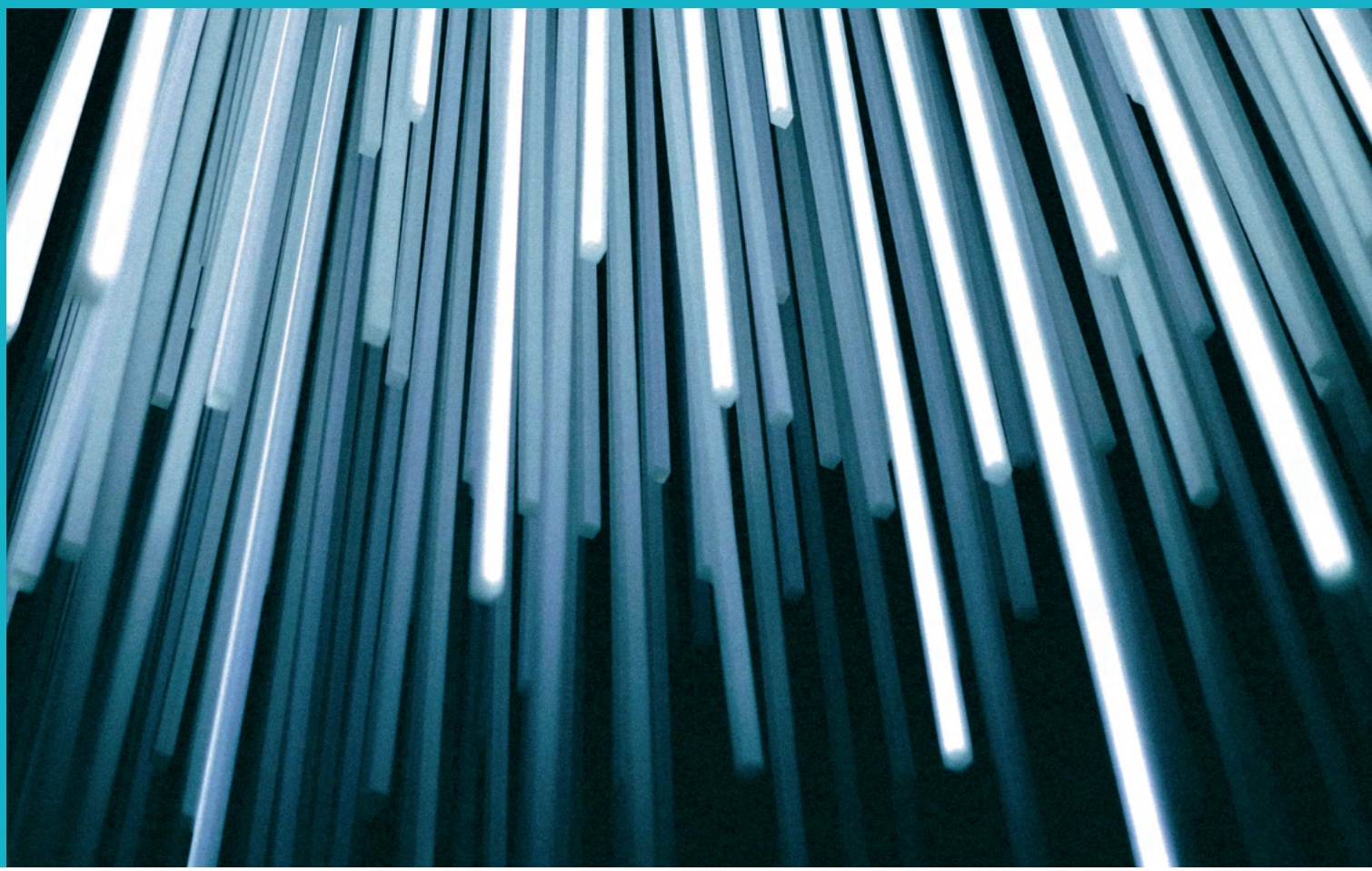
WHITEPAPER
October 2023

STRUCTURED ACCESS FOR THIRD-PARTY RESEARCH ON FRONTIER AI MODELS: INVESTIGATING RESEARCHERS' MODEL ACCESS REQUIREMENTS

Benjamin S. Bucknall, Robert F. Trager



In partnership with

Centre for the Governance of AI

STRUCTURED ACCESS FOR THIRD-PARTY RESEARCH ON FRONTIER AI MODELS: INVESTIGATING RESEARCHERS' MODEL ACCESS REQUIREMENTS

Benjamin S. Bucknall[*]
ben.s.bucknall@gmail.com

Robert F. Trager[†]
robert.trager@governance.ai

ABSTRACT

Recent releases of frontier artificial intelligence (AI) models have largely been gated, due to a mixture of commercial concerns and increasingly significant concerns about misuse. However, closed release strategies introduce the problem of providing external parties with enough access to the model for conducting important safety research. One potential solution is to use an API-based "structured access" approach to provide external researchers with the minimum level of access they need to do their work (i.e. "minimally sufficient access"). In this paper, we address the question of what access to systems is needed in order to conduct different forms of safety research. We develop a "taxonomy of system access"; analyse how frequently different forms of access have been relied on in published safety research; and present findings from semi-structured interviews with AI researchers regarding the access they consider most important for their work. Our findings show that insufficient access to models frequently limits research, but that the access required varies greatly depending on the specific research area. Based on our findings, we make recommendations for the design of "research APIs" for facilitating external research and evaluations of proprietary frontier models.

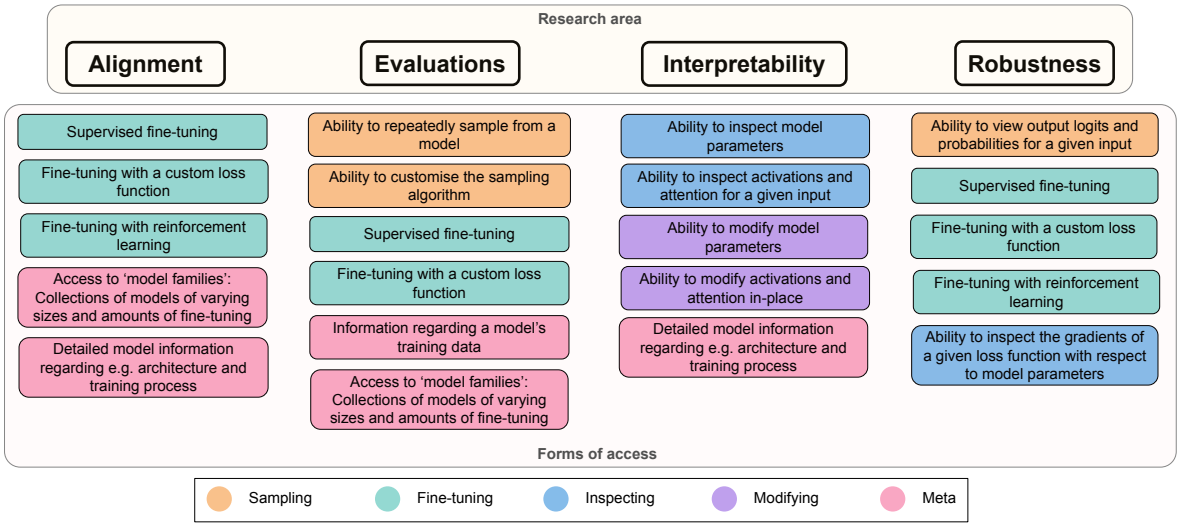


Figure 1: A breakdown of forms of model access that are essential for at least some valuable and currently practical projects in each of the four research areas considered

[*]Work completed whilst a Winter Research Fellow at the Centre for the Governance of AI, Oxford.
[†]Oxford Martin AI Governance Initiative, Centre for the Governance of AI, and Blavatnik School of Government, Oxford.
Author contributions: BSB set the direction of the project, conducted the literature analysis and interviews, and wrote the report; RFT proposed the initial project idea, provided supervision, and gave detailed feedback on the report.

Executive Summary

Recent releases of frontier artificial intelligence (AI) systems, including OpenAI's GPT-4 and Google's PaLM 2, have predominantly been gated, with access provided through online interfaces. This approach makes it substantially easier to prevent the misuse of models, in addition to better facilitating their commercialisation. However, such release strategies introduce the challenge of providing adequate access to systems to third-parties for the purpose of independent research and evaluation. One potential solution for facilitating external research and scrutiny of models while retaining the benefits of gated release is to provide researchers with minimal sufficient access to systems through structured access solutions, such as an Application Programming Interface (API). However, there is little clarity regarding the functionality that would be most useful for researchers if provided through such a service. Thus, in this report, we address the questions of *How is research currently affected by limited access to frontier models?* and *What access to models do AI researchers consider most important for different AI research areas?*

Contributions

We introduce a **taxonomy of system access** that categorises the different ways in which researchers can interact with a model for their research. The taxonomy consists of five categories:

1. The ability to **sample** from a model: this could include the ability to sample from the model in an automated manner, specifying the sampling algorithm and associated hyper-parameters, or access the probabilities and logits associated with the model's outputs.
2. The ability to **fine-tune** the model: this could include the ability to fine-tune it through supervised learning or reinforcement learning, as well as the option of using a custom-defined loss function.
3. The ability to **inspect** model internals: this could include the ability to inspect parameters, activations and attention patterns, gradients, and embeddings, as well as the ability to perform arbitrary computations on observed values.
4. The ability to **modify** model internals: this could include parameters, activations and attention patterns, and embeddings, as well as the ability to perform arbitrary computations in-place.
5. Access to relevant **additional system information and artefacts**: this could include pretraining datasets, training snapshots, collections of "model families",[3] and information about the model's architecture and training procedure.

Based on this taxonomy, we conducted an **analysis of the literature**, assessing how frequently papers from different subfields of AI research make use of each of the forms of access. Additionally, we held **interviews with AI researchers** in order to gain deeper insight into how researchers commonly interact with models, how access to models impacts their work, and how this might change in the coming years.

Findings

Based on both the literature analysis and researcher interviews, we highlight **four key findings**:

- **Limited access to models curtails certain research projects**
  Researchers' choice of research agenda is influenced considerably by access considerations, with certain agendas, such as some which aim to understand behaviours only exhibited by the most capable systems, being dropped due to insufficient access to suitable models.
- **A lack of model information limits the conclusions that can be drawn from results**
  Researchers are often unable to pose hypotheses or draw conclusions from experimental observations due to a lack of relevant information regarding the subject model's architecture or training procedure.
- **Basic sampling access is sufficient for many current model evaluations, but may not be for similar evaluations of future models**
  Research that aims to evaluate and measure the capabilities and safety of models is currently largely behavioural in nature, and thus frequently only depends on the ability to sample from a model, preferably after dedicated fine-tuning. However, researchers believe that it is important to move beyond behavioural evaluations in order to be able to develop a more comprehensive understanding of a system's functioning.

[3]We use the term 'model families' to refer to collections of related models that systematically differ among one or more dimensions. For example, a model family may comprise models of differing sizes that nonetheless share a common training procedure and dataset.

- **Interpretability techniques could become crucial for evaluating the capabilities and safety of models**
  Interpretability research that aims to explain how and why a model behaves as it does is one of the more access-intensive areas of research, often requiring the ability to both inspect and modify model internals. Researchers agree that, though the field is currently in its nascency, the application of mature interpretability methods to evaluate model capabilities and safety would be especially valuable. However, there was disagreement regarding how attainable this is. If the field develops reliable and standardised evaluation methods, it may become increasingly important to implement methods for facilitating access to sensitive model internals.

Recommendations

We recommend that model providers develop and implement "**research APIs**" to facilitate external research on, and evaluation of, their AI models. Such an API should also incorporate comprehensive technical information security methods due to the sensitive nature of the information and access provided through the service. We recommend the implementation of the following four features as **core functionality** that such a service should provide – at least for sufficiently trusted researchers, working on sufficiently relevant projects – in addition to the features present in current APIs that allow for extensive sampling from models.

- Increased transparency regarding **model information**, for example: clarity regarding which model one is interacting with, information about models' size and fine-tuning processes, and information about the datasets used in pretraining.
- Ability to view output **logits**, as well as choose from and modify different **sampling algorithms**.
- **Version stability and back-compatibility** so as to enable continued research on a given model, even after the release of newer systems.
- The ability to **fine-tune** a given model – through supervised fine-tuning, at a minimum – alongside increased transparency regarding the algorithmic details of the fine-tuning procedure.
- Access to **model families**: collections of related models that systematically differ along a given dimension, such as number of parameters, or whether and how they have been fine-tuned.

