offensive cyber capabilities	GPT-4 System Card • Cybersecurity	System Card
Weapons acquisition	GPT-4 System Card • Proliferation of Convention and Unconventional Weapons	System Card
Self and situation awareness	Big-bench • Self-Awareness	Benchmarking
Autonomous replication / self-proliferation	ARC Evals • Autonomous replication	Manual Red Teaming

Persuasion and manipulation	HELM • Narrative Reiteration • Narrative Wedging	Benchmarking (with human scoring)
	Big-bench • Convince Me (specific task)	Benchmarking
	Co-writing with Opinionated Language Models Affects Users' Views	Manual Red Teaming
5.1. Misinformation	HELM • Question answering • Summarization	Benchmarking
	Big-bench • Truthfulness	Benchmarking
	Red Teaming Language Models to Reduce Harms	Manual Red Teaming
5.2. Disinformation	HELM • Narrative Reiteration • Narrative Wedging	Benchmarking (with human scoring)
	Big-bench • Convince Me (specific task)	Benchmarking
5.3. Information on harmful, immoral or illegal activity	Red Teaming Language Models to Reduce Harms	Manual Red Teaming
5.4. Adult content	Red Teaming Language Models to Reduce Harms	Manual Red Teaming

At IMDA, we see ourselves as Architects of Singapore's Digital Future. We cover the digital space from end to end, and are unique as a government agency in having three concurrent hats – as Economic Developer (from enterprise digitalisation to funding R&D), as a Regulator building a trusted ecosystem (from data/AI to digital infrastructure), and as a Social Leveller (driving digital inclusion and making sure that no one is left behind). Hence, we look at the governance of AI not in isolation, but at that intersection with the economy and broader society. By bringing the three hats together, we hope to better push boundaries, not only in Singapore, but in Asia and beyond, and make a difference in enabling the safe and trusted use of this emerging and dynamic technology.

Recognising the importance of collaboration and crowding in expertise, Singapore set up the AI Verify Foundation to harness the collective power and contributions of the global open-source community to build AI governance testing tools. The mission of the AI Verify Foundation is to foster and coordinate a community of developers to contribute to the development of AI testing frameworks, code base, standards and best practices. It will establish a neutral space for the exchange of ideas and open collaboration, as well as nurture a diverse network of advocates for AI testing and drive broad adoption through education and outreach. The vision is to build a community that will contribute to the broader good of humanity, by enabling trusted development of AI. IMDA is a member of the Foundation.

Disclaimer