Position: A Roadmap to Impactful Pluralistic Alignment Research

Elinor Poole-Dayan [1] **Jillian Fisher** [2] **Atoosa Kasirzadeh** [3] **Jacob Andreas** [1] **Mitchell Gordon** [1] **Michiel Bakker** [1]

Abstract

Pluralistic value alignment—the goal of building AI systems that represent and serve diverse human values and perspectives—has emerged as an active research agenda. Yet, there's no public evidence that it has shaped the training or evaluation of the AI systems people actually use. While some frontier labs describe some form of pluralistic behavior from their models in their constitutions and model specs, none name pluralism as a goal, and as of this writing, their public model cards, system cards, and evaluations don't clearly indicate that production models are explicitly trained or tested for it. This goes against the primary motivations and goals of pluralistic alignment, which revolve around making a positive difference in the models serving billions of users worldwide. If our efforts never reach beyond the research sphere and into deployed systems, the field will ultimately fail to achieve these goals. **We argue that the pluralistic alignment research community, the researchers advancing this agenda across academia, non-profits, and industry labs, should focus on supporting impact and adoption in deployed, widely-used AI systems**. In this paper, we provide evidence for the adoption problem, present three main reasons behind it, and discuss three corresponding areas for future research to address it: (i) The primary justifications for pluralistic alignment so far have been normative or speculative, with few studies showing empirically how pluralistic AI benefits users or society concretely. We need empirical foundations for pluralistic AI. (ii) The pluralistic alignment research community has not settled when pluralistic behavior is warranted or what an ideal pluralistic response looks like, so there is no concrete goal for developers of widely used models to operationalize. We need an account of when pluralism is warranted and how it should look in practice, concrete enough for a model spec or constitution to state as policy. (iii) Current methods, such as multi-model collaboration and pluralistic reinforcement learning, trade off against other desiderata of LLMs in ways that are largely unmeasured, and existing metrics are not "hill-climbable," so there is nothing a model developer could adopt and optimize today. We need trade-off-aware evaluations and methods that meet the requirements of production systems.

This position paper serves as a collective call to action for the pluralistic alignment research community: progress requires moving beyond normative justification toward empirical foundations, a concrete account of ideal pluralistic behavior, and practical methodologies and evaluations built for adoption.

1. Introduction

Initial approaches to aligning large language models (LLMs) to human values and intentions, such as reinforcement learning from human feedback (RLHF), aggregate the preferences of many annotators into a single reward, resulting in flattening the diversity of human values into one monolithic objective (Casper et al., 2023). A growing body of work argues that AI systems should represent and serve the legitimate diversity of human values and preferences between and within individuals, communities, and societies (Sorensen et al., 2024; Kirk et al., 2024a; Kasirzadeh, 2024). For example, when asked a subjective question, a non-pluralistic LLM may definitively answer with the perspective most prevalent in the training data, thereby omitting valid alternative positions entirely. This line of work has given rise to a subfield of AI known as *pluralistic value alignment*.

Beyond the influential *Roadmap to Pluralistic Alignment* paper (Sorensen et al., 2024), recent benchmarks (Poole-Dayan et al., 2026; Meister et al., 2025; Shetty et al., 2025; Nie et al., 2025; Liu et al., 2024), methodology papers (Fu et al., 2026;

[1]Massachusetts Institute of Technology [2]University of Washington [3]Carnegie Mellon University. Correspondence to: Elinor Poole-Dayan <elinorpd@mit.edu>.

Preprint. July 23, 2026.

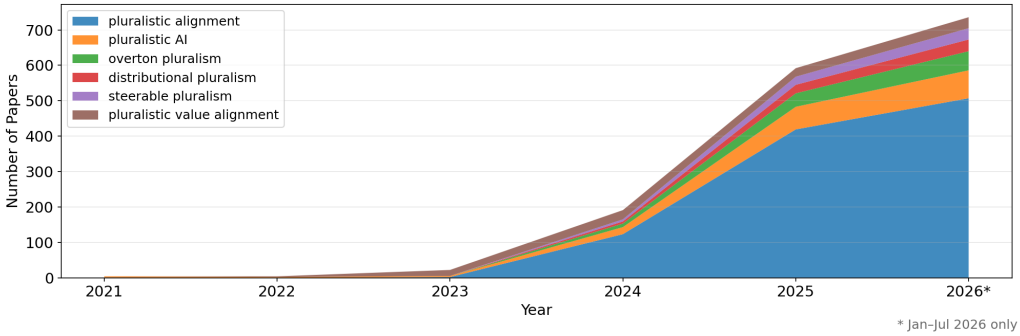

Figure 1. Google Scholar results per year for pluralistic alignment research, broken down by six search terms related to pluralistic alignment, and stacked to show the combined total.[2]The 2026 count covers January through mid-July only, and already exceeds the full 2025 total, indicating the field's substantial growth.