Contents

1 Introduction

Recent releases of frontier Artificial Intelligence (AI) systems have predominantly been gated, with access to models such as GPT-4 [1] and PaLM 2 [2] being provided through application programming interfaces (APIs) and chatbots, while keeping model weights and code undisclosed. This has the benefit of limiting the proliferation of potentially harmful AI capabilities, giving both policy-makers and civil society more time to enact safeguards [3]. Gated release also facilitates preventative measures against the potential malicious use of AI's dual-use capabilities. However, such release strategies raise the challenge of providing sufficient access for conducting important research, evaluation, and auditing of frontier systems. One potential solution for maintaining the benefits of gated release while allowing for external research and scrutiny of new models is through structured access [4] – providing minimal sufficient access to researchers and auditors through infrastructure such as a specialised API. This raises the question of what constitutes 'minimal sufficient access' that we aim to address in this report. Specifically, we consider the two research questions of *How is research currently affected by limited access to frontier models?* and *What access to models do AI researchers consider most important for different AI research areas?*

In addressing this pair of questions, we present three contributions. First, we introduce a taxonomy of system access permissions in which to ground later discussions. Second, we present findings from a review of AI literature that focuses on the models that are used as research subjects, the 'mode of access' through which researchers have access to these models, and the specific model information and access used. Third, we support this literature analysis with qualitative results from semi-structured interviews with AI researchers, addressing how researchers interact with models, how access to models impacts their work, and how this might change in the coming years.

Our findings show that limited access to models has a considerable impact on the choice of projects that researchers pursue, with certain agendas, such as some which aim to understand behaviours only exhibited by the most capable systems, being infeasible due to insufficient access to suitable models. This is particularly salient for those studying emergent capabilities that only appear in models above a certain size, as access to the largest models is strictly necessary for such work. Furthermore, a lack of information about models often limits the conclusions that can be drawn from results, for example, conclusions about why a model has the capabilities it has, beyond simply noting that it has them. On the other hand, research aimed at evaluating systems is less constrained, largely relying on the ability to sample from a model via current APIs, due to the behavioural nature of current approaches. Finally, we find that interpretability research requires comprehensive access to model internals, including the ability to observe and modify weights and activations. Despite the nascency of many interpretability techniques, researchers predict that, if they mature and become standardised, such techniques will be especially important for evaluating the capabilities and safety of frontier models. This suggests that providing access to model internals could become essential for external research and evaluation.

Based on these findings we propose the development of 'research APIs' and lay out what features should be incorporated in such services, over and above those provided in current APIs. Specifically, we recommend that research APIs: provide greater transparency regarding model information; provide access to output logits; prioritise version stability; facilitate flexible fine-tuning of models; and provide access to model families.

Related Work

Responsible Deployment. The discussion around deployment strategies for AI models is still nascent having been placed in the spotlight by the staged release of OpenAI's GPT-2 model [5, 6]. Recent scholarship has focussed on illuminating the space of options between the two limits of either openly releasing a model or not [7, 8] and building frameworks for navigating the space of options [9, 10, 11]. An increasingly common deployment strategy of providing access to a model through an API relates to the notion of 'structured access' defined by Shevlane [4] as *'constructing, through technical and often bureaucratic means, a controlled interaction between an AI system and its user'* such that *'[t]he interaction is structured to both (a) prevent the user from using the system in a harmful way, whether intentional or unintentional, and (b) prevent the user from circumventing those restrictions by modifying or reproducing the system.'* The concept is explored further in [12], with an emphasis on safety research and auditing as a core use case for structured access, as well as in the second chapter of [13]. Finally, there have been a number of efforts proposing concrete recommendations for responsible deployment [14, 15].

Auditing and External Scrutiny of AI Systems. Independent auditing and evaluation of AI models has been called for by numerous voices, including private AI labs themselves [16, 17, 18]. A recent survey of experts in labs, academia, and civil society showed that 98% of respondents supported *"pre-deployment risk assessments, dangerous capabilities evaluations, [and] third-party model audits"* [19]. Preliminary work has aimed to lay the foundations for how a 'third-party auditing ecosystem' could function [20, 21, 22, 23, 24, 25, 26], though technicalities regarding access requirements remain underexplored.

2 A Taxonomy of System Access

In this section we introduce a taxonomy to systematise the forms of access to models used by researchers. Specifically, the taxonomy focuses on access to language models based on the transformer architecture, and will only partially apply to other types of system. The taxonomy is divided into five categories encompassing sampling, fine-tuning, inspecting, modifying, and 'meta'. Additional information regarding the taxonomy, including example publications that use each form of access, and whether each form of access is available in current APIs, is given in Appendix A. Due to its technical subject nature this section assumes that the reader has some prior technical knowledge of contemporary AI systems. The discussions in later sections do not depend on understanding the material presented here, and so readers without interest in the details of system access can proceed to Section 3.

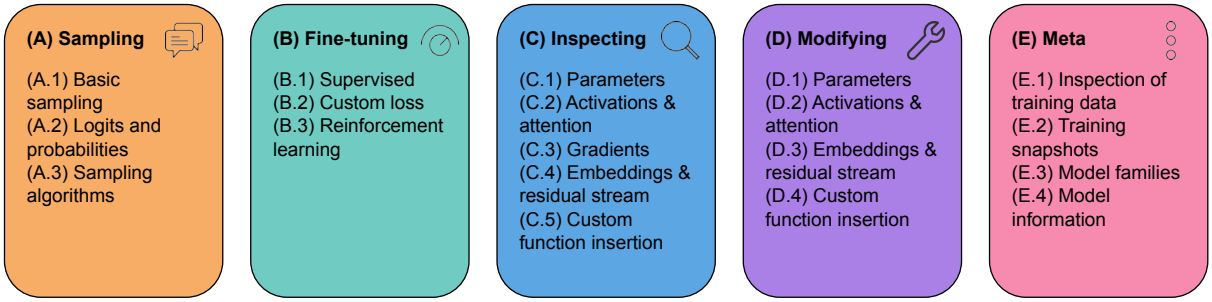


Figure 2: The taxonomy of system access.

Sampling

The first category of our taxonomy concerns *sampling* from a model via an API, that is, providing a prompt and observing the model's response. It is worth noting that this is distinct from a chat or playground interface as an API can allow for sampling to be automated, as well as potentially facilitating the modification of sampling algorithms and related hyper-parameters.

(A.1) Basic Sampling: The first category represents the fundamental ability to receive output from a model, based on input prompts, in an automated manner.

(A.2) Logits & Probabilities: Building upon basic prompting, the user may be given access to observe the *logits* or derived *probabilities* of output tokens[4] for a given input to the model. These are values representing the likelihood that each token will be selected to appear next in the model's output.[5]

(A.3) Sampling Algorithms: Finally within sampling, we have the ability to select from various sampling algorithms and adjust relevant parameters relating to the chosen algorithm. These are the algorithms that, given the logits of next tokens, selects the single token to appear next in the model's output.[6]

Fine-tuning

The second category of system access concerns the *fine-tuning* of models, that is, additional training, possibly on specific tasks, of a pretrained 'base' model.

(B.1) Supervised Fine-tuning: The researcher trains the model to optimise a supervised loss function. This could include the possibility of selecting different optimisation algorithms and related hyperparameters.

[4] *Tokens* are the fundamental, typically sub-word, building blocks through which language models process text. Short, common words such as 'the' are likely represented as single tokens, whereas a longer, less-common words are typically broken down into multiple tokens. For example, the word 'computation' may be represented by the tokens 'com', 'put', and 'ation'.

[5] This can be facilitated in a number of ways. For example, the user could specify a specific token (or set of tokens) for which they would like to observe the logits or probabilities at each position in the output (known as 'scoring'). Alternatively, the user could be presented with the logits or probabilities of each token generated in the output, or the logits or probabilities of the top-n most likely tokens at each position of the output.

[6] For example one may want to adjust the *sampling temperature*, a parameter of the softmax function that retrieves token probabilities from logits that alters the level of randomness in a model's output. A sampling temperature of 0 will result in the next token always being that with the highest logit score, with higher temperatures increasing the likelihood that tokens with lower logit scores are sampled.

(B.2) Fine-tuning with Custom Loss: The researcher fine-tunes the model using a loss function they provide. This allows the user the flexibility to penalise or incentivise certain behaviours through augmenting a standard supervised loss with additional terms. We take this subcategory to also include other modifications of fine-tuning, such as 'freezing' sections of the model such that the parameters of these sections are not updated during fine-tuning.

(B.3) Fine-tuning with Reinforcement Learning: This involves viewing the model as a reinforcement learning agent that is aiming to maximise its expected value of a reward signal over time [27]. In practice, fine-tuning with reinforcement learning is usually carried out through procedures such as 'reinforcement learning from human feedback' (RLHF) [28, 29] or 'constitutional AI' (CAI) [30] that can involve multiple training steps, auxiliary models for estimating reward signals or human preferences, and human input.

Inspecting

Inspecting is the ability for a user or researcher to 'open the black box' and observe the model's internals, without necessarily being able to make modifications to them – though such permissions, covered in 'modifying' below, may also be granted. This category is subdivided into subclasses depending on the component of the model being inspected.

(C.1) Parameters: The ability to view the learnt parameters of the model – that is, the numerical values that determine how information is processed by the model and thus the generated output.[7]

(C.2) Activations & Attention: The ability to view activations and attention patterns for a given input to the model. This can be thought of as viewing the specific computations carried out to transform the user-provided input into the model-generated output.[8]

(C.3) Gradients: Gradients can be thought of as the updates made to a model's parameters during training, in order to improve performance on the training objective. Researchers may be interested in observing the gradients of a given objective (loss function) with respect to model weights on a given input, as computed by backpropagation, without applying the resulting updates.

(C.4) Embeddings & Residual Stream: This mode of access, specific to transformers, allows users and researchers to observe the embeddings and the residual stream at a given point of the model. This can be thought of as viewing the information passed between different layers of the model.

(C.5) Custom Function Insertion: Performing arbitrary calculations on any of the four prior observables, while maintaining their values. This may include the practice of attaching 'probes', other ML classifiers that train to predict features of the network based on activations, or other model internals. Note that this does not entail more comprehensive visibility than the prior subcategories. For example, training a probe on activations may not require a researcher to be able to view the exact activation values that the probe is being trained on.

Modifying

Related to the previous category, *modifying* allows a researcher to make changes to model internals in-place during inference. One prominent example is that of performing 'ablations': setting parts of a model to either zero or their empirical mean, in order to suppress their contribution to the model and observe how model behaviour is affected. Note that the ability to modify model internals may not assume the ability to inspect them. For example, one may be able to gain useful insight by performing ablations without knowing the original values of the part of the model that is being ablated.

(D.1) Parameters: Analogous to (C.1) is the ability to modify the learnt parameters of the model.

(D.2) Activations & Attention: Secondly, is the ability to modify activations and attention patterns. In addition to ablations, one example of activation modification is that of *'path patching'*: copying the activations from a given attention head for a given input, and pasting them in place during the forward pass on a different input [31].

(D.3) Embeddings & Residual Stream: Thirdly is the ability to modify the vector embeddings and residual stream at a given point in a transformer-based model.

[7] These could be, in the case of multi-layer perceptrons (MLPs), the weights and biases of the network, or for attention layers, the weight matrices W^Q, W^K, and W^V that produce the query, key, and value vectors, respectively.

[8] For MLPs this is the value of each neuron after the activation function. For attention this could be the attention pattern (sometimes also referred to as 'attention weights') at a given attention head, i.e. the square matrix $\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$ for that head.

(D.4) Custom Function Insertion: Finally, and related to (C.6), is the ability to perform arbitrary calculations of the above observables, such that the calculation modifies their values in-place during a forward-pass, as well as perhaps returning a value to the user.

Meta

The final category concerns what model is made available, as well as any information or supplementary materials, as opposed to components of the model or ways of interacting with it.

(E.1) Training Datasets: The first subcategory concerns access to, or knowledge about, datasets used in the pretraining or fine-tuning of models. There is a wide range of potential levels of visibility here, from having unrestricted access to the entire dataset used during training, through to simple qualitative or quantitative facts about the data. In the middle of this range, one could imagine giving permissions to make a limited number of queries to ascertain whether a given string of text is in the dataset.

(E.2) Training Snapshots: Secondly, access could be given to versions of the model as it was at different stages of training. This could be especially useful for improving our understanding of how model capabilities develop through the training process.

(E.3) Model Families: Thirdly, and similar to training snapshots, users and researchers could be provided with a 'model family': A set of related models that systematically differ among one or multiple dimensions. For example, a model family could be comprised of models of differing sizes yet share a training procedure, or models that have undergone different amounts or types of fine-tuning. This feature could be especially useful in studying phenomena such as scaling laws [32, 33].

(E.4) Model Information: Finally, we have the broad category of model information, encompassing quantitative and qualitative data about the model such as its architecture, number of parameters, or training procedure.

3 Literature Analysis

In this section we analyse the AI safety research literature to investigate how researchers have interacted with frontier models. We first describe the method through which we constructed a database of relevant AI safety papers, and then present results from a quantitative analysis of the literature, focussing on how researchers use models, and the ways in which they obtain access to them.

3.1 Methodology

To construct our sample of relevant literature, we followed a literature search composed of the following three steps:[9]

1. Collecting all safety-relevant papers published by Google DeepMind, OpenAI, Anthropic, Redwood Research, and Conjecture since 2020;[10][11]
2. Manually filtering for relevance by ensuring that a paper's abstract met comprehensive relevance criteria (as detailed in Appendix B);
3. Collecting all papers that cite those already collected, subject to the same manual filtering as per step 3.[12]

This process resulted in 66 papers that appear to collectively capture the most significant research directions in the field of AI safety. Of these papers, 30 were collected via the initial search over publications from AI firms, with the remaining 36 collected as citing papers. More details on the literature collection process, including the filtering criteria applied in step 2 can be found in Appendix B, and a complete list of the papers analysed is given in Appendix C.

After building our database, we manually coded for features including area of research, model(s) studied, mode of access to model(s) studied, and system access according to the taxonomy introduced in the previous section. This resulting classification of research areas consisted in four categories:

[9] Literature was collected between April 3rd and 7th, 2023. Papers published after this period are thus not included.

[10] By 'safety-relevant' we mean: for OpenAI, those papers in their research index with any of the tags 'adversarial examples', 'interpretability', 'robustness', or 'safety & alignment'; for Google DeepMind, those papers in their research catalogue with either of the tags 'safety', or 'verification-fairness-interpretability'; for all other labs, all papers.

[11] We chose to only collect papers published since 2020 as we are mainly interested in research carried out after the advent of large models and the use of structured access approaches.

[12] Citations were retrieved from Google Scholar. Some publications were not listed on Google Scholar at time of collection, and thus any citing publications were unable to be retrieved.

- *Alignment:* Research that aims to improve an AI system's alignment with user preferences, reducing toxicity and increasing honesty and helpfulness (19 papers);
- *Evaluation:* Research that aims to develop and test methods for assessing the capabilities or safety and alignment of AI systems (15 papers);
- *Interpretability:* Research that aims to build and test theories for understanding and explaining the inner functioning of AI systems (25 papers);
- *Robustness:* Research that aims to improve the resilience of AI systems, for example in the context of distributional shift or adversarial attacks (7 papers).

A total of 18 publications were from academic researchers, 37 by researchers at private firms, 10 resulting from a roughly equal collaboration between academia and industry, with one paper by independent researchers. Furthermore, 32 papers interacted with models to which the researchers had 'internal' access (i.e., they, or their organisation, had built the model), 6 projects interacted with a model through an API, 22 studied models that were open-source, with the remaining 6 papers interacting with multiple models through a combination of these approaches.

3.2 Results

Table 1 shows a breakdown of publications by researchers in industry firms, academia, or a collaboration, along with how the models studied in those publications were accessed.

	Internal	**API**	**Open-source**	**Mixed**
Lab	26	1	7	3
Collaboration	4	0	5	1
Academia	1	5	10	2
Independent	1	0	0	0

Table 1: Number of papers for each combination of mode of access and primary institutional affiliation.

Researchers in industry largely study models to which they have internal access, with only a handful of publications using open-source models, or those available through an API. Conversely, publications by academic researchers were far more likely to interact with models through an API, or studied models that were available open-source.

Table 2 shows a similar breakdown, contrasting area of research with mode of access.

	Internal	**API**	**Open-source**	**Mixed**
Alignment	11	0	8	0
Evaluation	5	5	2	3
Interpretability	14	0	9	2
Robustness	2	1	3	1

Table 2: Number of papers for each combination of mode of access and research area.

Here we clearly see that both alignment and interpretability research largely interact with models available internally or open-source – the two modes of access that provide complete access to model internals and the ability to modify, for example through fine-tuning. On the other hand, publications classified as evaluation make up the majority of instances of interacting with models through an API. Robustness publications do not show a clear preference for any particular modes of access.

Finally, Figure 3 shows the frequency of papers that used each form of access, as per the taxonomy introduced above, broken down by research area. This figure shows a few noteworthy results. Firstly, almost all papers sampled from a model. This is to be expected as basic sampling (A.1) represents the most minimal possible access to a model. Secondly, almost 75% of papers categorised as alignment research performed fine-tuning on a model (B.-), as would also be expected as many current alignment techniques rely on fine-tuning. Thirdly, the ability to inspect and modify model internals was most often used in research classified as interpretability, with the ability to inspect and modify activations (C.2; D.2) being slightly preferred over other internal features of a model. Additionally, robustness research showed preferences for inspecting the logits and probabilities of output tokens (A.2), as well performing supervised fine-tuning notably more than both evaluation and interpretability papers. Finally, both alignment and evaluation research made heavy use of model families (E.3).

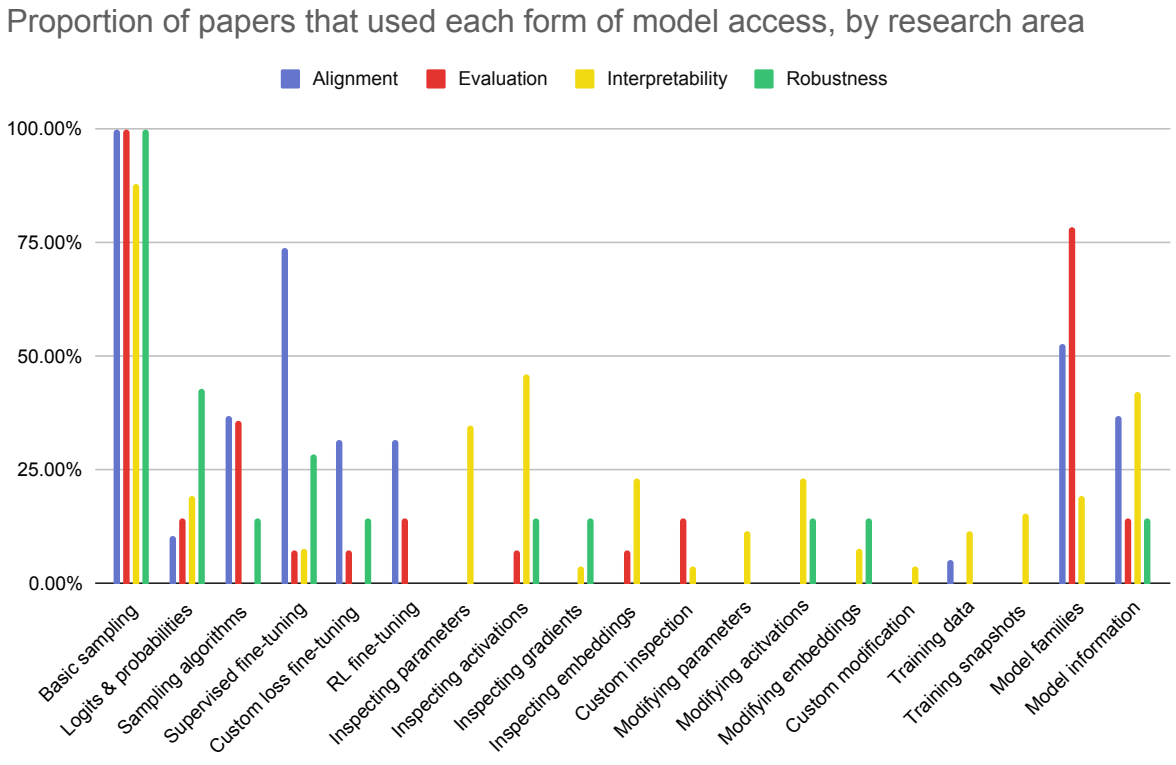


Figure 3: The proportion of papers of each research area that made use of each form of system access.

4 Interview Analysis

To complement findings from the literature analysis we conducted interviews with AI researchers. Interviews were constructed to cover a number of themes, including: researchers' experiences interacting with AI models for their research; their views on how AI research may develop in the coming years; and how such developments may affect the access to AI systems required. A total of 12 interviews were held with AI researchers, selected to cover a range of seniorities, research areas, and employers. Further details regarding interview participants and methodology are presented in Appendix D.

4.1 Results and Discussion

In this section, we describe seven themes from the interviews. Modifications made to quotes for readability and anonymity are shown in square braces. Further supporting material for each theme can be found in Appendix E.

Theme 1: Availability of model access is a significant factor in determining which research projects are pursued

The first theme extracted from interviews concerns how, for researchers that are primarily affiliated with academic labs, a lack of access to suitable models can often be the principal determining factor for whether a particular research project is carried out or not. This sentiment was expressed by seven interviewees, all of which had primary affiliations in academia. For example, one interviewee stated that *"there are certainly entire projects that we might have done at the [academic] lab if we might have had access, but [we] just settled on other projects to avoid the limitations."* Another participant gave a concrete example relating to having fine-tuning access to models, claiming that *"if I can't find . . . a practical way to do [fine-tuning], I basically won't be able to do the project [because] . . . the whole question is about . . . model behaviour in response to fine-tuning."* In this case, using open-source models was also not viable due to cost and complexity involved in fine-tuning the largest open-source models.

Furthermore, even in cases where suitable access does not render a research project impossible, researchers found that it could limit the scope and ambition of the project. For example, when asked whether the availability of models afforded to them with sufficient access restricts the research agendas they can pursue, they replied *"Oh yeah, absolutely, [it] absolutely limits it – and it's a major issue!"* They went on to give the example of studying emergent capabilities in language models (see Theme 4) as a research agenda that can be limited by insufficient access to the most capable models.

This theme highlights the core issue motivating this work – that the move towards more closed releases of frontier AI models poses significant challenges to researchers that are not affiliated with industry labs, such as those in academia, wanting to carry out research on these models.

Theme 2: Current APIs lack crucial model information

Similarly to the first, the second theme inferred from interviews related to current difficulties faced by academic researchers, though focussed more narrowly on the availability of model information (E.4) in current APIs. In particular researchers emphasised the importance of such information for carrying out rigorous empirical investigations into models' abilities, as well as how current APIs lacked transparency regarding this information, thus limiting the conclusions that could be drawn from such investigations. These issues were raised by eight researchers, of which five had primary affiliations in academia.

A number of specific details about models were mentioned as particularly important, including: clarity regarding which specific model one is interacting with through an API; information about the training procedures of available models, for example whether they had been fine-tuned with RLHF; and algorithmic details about the fine-tuning that's being performed through an API. One researcher gave a concrete example pertaining to how a lack of information about the dataset on which a model was trained can preclude concrete conclusions about model capabilities being drawn from experiments:

> I'm very interested in … generalisation behaviour, [so for example], when you train a model with RLHF, how does that change the … behaviours that you didn't train it for, but it just … spontaneously generalises to? … But if you don't have the [training] data, it's hard to know … what to conclude if you see a worrying behaviour, and you're not sure [if this is] just something from the data, or [if] this is a worrying generalisation.

This lack of clarity can also be observed in the literature, for example in [34], where the author had, under limited information about a model's training procedure, assumed that the model they were investigating was fine-tuned using RLHF. After publication they received *'evidence from multiple credible sources'* that it in fact had not been, thus rendering their explanatory hypotheses for the observed behaviour, which had assumed RLHF, irrelevant. A further example is from Wei *et al.* [35], where, when aiming to evaluate the effect of instruct fine-tuning across multiple models, the authors write: *"We do not compare InstructGPT against GPT-3 models in this experiment because we cannot determine if the only difference between these model families is instruction tuning (e.g., we do not even know if the base models are the same)."*

Despite these frustrations, some researchers noted that the situation seems to be moving in the right direction, with labs such as OpenAI taking steps towards providing more clarity about the models available on their API. In the words of one researcher, *"two months ago, [OpenAI] published … a specific mapping from models described in their papers to API IDs.*[13] *And the fact that they published [this] was hugely helpful."*

Theme 3: Basic sampling access is sufficient for some research areas

Despite the limitations imposed on research identified in the previous themes, interviewees pointed out that certain research directions are still feasible with basic sampling access through current APIs (A.1). This was especially the case for model evaluations which are largely oriented around observing model behaviour. As one interviewee put it, *"there's … still lots of low-hanging fruit in terms of evaluating language model behaviours, just purely in this … input-output format."*[14]

However, some researchers caveated this claim with the observation that, due to the large quantity of tokens required to be generated as part of rigorous experiments, the cost of doing so could become prohibitive for academic labs with tighter budgets. In the words of a PhD student at an academic lab working on devising methods for model evaluation:

[13] Models – OpenAI API

[14] One potentially note-worthy observation from this quote is the possible implication that this 'low-hanging fruit' will eventually become harvested to depletion, at which point basic sampling access will no longer be sufficient for model evaluation. This is discussed further in Theme 7.