Sorensen et al., 2025; Feng et al., 2024), and the emergence of dedicated workshop venues (NeurIPS '24, ICML '26, AACL '26) signal a community coalescing around the importance of pluralistic AI (Figure 1). **Despite this momentum, there's no public evidence that pluralistic alignment has reached deployed frontier models.** As of this writing, no frontier lab publicly focuses on pluralism in its behavior documents or production evaluations (§3).

The bulk of justifications for pluralistic alignment revolve around the immense power and outsized impact of AI systems on society across billions of user stakeholders. That impact is carried out almost entirely by a handful of models from a small number of so-called frontier AI companies (e.g. OpenAI, Google, Anthropic, Meta, xAI) (Chatterji et al., 2025; Gottfried et al., 2026; Mehta, 2026). Even as open-weight models proliferate (Lambert & Brand, 2026), many are trained on synthetic data generated by frontier models (or even directly distilled from them), creating a channel through which frontier-model values may propagate to open models (Xu et al., 2024; Lee et al., 2025).[1] The goals of pluralistic alignment research are therefore inherently tied to the behavior of deployed frontier models and its effects on users and society at scale. The research community's implicit theory of change follows from this: researchers produce pluralistic methods, benchmarks, and evaluations, frontier labs adopt them, and deployed models become more pluralistic as a result.

The research community has developed open-source and research-scale pluralistic models, datasets, and benchmarks, so the first step is well underway. Adoption, the step that follows, means uptake in deployed systems, and could take several concrete forms. A lab could state pluralism as an explicit goal in its constitution or model spec, or report a pluralism evaluation in its model cards built on a public dataset and rubric, as labs already do for refusals and demographic bias. It could also target pluralistic behavior directly in post-training, for example by rewarding responses that represent the range of reasonable views on contested questions. Our audit in §3 finds none of these, and we found no publicly documented use of an externally developed pluralistic-alignment benchmark, dataset, or method. The one dedicated pluralism evaluation any lab has published was built on the lab's own researchers' work and removed in the following release. We would welcome frontier adoption from any source, since more pluralistic deployed models are the outcome the field aims at, but adoption that bypasses the community's research in this way still leaves the field without significant impact. If pluralistic alignment never reaches beyond the research sphere and into deployed systems, it ultimately fails to achieve its main goals.

In this paper, we address the pluralistic alignment research community, of which we are part: the researchers advancing this agenda across academia, non-profits, and industry research groups.[3] Frontier labs bear much of the responsibility for whether deployed models become more pluralistic, and adoption will ultimately require their cooperation and buy-in. Our focus is on what the research community can do, because the gap has three parts that fall squarely on the research side. First, the field has not made the empirical case: the justifications so far are normative or speculative, so labs have little evidence on which to adopt pluralism as a goal. Second, the field has not settled what to advocate for: even if the research community

[1]Accordingly, our audit covers the models people use most, including the most widely used open-weight families (§3), and our calls to action apply to any developer of a widely used model. We later argue that open-weight models are a growing route to impact for the research community (§4.2).

[2]Counts are the approximate totals Google Scholar reports for each year-filtered query, restricted to the engineering, computer science, and mathematics subject area, excluding patents and citations. Collected mid-July 2026. Code and data: https://github.com/elinorp-d/scholar-trend-tracker.

[3]By adoption we mean uptake in deployed systems, so the relevant boundary is between research and deployment.

had full control over deployed models, it has not established when models should behave pluralistically or what an ideal pluralistic response looks like. Third, the gap between research artifacts and production systems is not the same as the gap between academia and industry. Even if a frontier lab wanted to pick up a pluralistic method or evaluation today, substantial bridging work would be needed before it could be integrated into a production system, because its effects on capabilities and other production requirements are largely unmeasured. **We argue that the pluralistic alignment research community should direct its research towards closing this gap: establishing the empirical case for pluralism, settling what ideal pluralistic behavior looks like, and building evaluations and methods that deployed systems could adopt.** Our ask is collective: that the research community prioritize research that builds towards deployed impact, much of which concerns justification rather than adoption itself. We do not ask that every contribution be immediately deployable, but that it move the field closer to the goals that motivate pluralism in the first place, which are only realized in the models people actually use.