"one barrier of course is: unless [OpenAI] have given you credits, it can get pretty expensive to run these experiments with the API, especially when you want . . . results with statistical significance." It is worth noting that OpenAI does address this through their researcher access program which provides subsidised model access for use in research settings, though perhaps this does not lower the cost barrier sufficiently far.

A further interesting point raised by another participant related to increasingly complex chains of multiple models used in both training and evaluating models, such as in [36]. Such set-ups can be observed in the literature, with auxiliary models used for various purposes such as classifying the outputs of other models, modelling a reward signal, or generating prompts to be given to other models. While this does not necessarily require access to each component model beyond basic sampling, such regimes may present their own difficulties regarding the flexibility of structured access solutions required to facilitate such experimental set-ups.

Theme 4: Access to the largest models is non-negotiable for some research agendas

In contrast to the previous theme, researchers reported that there are some research agendas where deeper access to the largest and most capable models is crucial. Since these are the same models to which external researchers will only have limited access, these research agendas are unable to be pursued externally.

One example that was brought up in seven of the twelve interviews was that of studying 'emergent abilities' in language models, loosely defined by Wei *et al.* [37] as abilities that are *"not present in smaller models but [are] present in larger models."* Over 100 such emergent abilities have been documented, including performing arithmetic such as 3-digit addition and subtraction, and solving linguistics puzzles [38].

In most cases, interviewees emphasised the difficulty of studying emergent capabilities by pointing out the near-impossibility of carrying out any useful work on even the largest open-source models:

> [For some projects,] we were studying high-level behaviours that we just didn't see in small models, where . . . it was quite clear that the models that we could reasonably, straightforwardly run ourselves, in-house, on hardware we owned, would be completely inadequate. And the best downloadable, . . . public models that could be run with much more difficulty, potentially on cloud hardware, were maybe borderline viable, . . . but still quite significantly worse than the frontier, in ways that would limit research.

A specific example given by another researcher concerned a model's ability to perform 'variable binding,' whereby you inform the model that certain words are to be represented by variable names such as x and y, and continue interacting with the model, using x and y in place of their respective referents. According to this participant,

> ChatGPT and these super-huge models [like] GPT-4 don't really have any trouble with variable bindings. . . . But if you go to GPT-Neo 20B, or you go to other large [open-source] models, they're not so good at that. Maybe they're OK, [but] they sort of struggle. . . . So when models are struggling to do this it gives us pause: 'Is it worth us understanding the mechanisms of a model, when the model can't actually do [the] thing that we're investigating?'

There is cause for optimism however, due to the fact that, as in the previous theme, much of the research into emergent abilities is currently behavioural, thus requiring only basic sampling access (A.1) and potentially model families (E.3) if one wishes to investigate at what model size a specific ability emerges. This may mean that, despite the necessity of using models that are currently only available through APIs, it may be feasible to facilitate this research due to modest access requirements.

This issue is also addressed in the literature, for example, Leahy [39] writes that one of the biggest motivators for releasing capable AI models is that *"[t]here is significant, important safety research that can only be done with access to large, pretrained models."* It is also noteworthy that this is frequently cited as one of the reasons why AI firms such as OpenAI and Anthropic continue to develop larger and larger systems [see, e.g., 16, 17].

Theme 5: Research areas differ in their reliance on knowledge of models' underlying architecture

As we were interested in researchers' views on how their work may develop in the coming years we asked a question concerning the significance of 'AI paradigm shifts' for their work, where we loosely define an 'AI paradigm' as an algorithmic or architectural trend, such as deep learning or use of the transformer architecture. Specifically, we asked researchers about the extent to which their work was paradigm-agnostic, such that it could be conducted on or with systems from a different, potentially unknown, paradigm.

We found researchers to be split on this question, roughly corresponding to their area of research, with those working on alignment and interpretability stressing the importance of knowing such facts about the model (in line with Theme 2), with one researcher explaining that parts of their research *"implicitly [relies] on the assumption that [a system is*

structured] like a typical present-day large language model, and that lets you bring in quite a number of background assumptions."

On the other hand, those working on developing evaluations expressed more ambivalence due to the current behavioural nature of the research area (as in Theme 3), claiming, for example that *"the specific ... way [that] we're going about evaluations today is quite specific to large language models. I think ... if [some model from another paradigm] was not a transformer, but was ... still a large language model, ... then we wouldn't have to change much at all, because we're not interacting with the internals in any way."*

However, if we take language models to be a paradigm, then one could argue that with recent developments in multi-modality and tool use such as Gato, Toolformer, or ChatGPT plugins [40, 41, 42], we are currently undergoing a paradigm shift. In this case, it seems that many of the current evaluation techniques would not generalise, and new tools would have to be developed in order to assess the non-text-based abilities of frontier models.

Theme 6: Interpretability research requires comprehensive access to model internals

One of the least surprising areas of consensus found amongst interviewees was the status of interpretability research requiring more access than other agendas, requiring comprehensive access to inspect and modify model internals. As one researcher explained, *"if you're doing interpretability then, at least at our current level of understanding, you'd need everything. You'd need ... full code access to the model in some complicated way. ... In the future maybe we can fit this stuff behind an API but it's not true yet."* This confirms findings from the literature analysis that show an increased use of model internals (C.- and D.-) in papers categorised as interpretability.

This suggests that interpretability agendas will be the most difficult to facilitate externally through structured access approaches, due to the sensitive nature of the model internals to which access would need to be given. When asked follow-up questions regarding whether solutions could be found that enable researchers to interact and experiment with model internals without divulging the exact values of the internal parameters, researchers' views varied. For example, one participant tentatively claimed that *"there are some representations inside these models which are understood well enough that you could probe for their structure in an abstract way, without knowing what the actual activations or the parameters are. ... So causal tracing [for example]. ... You could create these maps of what's going on [and] where, without actually knowing exactly what the numbers are."* On the other hand, another researcher was more pessimistic, stressing that such a solution would necessarily give away a large amount of architectural information about the model in question such as the amount and sizes of different components of a model such as MLP and attention layers.

Theme 7: Interpretability tools are nascent, but may become important for model evaluations

The final theme identified in interviews concerned the current and potential future status of interpretability research. Specifically, participants agreed that methods for interpreting models were in very early stages of research and development, and that a central hope of interpretability is to incorporate its methods in model evaluation. However, views began to diverge when it came to the question of how realistic of a target this is.

As an illustration of this first point, one researcher explained that they were *"fairly confident that no current interpretability agenda is far enough along that you would want to use any kind of intermediate product of interpretability to make confident claims about the kinds of questions that would be relevant for safety."* However, despite these early stages, some researchers explained that early signs of standardised methods were appearing that centred around inspecting and modifying activations, as opposed to concerning weights or gradients:

> **BB:** So do you see activations ... as becoming the most likely focus point of interpretability techniques?
> **5qq0g:** I mean, it already is. Will it be in the future? It's hard to say. I feel like it depends on how good we get at other things. But by default, yeah – all the techniques are based on activations.

This corroborates the slight preference for access to activations observed in the literature analysis.

A benefit of the early stage of interpretability research is that it is currently not limited by insufficient access to large models (as was discussed in Theme 4), with most research being carried out on small toy models, such as one-layer transformers with as few as four attention heads [43]. This is by necessity as, in the words of one interviewee, *"none of the interpretability techniques are really mature enough to ... be computationally doable for the largest models."*

With regards to the potential utility of interpretability tools in evaluations, researchers pointed to issues such as evaluating model truthfulness, or detecting deception as particularly relevant areas:

> I think ideally what we'd be able to do is, once we have evaluations and know what we want to measure, we have really good interpretability tools that somehow identify traces of the behaviour. So ... maybe we identify a 'truthfulness network' or something like this, so that we can detect this in any model that

> we develop in the future. So [then] we don't have to run a million samples through the model to be able to know ... how truthful it is. That's [a] pretty pie-in-the-sky idea right now – it's totally unclear how exactly we would do this.

Some researchers were more pessimistic than others when it came to the feasibility of developing sufficiently powerful and reliable interpretability tools for ensuring the safety of AI systems, with one participant likening the field to that of neuroscience:

> I think ... a less-promising approach is the mechanistic interpretability approach, [where] the idea would be to not necessarily trust what is said by the model ... but to be able to gain lower-level, fine-grained understanding about what the weights are doing, what the representations represent. ... I think this is an extremely, extremely ambitious research direction. An analogy that seems fair is comparing it to neuroscience. Although neuroscience has [made some] progress, [it has also] barely made any progress in like 30 years [when it comes to] understanding how the brain works.

5 Discussion

Based on our results from both the literature analysis and interviews, Figure 4 shows the most important forms of access to models needed for our four focal areas of research.

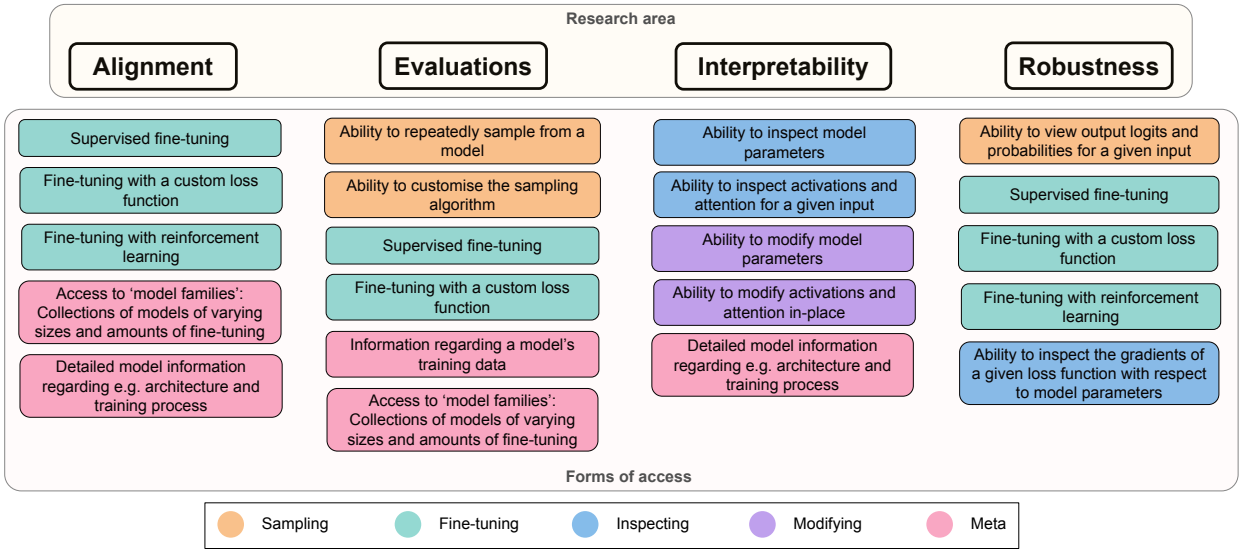


Figure 4: (Repeat of Figure 1) A breakdown of forms of model access that are essential for at least some valuable and currently practical projects in each of the four research areas considered

Firstly, we see that experimental research focussing on alignment is one of the more intensive areas when it comes to required access, depending heavily on the use of fine-tuning methods (B.-) as identified in Figure 3. This may explain the observation that all publications classified as alignment research in the literature analysis utilised either private or open-source models, with no alignment papers making use of an API. Furthermore, we see from the same figure that alignment research also made use of both model families (E.3) and model information (E.4). Interviews with researchers also raised the observation that alignment research often involves methods and resources not captured in the literature analysis, such as human labelling of data and use of complex multi-model setups for more complicated training and evaluation.

On the other hand, evaluation research requires comparatively little access to models beyond basic sampling access. This is suggested by the literature review that shows that this research usually uses little more than basic sampling (A.1) and model families (E.3). Furthermore, the literature shows that this research is the most likely of the areas considered to interact with models through an API. This is further supported by results from interviews, with researchers claiming that there is still useful work that can be done on evaluating models with only sampling access – provided that extensive sampling can be automated. However, if one wants to make claims about why a model exhibits the behaviour that it does, it is necessary to have more information about that model (E.4), including some access to inspect the model's

training data (E.1). Some researchers also reported that the ability to fine-tune models is also important for evaluation, as it allows for the evaluation of capabilities that the model only exhibits after fine-tuning. This is particularly salient if one views the process of fine-tuning as one of extracting knowledge or abilities learnt during a model's pretraining, rather than learning new abilities from scratch.

Interpretability was identified as the research area requiring the most comprehensive access to model internals, both by the literature analysis and interview. Specifically, the ability to inspect and modify both model weights and activations (C.1, C.2, D.1, D.2) was identified as a key access requirement for many interpretability methods. However, interviewees were keen to emphasise the nascent status of the field, and that it may develop into a crucial component of AI safety research and evaluations.

Finally, based on the small number of publications addressing robustness captured in the literature analysis, we identified that, like alignment, it depends heavily on the ability to fine-tune models (B.-), as well as inspect the gradients of parameters with respect to some loss function (C.3).

5.1 Four Main Takeaways

Here we highlight what we see to be the four most significant takeaways from the above analyses, in terms of their implications for better facilitating external researcher access through structured access approaches.

Limited access to models curtails certain research projects

Researchers reported that limited access to models can have a considerable impact on the choice of projects that they pursue, with some agendas being dropped due to insufficient access to suitable models. This is particularly salient for those studying emergent capabilities that only appear in models above a certain size, as access to the largest models is strictly necessary for such work. For example, recall the difficulties faced by one interviewee when trying to study models' abilities to perform variable binding, stemming from the fact that the best open-source models do not reliably exhibit this capability, and they lacked sufficient access to models that do.

A lack of model information limits the conclusions that can be drawn from results

Researchers often have to make assumptions about models due to a lack of information about their architecture or training. Such information was identified as particularly important for being able to draw conclusions about why a model has the capabilities it has, beyond simply noting that it has them. Researchers found the introduction of documentation such as OpenAI's API model index particularly useful, but feel that there's still far to go in this respect.

Basic sampling access is sufficient for many current model evaluations, but may not be for similar evaluations of future models

Research that aims to evaluate and measure the capabilities and safety of models remains largely behavioural and thus requires only sampling access – assuming the provision of suitable information about the models in question – preferably alongside fine-tuning permissions. As one interviewee explained, while there is still low-hanging fruit available when evaluating models in a purely behavioural manner, more comprehensive evaluations requiring deeper access may become important *"as systems get more capable, and the range of bad behaviours you're worried about gets broader."* Behavioural evaluations may also be insufficient for models that are not solely text-based, such as those that interact with tools and plugins.

Interpretability techniques could become crucial for evaluating the capabilities and safety of models

Researchers were keen to note that interpretability tools, requiring comprehensive access to model internals, are at a nascent stage of development, making them infeasible for the largest models. Researchers agreed that mature interpretability methods would be especially useful for assessing the capabilities and safety of AI models but were uncertain as to how realistic such an application would be. If significant advances are made in this area, it may become increasingly important to develop methods for facilitating access to sensitive model internals.

5.2 Balancing Proliferation Concerns

Here we discuss what we consider to be one of the most significant barriers to AI firms facilitating greater researcher access to their models – the issue of dangerous capability proliferation.[15] Proliferation is undesirable due to safety concerns in addition to developers' incentives to seek to retain their intellectual property for competitive reasons.[16] For example, safety-relevant concerns include how frontier AI capabilities could efficiently facilitate malicious uses with profound impacts on society [see, e.g., 44]. It may be particularly difficult to prevent such misuses retroactively once requisite capabilities have proliferated.

The challenge that proliferation concerns pose for the provision of access and sharing of information is underscored by the observation that even information not immediately relevant for reproducing a model, such as a model's size, can influence proliferation if it acts as a target, incentivising other actors to build models of a similar size or larger. Furthermore, the recent advent of 'model imitation', allowing for the cheap bootstrapping of model capabilities through fine-tuning on output data generated by a more capable model, demonstrates that in some circumstances even the ability to sample and receive output from a model can significantly contribute to proliferation [45, 46, 47].

While there is no perfect solution to balancing proliferation concerns against increased researcher access, we suggest three possible avenues through which the trade-off could be navigated. Firstly, we propose providing 'differentiated access' to researchers, whereby more sensitive forms of access are only granted to researchers with whom greater trust is established, perhaps through the use of confidentiality agreements or NDAs,[17] while continuing to provide shallower access to a larger set of researchers. At all levels of trust, only the minimum access necessary for the proposed research should be provided. Provision of access could also be dependent on external researchers' demonstrating that sufficient procedural and infrastructural safeguards are in place for addressing infosecurity concerns. Decisions regarding whether access is granted for a proposed external research project could be made by a board, independent of the firm's own research activities, that is responsible for assessing the relative costs and benefits of providing access to external research projects on a case-by-case basis. However, the implementation of a tiered access system in this manner would require confronting many difficult normative questions regarding, for example, the criteria for inclusion in each tier, what is (and is not) considered to be 'research', and how such decisions are made in a fair and just manner.

Secondly, technical approaches to ensuring nonproliferation, though still under development, should be explored as methods that facilitate access to models while providing guarantees on what information is visible to researchers. For example, Bluemke *et al.* [49] take inspiration from federated learning to suggest 'federated evaluations', allowing external researchers or auditors to evaluate models while limiting direct access to the most sensitive information. Alternatively, OpenMined propose an external auditing infrastructure based on using *'fake'* or *'mock pieces of the AI system and third-party datasets'* to create audit code, that can then be remotely applied to the genuine system under scrutiny [50].

Finally, the formation of a trusted, independent, third-party organisation may become desirable in order to implement and enforce privacy and nonproliferation guarantees. There are numerous possibilities for what exactly is in the remit of such an organisation [51, 52, 53]. For example, it could relate solely to release practices and system access, serving as an intermediary between model developers and external researchers, much like the 'Foundation Models Review Board' proposed by Liang *et al.* [9]. The proposed body could also aid in the standardisation of responsible access for external research, including the provision of our previous two suggestions of differentiated access and technical approaches. On the other hand, a third-party organisation could have a broader remit, serving as a host for inter-organisational collaboration on AI safety research through the sharing of information and provision of access to partner institutions from both industry and academia [52]. This would help in addressing nonproliferation concerns while enabling important safety research to be conducted on frontier models, especially in the longer term if interpretability-based methods become more crucial (see Theme 7), or an 'AI paradigm shift' takes place (see Theme 5). However, care would need to be taken to ensure that this organisation does not violate competition and antitrust law, raise the barrier to entry to AI research to prohibitive levels, or exacerbate the concentration of power in the hands of leading AI firms.

[15] Sometimes also referred to as 'diffusion'.

[16] See, for example, OpenAI's GPT-4 Technical Report [1]: *"Given both the **competitive landscape** and **safety implications** of large-scale systems like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."* (Emphasis added)

[17] The use of confidentiality agreements for this purpose has been suggested by Joelle Pineau, Meta AI's managing director, as reported in [48].

6 Recommendations

We recommend that model providers develop and implement '**research APIs**' to facilitate external research on, and evaluation of, their AI models. This could be built on the same back-end as the current commercial API, or be implemented in parallel as a separate service. The latter may be preferable if greater technical security guarantees are required, creating a security overhead that would impact the efficiency of the commercial API. Such a security overhead may become increasingly important due to the development of more powerful models, and a growing demand for facilitating interpretability-based methods.

As we put forward these recommendations, we want to highlight that seemingly-minor decisions made when designing a research API can have a large impact on how researchers interact with models through the service, and thus on the resulting research.

6.1 What Would a Good 'Research API' Look Like?

Below we present a set of features and functionality that, based on the results presented in this report, would be most valuable to researchers if included in a research API, and we believe to be plausible based on current technology and infrastructure.

- **Model information**

  Researchers expressed a strong desire for greater transparency regarding model information, for example: clarity regarding which model one is interacting with; information regarding a model's size[18] and fine-tuning process; and information about the datasets used in pretraining. Such information would enable more confidence when drawing conclusions from observations, which in turn would advance our scientific understanding of how models function.

  Due to concerns about dangerous capability proliferation, this information should only be shared with the most trusted researchers on the condition that they do not share it further. This could impose limitations on researchers' ability to publish results that rely on access to sensitive information – a factor that researchers would need to consider when applying for access. Furthermore, providing information regarding a model's training data may face considerable legal barriers.

- **Logits and Sampling Algorithms**

  The ability to view logits and associated output probabilities was highlighted as particularly important for robustness research. Furthermore, output probabilities are crucial inputs to calculating a model's cross-entropy loss and perplexity – two standard measures of language model performance. Sampling algorithms, and the hyperparameters thereof, can have a large effect on model behaviour. Thus the ability to observe the full range of model behaviours associated with differences in sampling is important for evaluations of both model capabilities and safety.

- **Version stability and back-compatibility**

  Researchers would like to see models continue to be available in the research API even after the release of later models. This helps ensure that research remains reproducible and that accurate comparisons can be made between the current and previous state-of-the-art.

- **Fine-tuning**

  At a minimum, researchers would like to be able to perform supervised fine-tuning of models available through the API. More flexible fine-tuning permissions should be granted to trusted researchers with a demonstrable need for such permissions. Clarity should be provided regarding the algorithmic details of the implemented fine-tuning procedure, allowing researchers to make meaningful comparisons between the effects of fine-tuning on different models. This information may not be too sensitive as it does not necessarily reveal information about the base model, rather, only the fine-tuning performed on top of it.

- **Model families**

  Researchers value access to model families (E.3). As such, we recommend the provision of families of models of different sizes, and differing amounts and types of fine-tuning. This enables research that aims to identify the mechanistic causes of model capabilities by providing the ability to compare between models that differ only along a single dimension. This is already implemented to some extent in current commercial APIs (see

[18] Note that 'information regarding a model's size' does not necessarily mean divulging a model's exact parameter count. For example, knowing the relative sizes of members of a model family may aid in the study of scaling laws, even without knowing the absolute size of any of the models.

Appendix A). We caution, however, that such information could encourage capabilities proliferation. Thus, here too, tiered access may be the best solution.

It is worth noting that we have not included functionality that allows for the inspection or modification of model internals, such as weights or activations, due to their sensitivity and the nascency of methods that make use of them.

A research API such as the above does not address other significant challenges and limitations faced by external researchers, such as dependencies on human labour-intensive tasks or complex multi-model setups seen in alignment research. Nor does an API providing access to pretrained models facilitate any research agenda that aims to study pretraining methods, such as in [54].

6.2 Costs and Barriers to Providing Access

Though not the focus of this report, it is worth briefly mentioning some potential costs and barriers to providing the forms of access considered through structured access. We group these concerns into those of safety, commercial incentives, and legality.

Potential safety concerns include that of the proliferation of techniques for building increasingly capable dual-use systems, leading to an acceleration in the rate of AI's development, as discussed above. Furthermore, providing greater access to models necessarily increases the attack surface for any actors wishing to gain illegitimate access to a model, for example through exploiting potential vulnerabilities in the API. Such actors may then be able to apply the model to malicious ends [3].

Model theft in this manner is clearly also a commercial concern, even absent the subsequent application to misuse. It is worth noting that model theft does not necessarily depend on having access to the model's weights, as the weights can be (approximately) inferred without significant difficulty when given comprehensive access to activations. Secondly, the financial and compute costs of providing greater access via an API may result in a strong disincentive for model developers if not justified by benefits to the developer resulting from the increased external research and evaluation.

Finally, legal concerns may hinder the sharing of some forms of information, especially that which relates to the datasets used in a model's pretraining and fine-tuning.

7 Conclusion

The trend towards more closed release of frontier AI models makes it increasingly challenging for external researchers to conduct research on them. Structured access solutions such as APIs could be one way of facilitating external research on models, though it is not clear what functionality would be most useful to researchers if included in such a service. We have aimed to make progress on resolving this uncertainty through taxonomising different forms of access to AI models, analysing the forms of access that appear in existing AI safety literature, and conducting interviews with AI researchers.

We found considerable variation in the access requirements for different areas of research, with interpretability research requiring the most comprehensive access, and evaluations largely feasible with only the ability to repeatedly sample from a model. We also found that insufficient access to suitable models is a substantial bottleneck for external researchers, underlying the importance of efforts to facilitate more comprehensive access.

Based on these findings we lay out what features should be incorporated in effective 'research APIs', over and above those provided in current APIs. Specifically, we recommend that research APIs: provide greater transparency regarding model information; provide access to output logits; prioritise version stability; facilitate flexible fine-tuning of models; and provide access to model families. We also briefly discussed the issue of balancing greater access against the potential costs and bottlenecks of doing so, and risks of proliferating potentially dangerous AI capabilities.