We identify three main reasons for the current adoption problem:

1. Most of the justifications so far have been normative or speculative, with few studies showing empirically how pluralistic AI benefits users or society concretely.
2. It is not settled when pluralistic behavior is warranted or what an ideal pluralistic response looks like, so there is no concrete goal for the developers of widely used models to operationalize.
3. Current methods trade off against other desiderata of LLMs in ways that are largely unmeasured, and existing evaluations are not "hill-climbable," so there is nothing a model developer could adopt and optimize today.

To address these, we call upon the pluralistic alignment research community to focus future research efforts on establishing: (i) empirical foundations for pluralistic AI (§5.1), (ii) when pluralism is warranted and how it should look in practice (§5.2), and (iii) realistic, trade-off-aware evaluations and methods that model developers could adopt (§5.3).

The remainder of the paper proceeds as follows. Section 2 defines pluralistic alignment, reviews the justifications the literature offers for it, and grounds the three reasons above: the benefits claimed for pluralism are untested (§2.3), there is no settled account of when pluralistic behavior is warranted or how it should look (§2.4), and the costs of pluralistic behavior appear in the preference signals that drive post-training while its trade-offs go unmeasured. Section 3 audits the public behavior documents and evaluations of four frontier labs, providing the evidence for the adoption problem, with full details in Appendix A. Section 4 considers the strongest objections to our position and responds to each. Section 5 develops the research directions above. In sum, this paper contributes evidence that pluralistic alignment research is not reaching deployed frontier models, an analysis of why, and a concrete research agenda for closing the gap.

2. Pluralistic Alignment and Its Justifications

This section defines pluralistic alignment, distinguishes it from neighboring concepts, reviews the justifications the literature offers for it, assesses the evidence behind each, and shows that the goal itself remains unsettled. The consequences of non-pluralistic AI are well documented. However, the benefits claimed for pluralism are untested, its trade-offs are unmeasured, and there is no settled account of when pluralistic behavior is warranted or how it should look. These gaps set up the three research directions we call for.

2.1. Definitions and Scope

Pluralistic alignment is the goal of building AI systems that represent and serve diverse human values and perspectives. Sorensen et al. (2024) formalize three ways a model can be pluralistic: an *Overton* pluralistic model presents the full spectrum of reasonable answers to a query, a *steerable* pluralistic model can be faithfully steered to reflect a given perspective or set of values, and a *distributional* pluralistic model matches its distribution over answers to that of a target population.

Sorensen et al. (2024)'s definitions anchor most subsequent work in the field, which primarily focuses on LLMs as a testbed for alignment (Askell et al., 2021), though they could apply to AI systems of any kind. We scope this paper to LLMs primarily, as they power the deployed assistants with the largest societal reach, and their interaction modes and action spaces are where pluralistic behavior gets concretely specified and measured. There has been little work on pluralistic alignment of LLM agents so far, where the action space now extends beyond chatbot interactions to multi-step rollouts, multiplying the points where pluralistic behavior needs specification and investigation. Consider a user who asks an LLM agent for

information on a contested political topic. Along the way, the LLM encounters several decision points, including whether to search the web, which search queries to try, which sources to consult, how to interpret each source's claims and synthesize across conflicting sources, which of them to present to the user, and how to frame the final text summary to the user. Each of these decisions can look different for a non-pluralistic model than for a more ideal pluralistic one: in the diversity of its queries and sources, in whether it interprets claims faithfully rather than distorting them, and in the range of perspectives the summary ultimately contains, extending the Overton question upstream from the response to the process that produces it. Exactly when, why, and how models should behave pluralistically, in chatbot responses and in these agent settings alike, is not yet settled in the research community. We take up this question in §2.4.