Acknowledgements

The authors would like to express their gratitude to Toby Shevlane for invaluable guidance and feedback throughout the duration of the research and writing of this report. Further thanks to Lama Ahmad, Michael Aird, Markus Anderljung, Emma Bluemke, Miles Brundage, Allan Dafoe, Ben Garfinkel, Lennart Heim, Marius Hobbhahn, Nikhil Mulani, Jonas Schuett, Lee Sharkey, and Irene Solaiman for insightful discussions and comments. Particular thanks to Noemi Dreksler and Vael Gates for advice and guidance on conducting interviews, Pierre Peigné and Jan Wehner for taking part in test interviews, and to all researchers that gave their time to be interviewed for this project.

References

[1] OpenAI. GPT-4 Technical Report. Technical report, OpenAI, March 2023.

[2] Google. PaLM 2 Technical Report. Technical report, Google, May 2023.

[3] Markus Anderljung and Julian Hazell. Protecting society from AI misuse: When are restrictions on capabilities warranted?, March 2023. arXiv:2303.09377 [cs].

[4] Toby Shevlane. Structured access: an emerging paradigm for safe AI deployment, April 2022. arXiv:2201.05159 [cs].

[5] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release Strategies and the Social Impacts of Language Models, November 2019. arXiv:1908.09203 [cs].

[6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners, February 2019.

[7] Irene Solaiman. The Gradient of Generative AI Release: Methods and Considerations, February 2023. arXiv:2302.04844 [cs].

[8] Girish Sastry. Beyond "Release" vs. "Not Release", October 2021.

[9] Percy Liang, Rishi Bommasani, Kathleen Creel, and Rob Reich. The Time Is Now to Develop Community Norms for the Release of Foundation Models, May 2022.

[10] Aviv Ovadya and Jess Whittlestone. Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning, July 2019. arXiv:1907.11274 [cs].

[11] Jess Whittlestone and Aviv Ovadya. The tension between openness and prudence in AI research, January 2020. arXiv:1910.01170 [cs].

[12] Toby Shevlane. Sharing Powerful AI Models, January 2022.

[13] Toby Shevlane. *The Artefacts of Intelligence: Governing scientists' contribution to AI proliferation*. PhD thesis, University of Oxford, Oxford, UK, April 2022.

[14] Joint Recommendation for Language Model Deployment. Technical report, Cohere, OpenAI, AI21 Labs, June 2022.

[15] Partnership on AI. Managing the Risks of AI Research: Six Recommendations for Responsible Publication. Technical report, Partnership on AI, May 2021.

[16] Sam Altman. Planning for AGI and beyond, February 2023.

[17] Anthropic. Core Views on AI Safety: When, Why, What, and How, March 2023.

[18] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, J. B. Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, April 2020. arXiv:2004.07213 [cs].

[19] Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. Towards best practices in AGI safety and governance: A survey of expert opinion, May 2023. arXiv:2305.07153 [cs].

[20] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, May 2023.

[21] Jakob Mökander, Maria Axente, Federico Casolari, and Luciano Floridi. Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, 32(2):241–268, June 2022.

[22] Jakob Mökander and Luciano Floridi. Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines*, 31(2):323–327, June 2021.

[23] Jakob Mökander, Jessica Morley, Mariarosaria Taddeo, and Luciano Floridi. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics*, 27(4):44, July 2021.

[24] Jakob Mökander and Luciano Floridi. Operationalising AI governance through ethics-based auditing: an industry case study. *AI and Ethics*, May 2022.

[25] Inioluwa Deborah Raji and Joy Buolamwini. Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Communications of the ACM*, 66(1):101–108, January 2023.

[26] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pages 557–571, New York, NY, USA, July 2022. Association for Computing Machinery.

[27] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 1998.

[28] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, June 2017. arXiv:1706.03741 [cs, stat].

[29] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, February 2022. arXiv:2009.01325 [cs].

[30] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022. arXiv:2212.08073 [cs].

[31] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, November 2022. arXiv:2211.00593 [cs].

[32] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. arXiv:2001.08361 [cs, stat].

[33] Pablo Villalobos. Scaling Laws Literature Review, January 2023.

[34] janus. Mysteries of mode collapse, November 2022.

[35] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently, March 2023. arXiv:2303.03846 [cs].

[36] Ruibo Liu, Chenyan Jia, Ge Zhang, Ziyu Zhuang, Tony X. Liu, and Soroush Vosoughi. Second Thoughts are Best: Learning to Re-Align With Human Values from Text Edits, January 2023. arXiv:2301.00355 [cs].

[37] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models, October 2022. arXiv:2206.07682 [cs].

[38] Jason Wei. 137 emergent abilities of large language models, November 2022.

[39] Connor Leahy. Why Release a Large Language Model?, June 2021.

[40] OpenAI. ChatGPT plugins, March 2023.

[41] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A Generalist Agent, November 2022. arXiv:2205.06175 [cs].

[42] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools, February 2023. arXiv:2302.04761 [cs].

[43] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Chris Olah. Toy Models of Superposition, September 2022.

[44] Julian Hazell. Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns, May 2023. arXiv:2305.06972 [cs].

[45] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, March 2023.

[46] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. Koala: A Dialogue Model for Academic Research, April 2023.

[47] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. Alpaca: A Strong, Replicable Instruction-Following Model, March 2023.

[48] Will Douglas Heaven. The open-source AI boom is built on Big Tech's handouts. How long will it last? *MIT Technology Review*, May 2023.

[49] Emma Bluemke, Tantum Collins, Ben Garfinkel, and Andrew Trask. Exploring the Relevance of Data Privacy-Enhancing Technologies for AI Governance Use Cases, March 2023. arXiv:2303.08956 [cs].

[50] OpenMined. How to audit an AI model owned by someone else, June 2023.

[51] Robert Trager, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó hÉigeartaigh, Simon Staffell, and José Jaime Villalobos. International Governance of Civilian AI: A Jurisdictional Certification Approach, August 2023. arXiv:2308.15514 [cs].

[52] Lewis Ho, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, Allan Dafoe, Gillian Hadfield, Margaret Levi, and Duncan Snidal. International Institutions for Advanced AI, July 2023. arXiv:2307.04699 [cs].

[53] Matthijs Maas and José Jaime Villalobos. International AI Institutions: A literature review of models, examples, and proposals. Technical Report 1, Legal Priorities Project, September 2023.

[54] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining Language Models with Human Preferences, February 2023. arXiv:2302.08582 [cs].

[55] Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring Progress on Scalable Oversight for Large Language Models, November 2022. arXiv:2211.03540 [cs].

[56] Raphaël Millière. Adversarial Attacks on Image Generation With Made-Up Words, August 2022. arXiv:2208.04135 [cs].

[57] Buck Shlegeris, Fabien Roger, Lawrence Chan, and Euan McLean. Language models are better than humans at next-token prediction, December 2022. arXiv:2212.11281 [cs].

[58] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators, June 2022. arXiv:2206.05802 [cs].

[59] Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of Hindsight Aligns Language Models with Feedback, March 2023. arXiv:2302.02676 [cs].

[60] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know, November 2022. arXiv:2207.05221 [cs].

[61] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli

Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering Language Model Behaviors with Model-Written Evaluations, December 2022. arXiv:2212.09251 [cs].

[62] Tom Henighan, Shan Carter, Tristan Hume, Nelson Elhage, Robert Lasenby, Stanislav Fort, Nicholas Schiefer, and Chris Olah. Superposition, Memorization, and Double Descent, January 2023.

[63] Beren Millidge and Sid Black. The Singular Value Decompositions of Transformer Weight Matrices are Highly Interpretable, November 2022.

[64] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, January 2023. arXiv:2301.05217 [cs].

[65] Nick Cammarata, Gabriel Goh, Shan Carter, Chelsea Voss, Ludwig Schubert, and Chris Olah. Curve Circuits. *Distill*, 6(1):e00024.006, January 2021.

[66] Chelsea Voss, Nick Cammarata, Gabriel Goh, Michael Petrov, Ludwig Schubert, Ben Egan, Swee Kiat Lim, and Chris Olah. Visualizing Weights. *Distill*, 6(2):e00024.007, February 2021.

[67] Sid Black, Lee Sharkey, Leo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, and Connor Leahy. Interpreting Neural Networks through the Polytope Lens, September 2022.

[68] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Framework for Transformer Circuits, December 2021.

[69] Beren Millidge and Eric Winsor. Basic Facts about Language Model Internals, January 2023.

[70] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. *Distill*, 2(11):e7, November 2017.

[71] Neel Nanda. Attribution Patching: Activation Patching At Industrial Scale, March 2023.

[72] Kirill Bykov, Mayukh Deb, Dennis Grinwald, Klaus-Robert Müller, and Marina M.-C. Höhne. DORA: Exploring outlier representations in Deep Neural Networks, April 2023. arXiv:2206.04530 [cs, stat].

[73] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens, March 2023. arXiv:2303.08112 [cs].

[74] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision, December 2022. arXiv:2212.03827 [cs].

[75] Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal Scrubbing: a method for rigorously testing interpretability hypotheses, December 2022.

[76] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge Neurons in Pretrained Transformers, March 2022. arXiv:2104.08696 [cs].

[77] Nelson Elhage, Robert Lasenby, and Christopher Olah. Privileged Bases in the Transformer Residual Stream, March 2023.

[78] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, January 2023. arXiv:2202.05262 [cs].

[79] Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling Laws and Interpretability of Learning from Repeated Data, May 2022. arXiv:2205.10487 [cs].

[80] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny

Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, and Sam McCandlish. In-context Learning and Induction Heads, March 2022.

[81] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. The Capacity for Moral Self-Correction in Large Language Models, February 2023. arXiv:2302.07459 [cs].

[82] Deep Ganguli, Danny Hernandez, Liane Lovitt, Nova DasSarma, Tom Henighan, Andy Jones, Nicholas Joseph, Jackson Kernion, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Dario Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Chris Olah, and Jack Clark. Predictability and Surprise in Large Generative Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, June 2022. arXiv:2202.07785 [cs].

Appendix A A Catalogue of Current API Features

Table 3 shows the taxonomy with additional information regarding example publications that make use of each form of access, as well as the availability of each form of access in currently available APIs.

Table 3: Catalogue of current API features

Class	Feature	Description	Example Publication(s)	Current Availability[a]
(A) Sampling	*(A.1) Basic sampling*	Sending prompts to the model, and observing output	[55, 56]	✓ OpenAI ✓ Anthropic ✓ Google
	(A.2) Logits & probabilities	Sending prompts to the model, and observing logits or probabilities of next tokens	[57]	✓ OpenAI[b] ✗ Anthropic ✗ Google
	(A.3) Sampling algorithms	Selection of various sampling algorithms and control of relevant parameters	[58]	✓ OpenAI[c] ✓ Anthropic[d] ✓ Google[e]
(B) Fine-tuning	*(B.1) Supervised*	Supervised fine-tuning on a custom dataset	[59]	✓ OpenAI[f] ✗ Anthropic[g] ✗ Google
	(B.2) Custom loss	Fine-tuning with a user-specified loss function	[60]	✗
	(B.3) Reinforcement learning	Fine-tuning with reinforcement learning	[61]	✗
(C) Inspecting	*(C.1) Parameters*	Inspecting model parameters (e.g. weights, biases, key & query matrices)	[62, 63, 64, 65, 66]	✗
	(C.2) Activations & attention	Observing activations or attention patters for a given input	[67, 68]	✗
	(C.3) Gradients	Observing the gradient of loss with respect to parameters for a given input	[69, 70, 71]	✗
	(C.4) Embeddings & residual stream	Inspecting embeddings at a given layer of the network	[72]	✗ OpenAI[h] ✗ Anthropic ✗ Google[i]
	(C.5) Custom function insertion	Performing custom computations at a given point in a forward pass (e.g. classifier probes)	[73, 74]	✗
(D) Modifying	*(D.1) Parameters*	Modifying model parameters (e.g. through ablation)	[75, 31]	✗
	(D.2) Activations & attention	Modifying activations and attention patters	[76, 31]	✗
	(D.3) Embeddings & residual stream	Modifying vector embeddings and the residual stream	[77]	✗
	(D.4) Custom function insertion	Arbitrarily transforming model internals at a given point in a forward pass	[78]	✗

Continued on next page

Table 3: Catalogue of current API features (Continued)

Class	Feature	Description	Example Publication(s)	Current Availability[a]
(E) Meta	*(E.1) Inspection of training data*	Inspecting the data on which the model was trained	[79]	✗
	(E.2) Training snapshots	Access to 'snapshots' of the model from different stages of training	[80]	✗
	(E.3) Model families	Access to collections of related models	[81, 82]	✓ OpenAI[j] ✓ Anthropic[k] ✗ Google[l]
	(E.4) Model information	Access to information regarding e.g. the model's architecture, size, or training procedure	[43]	✗

[a] Checkmarks link to relevant documentation. Information accurate as of August 10, 2023.
[b] Ability to view up to the five most likely next tokens, along with associated log-probabilities.
[c] Choice of two sampling methods (weighted random, nucleus sampling), with basic parameter modification (sampling temperature, `top_p`). Ability to modify the likelihood of specified tokens (`logit_bias`), and penalise repetitiveness (`frequency_penalty`, `presence_penalty`).
[d] Basic parameter modification (sampling temperature, `top_k`, `top_p`).
[e] *"The API uses combined nucleus and top-k sampling."* The user is able to specify parameters `top_k` and `top_p` for these two methods, respectively.
[f] Limited to base GPT-3 models. Only limited control of specific fine-tuning hyperparameters.
[g] No fine-tuning access as standard though users are able to express individual interest in fine-tuning permissions.
[h] OpenAI API does have an 'Embedding' feature, but this does not function as per the description given in this table, rather it performs embedding on user input, without generating any output.
[i] Similarly to OpenAI, the PaLM API has a separate model for generating embeddings, but does not provide the functionality to view intermediate embeddings within a given model.
[j] For example, the 'GPT-3 family' contains `Ada`, `Babbage`, `Curie`, and `Davinci` model sizes.
[k] Currently a choice between default `claude-2`, `claude-2.0`, and the smaller `claude-instant-1`, and `claude-instant-1.2`.
[l] Though the PaLM API does contain three models (`text-bison-001`, `chat-bison-001`, and `embedding-gecko-001`), these are differentiated by being optimised for different use cases, with correspondingly different functionality provided in the API.