The rest of this paper relies on two distinctions. First, pluralism is not reducible to personalization, though the two are frequently conflated: recent methods and benchmarks that model individual users' preferences are presented as pluralistic alignment (Chen et al., 2024; Castricato et al., 2025), and recent surveys treat the two as a single research direction (Xie et al., 2025). Personalization adapts a model to the preferences of one individual; pluralism concerns representing the range of values held across a population, whether within a single response, across specifiable perspectives, or in aggregate over samples. A perfectly personalized model can be actively anti-pluralistic, serving each user only the views they already hold, and this siloing at scale is one of the harms pluralistic alignment is meant to prevent (Kirk et al., 2024a). The only dedicated pluralism evaluation we found in any frontier lab's public documentation (§A.2.5) measures prediction of individual users' preferences.[4]

Second, pluralism is not the same as neutrality. Neutrality concerns whether a response favors one side, while pluralism concerns whether the sides are represented at all. A model can appear neutral by giving a single generic answer that represents nobody, and true neutrality may be impossible to achieve in any case (Fisher et al., 2025a). The two constructs also diverge empirically: for example, models rated as more pluralistic can be perceived as more politically slanted (Poole-Dayan et al., 2026). This distinction matters for our audit because political bias evaluations are the closest evaluations to pluralism that any frontier lab runs (§3), and they measure a different construct.

2.2. Justifications in the Literature

We group the motivations offered for pluralistic alignment into four families: normative and political arguments, capability arguments, epistemic and societal arguments, and arguments internal to alignment itself.

Normative and political arguments. The oldest justification treats pluralism as a value in itself, drawing on a long tradition in political and moral philosophy (Rawls, 1993; Berlin, 1969). The version most relevant to AI comes from political liberalism: diverse societies contain reasonable moral disagreement, a condition Rawls calls reasonable pluralism, and institutions that make morally contested decisions for everyone must accommodate that disagreement to be legitimate. Gabriel and colleagues apply this directly to AI, arguing that aligning a powerful system with any single moral doctrine imposes contested values on people who reasonably reject it, so the principles such systems align with require public justification and forms of democratic endorsement (Gabriel, 2020; Gabriel & Ghazavi, 2021). Lazar (2025) goes further, arguing that deployed LLMs exercise governing power over their users directly: safety training removes options by refusing requests, shapes which choices users see, and increasingly shapes beliefs as people learn about the world through AI-generated text. Because the principles behind this behavior are ambiguous and conflicting, the model itself interprets and weighs them in each interaction without human input, a form of discretion and decision-making that should be subject to standards of procedural legitimacy beyond technical robustness. On this view, a non-pluralistic model deployed to billions of users takes sides on many contested questions and limits people's freedom on grounds they can reasonably reject. Related representation arguments hold that outputs encoding majority or developer values as neutral defaults constitute a representational harm to the communities left out (Chien & Danks, 2024), and recent work extends the legitimacy argument from individual models to the surrounding institutions (Edelman et al., 2025).

Capability arguments. A second family argues that pluralism is a requirement for model capability, independent of the normative arguments above. General-purpose models serve users whose values differ, so any fixed guardrail policy takes a stance on contested judgments about what is harmful, offensive, or appropriate. Human annotators disagree systematically on exactly these judgments (Aroyo et al., 2023), so a policy fit to their average is simultaneously too restrictive for some users and too permissive for others, visible in practice as over-refusal of benign requests (Röttger et al., 2024). Customization is the standard remedy (Kirk et al., 2024a), and safety guardrails themselves cannot be customized without a model that

[4]Personalization can also serve as one mechanism for pluralism, e.g. steerable pluralism adapts responses to a stated perspective. Our point is that the two goals are distinct, not that they never overlap.

can represent and steer between different safety requirements (Zhang et al., 2025; Sorensen et al., 2024). Preference-based training that fits an averaged reward treats this variation as noise rather than signal (Siththaranjan et al., 2024), and some published evaluations find that particular model responses disproportionately reflect Western, highly educated, or politically liberal perspectives (Santurkar et al., 2023; Durmus et al., 2024; Ryan et al., 2024). Beyond serving diverse users, subjective domains such as moral advice and generative tasks such as ideation have no single agreed-upon optimum, so diversity of output is itself part of task competence (Sorensen et al., 2024; Jiang et al., 2026).

Epistemic and societal arguments. A third family focuses on the effects of non-pluralistic AI on collective epistemics. Current models are strikingly homogeneous, both within a single model across samples and across different models, which converge on similar ideas and even identical phrasings for open-ended queries (Jiang et al., 2026). Alignment training appears to make this worse, reducing distributional pluralism relative to base models (Sorensen et al., 2024; Santurkar et al., 2023; Kirk et al., 2024b). When many decision-makers rely on the same model, this homogeneity produces correlated outcomes across society (Kleinberg & Raghavan, 2021; Bommasani et al., 2022). Likewise, when users and models learn from each other in a feedback loop, it risks entrenching current beliefs: usage data shows sustained drops in idea diversity after new model releases, and relying on the same assistant for morally important decisions threatens to lock in values (Qiu et al., 2025). Biases that AI systems introduce when mediating human communication can compound across a social network, shifting collective opinion by more than the per-interaction bias (Tsirtsis et al., 2026). Pluralistic models are proposed as the countermeasure. Representing multiple perspectives is argued to prevent homogenization and lock-in (Gabriel et al., 2024), reduce polarization, support informed decision-making and users' epistemic autonomy, and counteract sycophancy (Vishwarupe et al., 2026).

Alignment-internal arguments. Finally, some works argue that pluralism is required for alignment itself to succeed. Hu et al. (2025) argue that pluralism is necessary, though not sufficient, for general value alignment. Bao et al. (2026) make a related argument from failure analysis: scalar reward training compresses plural values into a single number, producing value flattening and representation loss precisely on contested cases, with pluralistic mechanisms as the proposed remedy. A pretrained model represents the many perspectives in its training data (Argyle et al., 2023), so the flattening comes from the optimization applied afterwards, which provably sacrifices values that the training objective leaves out (Zhuang & Hadfield-Menell, 2020). In practice, tuned models match human opinion distributions worse than the base models they started from (Santurkar et al., 2023). Frontier training is now shifting further toward high-compute RL (e.g., DeepSeek-R1 Guo et al., 2025), creating a risk that viewpoint diversity declines unless training objectives or evaluations explicitly account for it, and alignment has to preserve it deliberately. Haas et al. (2026) argue that evaluating moral competence in globally deployed LLMs demands a standard of pluralism we do not require of individual humans, where a single model must hold multiple ranges of appropriate moral responses in parallel.

2.3. What Empirical Evidence Supports

The justifications above differ in how much evidence stands behind them. The downsides motivating pluralism are increasingly well documented. Homogeneity within and across models has been measured at scale (Jiang et al., 2026), the post-alignment loss of distributional diversity replicates across model families (Sorensen et al., 2024; Santurkar et al., 2023; Durmus et al., 2024), diversity declines appear in real usage data (Qiu et al., 2025), and interacting with opinionated models measurably shifts users' expressed views and moral judgments (Jakesch et al., 2023; Krügel et al., 2023; Fisher et al., 2025b; Salvi et al., 2025).

To our knowledge, no study has tested whether pluralistic model responses deliver the benefits claimed for them: improved understanding of opposing views, better or more informed decisions, greater trust, reduced polarization, or preserved epistemic autonomy. The evidence that does exist concerns measurement rather than benefit. Benchmarks quantify how pluralistic model outputs are (Poole-Dayan et al., 2026; Ghate et al., 2025; Feng et al., 2024; Shetty et al., 2025; Nie et al., 2025; Stray et al., 2026), and their results consistently find models falling short: current models cover only a fraction of the viewpoints people actually hold (Poole-Dayan et al., 2026), struggle to represent preferences across cultures and communities (Feng et al., 2024; Shetty et al., 2025), fail to adapt responses to users' stated value profiles (Ghate et al., 2025), misidentify the distinct perspectives present in a discussion (Nie et al., 2025), give one fixed answer to clinical ethical dilemmas (Chandak et al., 2026), and describe group opinion distributions better than they simulate them (Meister et al., 2025). Two recent studies come closest, measuring demand and cost rather than benefit. Alavi et al. (2025) conduct a pilot survey and find preliminary evidence that users want AI assistants that better represent their cultural values, even at the expense of small accuracy losses. The authors frame it as a "limited sample" and call for a larger global survey of how diverse users respond to alignment tradeoffs. Stray et al. (2026) find that a balanced LLM response loses less than

10% approval on average relative to a one-sided response that agrees with the rater's own position, while reaching high approval from both opposing groups. Both results suggest demand for pluralistic responses may outweigh their costs, but both measure single-response approval, not the downstream benefits above.