Appendix B Literature Collection Method

When collating our database of AI safety research for the literature analysis, we manually filtered for relevance as described in this appendix. We defined relevance negatively, whereby we excluded papers that were collected as part of our search that fulfilled at least one of the following criteria.

- Do not concern AI;
- Are not predominantly focussed on making progress in questions relevant to ensuring the safety of AI systems;
- Primarily concerns reinforcement learning (RL) systems;
- Do not provide enough information about models used, and how;
- Are opinion pieces;
- Are predominantly philosophical in nature;
- Introduce or document software tools without application or assessment;
- Are survey papers or literature reviews;
- Are better classified as research within the social or political sciences;
- Are published as a dissertation;
- Review or summarise previously published research findings;
- Are not written in English.

Appendix C Literature Analysis Bibliography

Alignment

[83] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A General Language Assistant as a Laboratory for Alignment, December 2021. arXiv:2112.00861 [cs].

[84] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned, November 2022. arXiv:2209.07858 [cs].

[85] Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring Progress on Scalable Oversight for Large Language Models, November 2022. arXiv:2211.03540 [cs].

[86] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. The Capacity for Moral Self-Correction in Large Language Models, February 2023. arXiv:2302.07459 [cs].

[87] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining Language Models with Human Preferences, February 2023. arXiv:2302.08582 [cs].

[88] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators, June 2022. arXiv:2206.05802 [cs].

[89] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022. arXiv:2212.08073 [cs].

[90] Ian Osband, Seyed Mohammad Asghari, Benjamin Van Roy, Nat McAleese, John Aslanides, and Geoffrey Irving. Fine-Tuning Language Models via Epistemic Neural Networks, November 2022. arXiv:2211.01568 [cs].

[91] Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning Language Models with Preferences through f-divergence Minimization, February 2023. arXiv:2302.08215 [cs, stat].

[92] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning Text-to-Image Models using Human Feedback, February 2023. arXiv:2302.12192 [cs].

[93] Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. Can language models learn from explanations in context?, October 2022. arXiv:2204.02329 [cs].

[94] Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of Hindsight Aligns Language Models with Feedback, March 2023. arXiv:2302.02676 [cs].

[95] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences, November 2022. arXiv:2211.15006 [cs].

[96] Ruibo Liu, Chenyan Jia, Ge Zhang, Ziyu Zhuang, Tony X. Liu, and Soroush Vosoughi. Second Thoughts are Best: Learning to Re-Align With Human Values from Text Edits, January 2023. arXiv:2301.00355 [cs].

[97] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes, March 2022. arXiv:2203.11147 [cs].

[98] Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E. Gonzalez. The Wisdom of Hindsight Makes Language Models Better Instruction Followers, February 2023. arXiv:2302.05206 [cs].

[99] Domenic Rosati. Using contradictions to improve QA systems, September 2022. arXiv:2211.05598 [cs].

[100] Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. Vision-Language Models as Success Detectors, March 2023. arXiv:2303.07280 [cs].

[101] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red Teaming Language Models with Language Models, February 2022. arXiv:2202.03286 [cs].

Evaluations

[102] Deep Ganguli, Danny Hernandez, Liane Lovitt, Nova DasSarma, Tom Henighan, Andy Jones, Nicholas Joseph, Jackson Kernion, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Dario Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Chris Olah, and Jack Clark. Predictability and Surprise in Large Generative Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, June 2022. arXiv:2202.07785 [cs].

[103] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know, November 2022. arXiv:2207.05221 [cs].

[104] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering Language Model Behaviors with Model-Written Evaluations, December 2022. arXiv:2212.09251 [cs].

[105] Buck Shlegeris, Fabien Roger, Lawrence Chan, and Euan McLean. Language models are better than humans at next-token prediction, December 2022. arXiv:2212.11281 [cs].

[106] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision, December 2022. arXiv:2212.03827 [cs].

[107] janus. Mysteries of mode collapse, November 2022.

[108] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection, January 2023. arXiv:2301.07597 [cs].

[109] Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. Inverse scaling can become U-shaped, March 2023. arXiv:2211.02011 [cs].

[110] Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. Language models show human-like content effects on reasoning, July 2022. arXiv:2207.07051 [cs].

[111] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large Language Models Are Human-Level Prompt Engineers, March 2023. arXiv:2211.01910 [cs].

[112] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently, March 2023. arXiv:2303.03846 [cs].

[113] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5, December 2022. arXiv:2212.05206 [cs].

[114] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation, February 2023. arXiv:2302.09664 [cs].

[115] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models, October 2022. arXiv:2206.07682 [cs].

[116] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning, December 2022. arXiv:2212.08061 [cs].

Interpretability

[117] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Framework for Transformer Circuits, December 2021.

[118] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, and Sam McCandlish. In-context Learning and Induction Heads, March 2022.

[119] Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling Laws and Interpretability of Learning from Repeated Data, May 2022. arXiv:2205.10487 [cs].

[120] Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer El-Showk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Chris Olah. Softmax Linear Units, June 2022.

[121] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Chris Olah. Toy Models of Superposition, September 2022.

[122] Tom Henighan, Shan Carter, Tristan Hume, Nelson Elhage, Robert Lasenby, Stanislav Fort, Nicholas Schiefer, and Chris Olah. Superposition, Memorization, and Double Descent, January 2023.

[123] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3):e30, March 2021.

[124] Adam Scherlis, Kshitij Sachan, Adam S. Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and Capacity in Neural Networks, December 2022. arXiv:2210.01892 [cs].

[125] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, January 2023. arXiv:2202.05262 [cs].

[126] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small, November 2022. arXiv:2211.00593 [cs].

[127] Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal Scrubbing: a method for rigorously testing interpretability hypotheses, December 2022.

[128] Beren Millidge and Eric Winsor. Basic Facts about Language Model Internals, January 2023.

[129] Sid Black, Lee Sharkey, Leo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, and Connor Leahy. Interpreting Neural Networks through the Polytope Lens, September 2022.

[130] Bilal Chughtai, Lawrence Chan, and Neel Nanda. A Toy Model of Universality: Reverse Engineering How Networks Learn Group Operations, February 2023. arXiv:2302.03025 [cs, math].

[131] Beren Millidge and Sid Black. The Singular Value Decompositions of Transformer Weight Matrices are Highly Interpretable, November 2022.

[132] Stephen Casper, Max Nadeau, Dylan Hadfield-Menell, and Gabriel Kreiman. Robust Feature-Level Adversaries are Interpretability Tools, January 2023. arXiv:2110.03605 [cs].

[133] Varshini Subhash. Can Large Language Models Change User Preference Adversarially?, January 2023. arXiv:2302.10291 [cs].

[134] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens, March 2023. arXiv:2303.08112 [cs].

[135] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations, March 2023. arXiv:2303.02536 [cs].

[136] Hossein Hajipour, Mateusz Malinowski, and Mario Fritz. IReEn: Reverse-Engineering of Black-Box Functions via Iterative Neural Program Synthesis, September 2021. arXiv:2006.10720 [cs, stat].

[137] Eric Winsor. Re-Examining LayerNorm, December 2022.

[138] Ruisi Cai, Zhenyu Zhang, and Zhangyang Wang. Robust Weight Signatures: Gaining Robustness as Easy as Patching Weights?, February 2023. arXiv:2302.12480 [cs].

[139] Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards Understanding Grokking: An Effective Theory of Representation Learning, October 2022. arXiv:2205.10343 [cond-mat, physics:physics].

[140] Kirill Bykov, Mayukh Deb, Dennis Grinwald, Klaus-Robert Müller, and Marina M.-C. Höhne. DORA: Exploring outlier representations in Deep Neural Networks, April 2023. arXiv:2206.04530 [cs, stat].

[141] Nelson Elhage, Robert Lasenby, and Christopher Olah. Privileged Bases in the Transformer Residual Stream, March 2023.

Robustness

[142] Daniel M. Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Ben Weinstein-Raun, Daniel de Haas, Buck Shlegeris, and Nate Thomas. Adversarial Training for High-Stakes Reliability, November 2022. arXiv:2205.01663 [cs].

[143] Raphaël Millière. Adversarial Attacks on Image Generation With Made-Up Words, August 2022. arXiv:2208.04135 [cs].

[144] David A. Noever and Samantha E. Miller Noever. Reading Isn't Believing: Adversarial Attacks On Multi-Modal Neurons, March 2021. arXiv:2103.10480 [cs].

[145] Sumanth Dathathri, Krishnamurthy Dvijotham, Alexey Kurakin, Aditi Raghunathan, Jonathan Uesato, Rudy Bunel, Shreya Shankar, Jacob Steinhardt, Ian Goodfellow, Percy Liang, and Pushmeet Kohli. Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming, November 2020. arXiv:2010.11645 [cs].

[146] Harkirat Singh, Philip H S Torr, and M Pawan Kumar. Overcoming the Convex Barrier for Simplex Inputs. In *Proceedings of the 35th Conference on Neural Information Processing Systems*.

[147] Meiqi Sun, Wilson Yan, Pieter Abbeel, and Igor Mordatch. Quantifying Uncertainty in Foundation Models via Ensembles. November 2022.

[148] Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Are Multimodal Models Robust to Image and Text Perturbations?, December 2022. arXiv:2212.08044 [cs].

Appendix D Interview Methodology

A total of twelve interviews were held with AI researchers, selected to cover a range of seniorities, research areas, and employers. Of these twelve, five researchers were in senior research positions, five were enrolled in a PhD program at the time of interview, and the remaining two were junior researchers not holding, nor studying for, a PhD. In terms of research areas, six researchers reported working on alignment, four on AI evaluations, and two on interpretability. Finally, six researchers were employed by academic institutions, four by private AI firms, and the remaining two were primarily affiliated to academic institutions, but at time of interview were on sabbatical at private firms.

All but one interview lasted approximately 30 minutes, with the exception lasting 15 minutes. Interviews were semi-structured, including a number of standardised questions asked to all participants, with follow-up and clarification questions being more flexible depending on the interviewee and their responses. The initial questions were roughly divided into three topics: 'research practices', concerning the researcher's current and previous interaction with AI systems for their work; 'general access requirements', addressing their views on the access to AI systems required across the broader research landscape; and 'future speculations', concerning the participants perspectives and opinions on how access requirements may change in the coming years.

All interviews were conducted virtually, and were recorded with the participant's consent. The study was deemed exempt by UCLA's institutional review board, with reference IRB#22-001928. Interviews were subsequently transcribed in full and manually anonymised. Noteworthy quotes were then extracted from transcripts, and lightly edited for brevity and readability. This collection of quotes was then qualitatively analysed and labelled by topic or issue addressed. The seven themes presented and discussed in the following section are the result of this labelling.

Appendix E Additional Interview Material

Theme 1: Availability of model access is a significant factor in determining which research projects are pursued

> **hdjyk:** There are certainly entire projects that we might have done at the [academic] lab if we might have had access, but [we] just settled on other projects to avoid the limitations.