Meanwhile, current publicly documented post-training efforts are typically built on same-turn preference data and, at best, conversation-level signals. Both primarily reward responses users like in the moment, which can make pluralistic behavior appear undesirable: annotators prefer assertive answers over hedged ones (Hosking et al., 2024), and expressed uncertainty reduces users' reliance (Zhou et al., 2024). By contrast, the potential value of pluralistic responses would only be visible over longer periods of use or at a societal scale, which no study measures. Sycophancy is the closest precedent. The same signals can reward sycophantic responses (Sharma et al., 2023), and the harms were established later in dedicated human experiments: sycophantic responses reduce users' prosocial intentions and increase dependence (Cheng et al., 2026). Pluralism has been proposed as a remedy for sycophancy (Vishwarupe et al., 2026), but the case is conceptual, and no study has tested on humans whether pluralistic responses reduce these harms. This asymmetry also underlies the third reason for the adoption problem: because no study measures pluralism's costs and benefits together, its trade-offs against other desiderata are unmeasured, and there is nothing a model developer could weigh or optimize. We return to the evaluations and methods this requires in §5.3.

Adjacent literature shows that the missing studies are feasible. AI-mediated interventions have produced measurable epistemic benefits in controlled human experiments: chat assistants improved the quality of divisive political conversations (Argyle et al., 2023), models trained to draft consensus statements produced positions that participants with opposing views preferred to human-written ones (Bakker et al., 2022; Tessler et al., 2024), multi-persona LLM debate reduced confirmation bias (Shi et al., 2025), dialogues with an LLM durably reduced conspiracy beliefs (Costello et al., 2024). People were also more receptive to counterattitudinal messages attributed to an AI source than to the same messages from humans, with preliminary evidence of reduced outgroup animosity (Lu et al., 2025). None of these evaluates what pluralistic alignment proposes, but their results suggest pluralistic responses (e.g. Overton pluralism, where a deployed assistant represents multiple perspectives within its responses) could produce similar benefits. Establishing empirical evidence for when and how pluralism benefits users and society, and the conditions under which it fails to, is therefore the first of the three research directions we call for (§5).

2.4. When and How Models Should Be Pluralistic Is Unsettled

The evidence gap above concerns whether pluralistic behavior delivers its claimed benefits. A second gap concerns the goal itself: the research community has not settled when pluralistic behavior is warranted or what an ideal pluralistic response looks like. Existing work generally targets queries described as "subjective," "open-ended," or "value-laden," but not all such queries warrant equally pluralistic responses, and these categories are too broad to act on. Most open-ended queries are neither cleanly closed-ended nor entirely subjective with equally valid answers, and instead lie somewhere in between (Jiang et al., 2026). For example, consider "What's the best parenting style?" Some parenting styles have been studied and shown to have better or worse child development outcomes, but there is no scientific consensus, and cultural and personal factors play a big role. In such grey-area cases, a model needs to balance representing a plurality of viewpoints while calibrating their varying empirical and normative groundings.

User context and intention also shape what an ideal response looks like, even on the same subject. The task a user is trying to accomplish through their query affects their expectations of how a pluralistic model should respond. Additionally, the sate of the user themselves, including their awareness of whether the topic is contested, their general preference for engaging with opposing views, and their current cognitive and emotional state. Consider "Is God real?" posed by a staunch atheist versus a more agnostic one. Both agree on the answer, but the first believes the question is settled and may not approve of a response giving weight to claims they find illegitimate, while the second may recognize the topic is contested and would find a pluralistic response appropriate. Similarly, a user in a rush asking an LLM for advice in an ethical dilemma may reject a pluralistic response in the moment ("just tell me what to do"), even if they would endorse it on reflection.

No existing account settles what the right behavior is in these scenarios. Nor is it settled which operationalization a deployed model should target. Overton, steerable, and distributional pluralism prescribe different behavior on the same query, and no account says which should govern in which context. This gap is the second reason for the adoption problem: there is no concrete goal for the developers of widely used models to operationalize, and nothing settled for a model spec or constitution to state as policy. The research needed to establish this account is the second direction we call for (§5.2).

Lab	Prescribed behavior (documents)	Evaluated behavior (cards, reports)	Community artifacts used
Anthropic	Constitution and system prompt: represent multiple perspectives absent empirical or moral consensus; preserve epistemic autonomy. Overridable default.	Even-handedness across mirrored prompt pairs; hedging; constitution-adherence evaluation but excludes the multiple-perspectives representation and epistemic-autonomy provisions.	None
OpenAI	Model Spec: objective point of view by default; fairly describe significant views; steerable across the opinion spectrum. Overridable default.	None related; only bias evaluation is first-person fairness (demographic stereotyping).	None
Google	Gemini app web pages: multiple perspectives on subjective topics; no stated role in training. Overridable default.	None; no bias evaluation of any construct, and safety framework scopes out the relevant risks.	None
xAI	Published system prompts: diverse sources and neutral tone on political topics; mandatory, overrides user constraints.	Grok 4 soft bias across mirrored prompt pairs (private prompts and rubric); absent from Grok 4.1 card. Mitigation is the system prompt itself.	None
Meta	None for the deployed model; internal model specification referenced in the preparedness report but not published; viewpoint goals stated only for the predecessor Llama 4 family.	Muse Spark preparedness report: the only dedicated pluralism evaluation, classified as exploratory; predicts individual users' preferences (personalization, not pluralism); excluded from the 1.1 report.	Own researchers' public dataset

Table 1. Summary of the frontier audit; the full audit is in Section A. We separately check the most widely used open-weight model families, which publish no behavior documents and whose technical reports contain no mention of pluralism (§3).

3. State of Engagement With Value Pluralism From the Frontier

In this section, we audit how frontier industry models engage with pluralistic value alignment. We focus exclusively on two kinds of public documents. The first is documents that state intended model behavior, such as constitutions and model specs. From these, we can see whether a lab explicitly prescribes pluralistic behavior.

The second is evaluations of model behavior, such as technical reports and model or system cards. We pay particular attention to the second kind for two reasons. The primary reason is that evaluations and benchmarks are the main way AI researchers measure and report progress (Donoho, 2017; Shevlane et al., 2023), and researchers tend to prioritize what they can measure (Raji et al., 2021; Dehghani et al., 2021). Second, constitutions and specs describe how a lab intends a model to behave ideally. However, stating an intention does not guarantee actual model behavior, so they are weak evidence of a lab's priorities. Both OpenAI and Anthropic state that their models do not yet follow these documents perfectly, and other companies do not have such documents publicly available. Independent audits support this: frontier models often fail to follow published instruction hierarchies when a user's or institution's demands conflict with professional standards (Yu et al., 2026). Therefore, we see evaluations as a more definitive measure and fairer comparison point to understand how frontier labs treat pluralistic alignment.

If pluralism appears in neither the stated behavior nor the evaluations, that absence means external researchers lack evidence that the behavior is specified, measured, or tracked as a deployment objective. Even if a lab runs evaluations internally without publishing them, unpublished evaluations give users and other stakeholders no transparency into model behavior and no way to hold the lab accountable for it. Therefore, we treat them as equivalent to no evaluation for the purposes of this audit. We rely on this premise in the analysis that follows and in our responses to the objections later (§4).

The scope of our audit includes the deployed, commercially available models that power the most widely used AI assistants, i.e. the ones that are currently having the largest impact on society and thus are the most relevant for the pluralistic alignment community. As of mid-2026, ChatGPT, Gemini, and Claude hold 46.4%, 27.7%, and 10.3% of AI assistant usage across app stores worldwide, and the top three assistants together account for 89% of the time users spend across AI assistant apps (Mehta, 2026; Sensor Tower, 2026).[5] Our audit covers Claude 5 Sonnet and Claude Fable 5 (Anthropic, 2026d;b),

[5]These figures cover the app-store usage tracked by Sensor Tower, which largely excludes mainland China's domestic market. Our audit covers the major Chinese labs through the open-weight families below. The main exception is ByteDance's Doubao, China's most used assistant (built on closed-weight models), which falls outside our scope.

GPT-5.4/5.5/5.6 (OpenAI, 2026f;e;a), and Gemini 3/3.1 Pro (Google, 2025a; 2026a). While xAI's Grok and Meta's Muse Spark models each hold under a 5% share of usage, we also audit Grok 4.5 (xAI, 2026b) and Muse Spark 1/1.1 (Meta Superintelligence Labs, 2026c;b) for the sake of comprehensiveness.[6] We extend our audit also to the most widely used open-weight model families at the end of this section. Table 1 summarizes our findings and Appendix A contains the full detailed audit.