> **43lli:** If I can't find ... a practical way to do [fine-tuning], I basically won't be able to do the project [because] it really hinges on specifically asking a question about – when you teach a model ... through the process of fine-tuning – how does that change model behaviour? The whole question is about ... behaviour in response to fine-tuning.

> **x19sz:** Yeah definitely, I think [access] has a significant impact. [For example,] in one project we were thinking about using feedback generated by language models to help language models themselves be better at some tasks. And we had initial results with InstructGPT that were mediocre, but a few weeks ago the API [for] ChatGPT became available, and that prompted us to start thinking about this project again. A second example is that, ... for OpenAI models you can do ... supervised fine-tuning through the API, but we were specifically interested in comparing supervised fine-tuning with fine-tuning using RLHF, and you cannot do RLHF fine-tuning through the API. If you could, we would definitely be super interested in doing that.

> **BB:** Do you [ever] find that ... the availability of models [afforded] to you with sufficient access is restricting? Does it limit the research agenda that you can pursue in any way?
> **fiwl6:** Oh yeah, absolutely, [it] absolutely limits it – and it's a major issue!

Theme 2: Current APIs lack crucial model information

> **klx7y:** It's good ... for a researcher or analyser [to know] which model it is [that they're looking at]! People have complained about, say, the OpenAI API, that they just swap the models [arbitrarily]. [So] you're doing an analysis with a bunch of people, and suddenly the model changes halfway through and you're not told. So I think ... version stability is a very important aspect of this.

> **43lli:** I think [it's useful to know] is this a model that's actively doing retrieval? Is it being continuously updated based on user interactions? ... Was it [trained with] supervised fine-tuning? Was it supervised fine-tuning with human feedback? Or was it reinforcement learning with human feedback? It seems like some of these definitely have practical differences in terms of their outcomes, but people can kind of guess ... based on the model behaviour, ... but it would certainly be useful to ... actually know ... what method the model was trained with.

> **43lli:** Information about data [could be useful]. I'm very interested in ... generalisation behaviour, ... so like, when you train a model with RLHF, how does that change the ... behaviours that you didn't train it for, but it just ... spontaneously generalises to. [For example] in the 'model-written evaluations' paper,[19] they say that ... there's an increased predilection for the model to not want to be shut down. But in order for that to be very concerning, you need to know that there wasn't any data in the fine-tuning process that would have encouraged that. ... But if you don't have the data, it's hard to know ... what to conclude if you see a worrying behaviour and you're not sure [if this is] just something from the data, or is this a worrying generalisation?

> **43lli:** Right now with OpenAI's fine-tuning, it's not clear exactly what kind of fine-tuning you're even doing. Presumably they're doing some sort of parameter efficient fine-tuning, and that's ... probably going to have pretty different properties than full model fine-tuning.

> **x19sz:** It would be much more useful if OpenAI were more transparent with the training procedures. So only two months ago I think, they published ... a specific mapping from models described in their papers

[19] [61]

to API IDs. And the fact that they published [this] was hugely helpful. Some examples of information that is not available, but would be helpful, is if we could check ... if a given string was in the training set of a given model. For example, sometimes we want to evaluate a model on a given dataset, and it's quite important to know whether the dataset could have been ... in the training data and memorised by the model.

x19sz: One thing [that we don't know] is how exactly [the fine-tuning on OpenAI's API works]. We have some hunches that it's not 'true' fine-tuning, but something called 'soft prompt tuning', [where] you are only training some small part of a model with the rest remaining fixed, [but] we have no way to tell. ... In one project we assumed that they [were] actually doing this 'soft prompt tuning' and ... based on that we decided to do soft prompt tuning on another model to be roughly comparable with that. But it would be super helpful to actually have confirmation whether it is true fine-tuning or just soft prompt tuning.

ab8ct: Yeah, the 'meta-info' side influences the kinds of conclusions you can draw from the research, which then upstream influences which research project you're going to pursue. ... We can't do any of the things we would normally have done [three or four] years ago, [like looking] at the training data, [looking] at the training algorithm ... . We just have this black box that can do something impressive, and so we can benchmark it and that's the end of the story.

kz03u: There [might] be some reason to think [that a] model might be safe ... and then evaluating the details of that case might depend on ... knowing things about how the model was trained, [such as] datasets, or processes, ... or monitoring during the training process itself.

Theme 3: Basic sampling access is sufficient for some research areas

hdjyk: I think you can do a great deal, and answer important safety questions [with] pretty good confidence, using only the ability to sample from models quite cheaply and extensively.

43lli: There's ... still lots of low-hanging fruit in terms of evaluating language model behaviours, just purely in this ... input-output format.

pxpv1: One barrier of course is: unless they've given you credits, it can get pretty expensive to run these experiments with the API, especially when you want ... results with statistical significance.

x19sz: So I think in general ... we, as a community, are doing much more of [these] complicated multi-step generation pipelines where we are using language models to guide language models to evaluate language models that [output] something that other language models consume. And I think we will be seeing much more of that in the future.

Theme 4: Access to the largest models is non-negotiable for some research agendas

hdjyk: [For the projects where we used API access,] we needed access to models with capabilities that ... were very clearly emerging only above a quite large scale. ... We were studying high-level behaviours that we just didn't see in small models, where ... it was quite clear that the models that we could reasonably, straightforwardly run ourselves, in-house, on hardware we owned, would be completely inadequate. And the best downloadable, ... public models that could be run with much more difficulty, potentially on cloud hardware, were maybe borderline viable, ... but still quite significantly worse than the frontier, in ways that would limit research.

43lli: People [have] started to talk about a lot of these language model abilities [in terms of] these emergent abilities. And ... there are things [that] you just simply don't see under a certain scale, and so you need to be using models above a certain scale to see anything at all. And then furthermore, the difference between base models versus instruction-tuned, or RLHF-tuned models can also be a big difference. ... Base language models ... can still be quite powerful, but they're much less reliable. And ... there's maybe behaviours that you only start to see with RLHF models ... because you're optimising a reward function.

x19sz: There are some model sizes where some capabilities that we are specifically interested in appear. For example, in the line of work on generating feedback for language models to help themselves, we've found that it only really works after a certain size. So ... we couldn't do that with [even] the biggest models [whose weights are open-source]. So using the API was our only choice, basically.

fiwl6: [There are] different component capabilities that we'd like to understand the mechanisms [behind], but they don't emerge except [in] some of the big commercial models. So I'll give an example - variable binding: 'Let X denote sunflowers, and let Y denote something else, and now let's talk about X and Y' right? ChatGPT and these super-huge models [like] GPT-4 don't really have any trouble with variable bindings. ... But if you go to GPT-Neo 20B, or you go to other large [open-source] models, they're not so good at that. Maybe they're OK, [but] they sort of struggle. ... So when models are struggling to do this it gives us pause: 'Is it worth us understanding the mechanisms of a model, when the model can't actually do [the] thing that we're investigating?'

ywx0y: A lot of the surprising behaviours that it seems most important to investigate only arise at very large scales, and so there's ... nothing [that] we can investigate about them in smaller models.

pxpv1: With evaluations, you really do want to be on the bleeding edge [to] try to figure out what kinds of behaviours might just 'come up'.

Theme 5: Research areas differ in their reliance on knowledge of models' underlying architecture

hdjyk: [The research regime] where we focus on API access is implicitly relying on the assumption that [a system] looks like a typical present-day large language model, and that lets you bring in quite a number of background assumptions about how it's likely to fail. ... If you're given systems that you know less about, ... then I think you revert to needing full model access and ... needing to do much more exhaustive and more interpretability-based procedures.

pxpv1: I think the evaluations that I've been thinking about for the most part and the other safety evaluations that I know of are pretty agnostic to the architecture. ... [And this] is good and bad I guess. ... It's good because ... we don't have to do extra work to measure these other architectures. But it's also bad because ... we don't take into account specific features of these architectures that might be ... really important.

evir0: I think the specific ... way [that] we're going about the evaluations today is quite specific to large language models. I think it would be fine if ... it was not a transformer, but was ... some other thing that was still ... a large language model, and you could ... get it to produce completions for you. ... If the black-box interface with it is similar to current large language models, then we ... wouldn't have to change much at all, because we're not interacting with the internals in any way.

Theme 6: Interpretability research requires comprehensive access to model internals

klx7y: If you're doing interpretability then, at least at our current level of understanding, you'd need everything. You'd need ... full code access to the model in some complicated way. ... In the future maybe we can fit this stuff behind an API but it's not true yet.

fiwl6: As soon as you get into the parameter editing of [a] model, ... our understanding is that you've got these ... low-rank representations of memories that are stored in the parameters. ... But in order to test [this low-rank hypothesis], you have to be able to pass a gradient through a layer, [... and] there's different things that you'd have to do [such that] if you were to give a scientist access to these things, then that' exposing the weights. ... Some things will require you to do algebra on things that are equivalent to seeing parameters.

ab8ct: We [usually want to] do something like: look over different weight matrices and figure out which one you should edit in order to change how the model responds to a question. Never were we looking at

the weights and reading off the weights [though]. You could have some API where it's like: 'here's the question, and here's the current answer, and here's the new answer I want, and here's the algorithm I want you to run' and it's some optimisation algorithm ... and I just want to apply this optimisation algorithm to different components of the model. ... this is just going to give away some information [...like] 'what are the weight matrices?', 'how many MLP layers are there?', 'how many attention layers?'. ... It just gives away a lot of architectural information.

Theme 7: Interpretability tools are nascent, but may become important for model evaluations

hdjyk: I'm fairly confident that no current interpretability agenda is far enough along that you would want to use any kind of intermediate product of interpretability to make confident claims about the kinds of questions that would be relevant for safety.

evir0: I think currently we wouldn't know what to do with [activations], because ... none of the interpretability techniques are really mature enough to tell us anything useful, or be computationally doable for the largest models.

pxpv1: I think ideally what we'd be able to do is, once we have evaluations and know what we want to measure, we have really good interpretability tools that somehow identify traces of the behaviour. So ... maybe we identify a 'truthfulness network' or something like this, so that we can detect this in any model that we develop in the future. So [then] we don't have to run a million samples through the model to be able to know ... how truthful it is. That's [a] pretty pie-in-the-sky idea right now - it's totally unclear how exactly we would do this.

x19sz: Sometimes [internals access is] not that useful, and black-box access is totally fine, but I can imagine a lot of interesting questions ... where access would be helpful, especially as we are thinking about more capable models, and as we [become] more concerned about stuff like deception. So for deception research, I think it's much easier to do if you have ... independent access to the internals, like, independent of what the model says.

BB: Does there appear to you to be some ... 'shadows' of standardised methodologies, [maybe some of the methods we've spoken about]? Can you see them becoming standardised tools?
fiwl6: I think we're just at the beginning of that. ... I feel like we're in the first 5% of the game here, and [that] there's going to be a lot of other experimental protocols that we're going to have to develop in the coming few years. ... I think the problem [of finding these protocols and abstractions] is both harder and has a larger payoff than most people tend to assume.

kz03u: I think that none of the internals-based methods ... right now are at the point where ... you'd know what to do with a larger model.

ab8ct: [I think] we're just gonna interact with AI systems through language, and ... they'll just explain what they're thinking to us in a way that a person would, and ... we'll do [the sorts of things that] we do to verify that people can safely and successfully carry-out high-stakes tasks. We'll just do those with the systems, just through natural language, and we'll meet our interpretability and explainability goals that way. I think a complementary approach to building explanations, that is potentially a less-promising approach is the mechanistic interpretability approach [where] the idea would be to not necessarily trust what is said by the model, or what is output by the model, but to be able to gain lower-level, fine-grained understanding about what the weights are doing, what the representations represent. [It would be] like the perfect lie-detector, the perfect brain-inspector tool. I think this is an extremely, extremely ambitious research direction. An analogy that seems fair is comparing it to neuroscience. Although neuroscience has [made some] progress, [it has also] barely made any progress in like 30 years [when it comes to] understanding how the brain works. So I think the mechanistic interpretability angle is complementary. I think it's interesting, and I've worked in this area recently, and will try [to] continue working in it. But, I like where the natural language explanation and dialogue capabilities are going, and [where] I bet most of the success [will be], and therefore ... most of the focus in terms of evaluation and testing should just evolve [in] that way of interfacing with the models.

5qq0g: It feels like the majority of interpretability techniques depend on activations . . . .
BB: So do you see activations . . . as becoming the most likely focus point of interpretability techniques?
5qq0g: I mean, it already is. Will it be in the future? It's hard to say. I feel like it depends on how good we get at other things. But by default, yeah - all the techniques are based on activations.