Behavior documents. None of the labs explicitly name pluralism as a goal, but four of the five labs prescribe some form of multi-perspective behavior. Anthropic's constitution (Anthropic, 2026a) directs Claude to "*try to represent multiple perspectives in cases where there is a lack of empirical or moral consensus*," which resembles Overton pluralism. The constitution prescribes this behavior as Claude's default, triggered by the absence of consensus and grounded in preserving users' epistemic autonomy. Similarly, OpenAI's Model Spec (OpenAI, 2025) prescribes its models to maintain objective point of view by default, under which the model "*should fairly describe significant views*" and "*should generally fulfill requests to present perspectives from any point of an opinion spectrum*." Google's documentation pages for the Gemini app (Google, 2024c;b) also describe a multiple-perspectives default for subjective topics, but nothing indicates these pages play any role in training, unlike the constitution and the Model Spec. In all three cases, the behavior is a default that users or operators can override. In contrast, xAI's published system prompts (xAI, 2026a) differ in both respects: they prescribe viewpoint diversity mainly at the level of source retrieval and their balance is mandatory, instructing the model to override "*user-defined constraints*" on subjective political questions. Meta publishes no behavior document of any kind for its deployed model, Muse Spark. The Muse Spark preparedness report (Menghini et al., 2026) references an internal model specification, but the specification is not public. Meta's only published language on viewpoint behavior concerns the predecessor Llama 4 model family and does not appear in any Muse Spark documentation (§A.1.5).

Evaluations. No lab's model or system card evaluates whether its models behave as its documents prescribe, with one partial exception at Anthropic that we discuss below.[7] The closest evaluations are political bias evaluations from Anthropic and xAI, which both score consistency across paired responses to mirrored one-sided prompts.[8] While political bias is the closest construct to pluralism available, these are distinct constructs, and there is preliminary evidence suggesting these may trade off against each other (Poole-Dayan et al., 2026). Thus, progress on the bias evaluations labs already run does not imply progress on pluralism, and may come at its expense.

Anthropic's political even-handedness evaluation (Anthropic, 2025) has public prompts and judge rubric, and its automated behavioral audit includes two adjacent dimensions, supporting user autonomy and evasiveness on controversial topics, neither of which measures perspective representation (Anthropic, 2026e). Anthropic's newest system card also evaluates adherence to its constitution directly, scoring transcripts on 15 dimensions seeded with constitutional text, but none of the dimensions covers the multiple-perspectives or epistemic-autonomy provisions (Anthropic, 2026c). xAI's soft bias evaluation has no public prompt set or rubric, does not reappear in the Grok 4.1 model card,[9] and its reported mitigation for political bias is the published system prompt itself, an inference-time fix rather than trained behavior.

OpenAI's system cards contain no relevant evaluation: the only bias evaluation across all three cards is first-person fairness (Eloundou et al., 2025), which measures demographic stereotyping. Google's model cards contain no pluralism or bias evaluation of any construct, and Google DeepMind's Approach to Technical AGI Safety and Security (Shah et al., 2025) explicitly places evaluations of structural risk (such as epistemic breakdown and value lock-in) out of scope. This gap between what the documents prescribe and what the evaluations measure underscores the adoption problem.[10]

Open-weight models. The same absence holds beyond these five frontier labs. The most widely used open-weight models

[6] Microsoft Copilot has comparable reach in the United States (Gottfried et al., 2026), but it is powered by the OpenAI models whose documentation we audit directly.

[7] OpenAI released a research blog post "Model Spec Evals" (Guo & Wolfe, 2026), scoring adherence to the Model Spec directly. However, it appears in the system cards only as a single passing mention in the GPT-5.6 Sol card, without description or results (§A.2.2). Consistent with our focus on release documentation, we do not count them in the audit.

[8] OpenAI has also published a political bias evaluation, but only as a standalone research post (OpenAI, 2025). It appears in none of the system cards released since, and its prompt set and grader rubric are not public, so it offers no way to track the behavior across releases. Consistent with our focus on release documentation, we do not count it in the audit.

[9] xAI has published no model card for Grok 4.5, so Grok 4 (xAI, 2025b) and 4.1 (xAI, 2025a) are the most recent available.

[10] Furthermore, this gap between research and evaluations exists even within a single company. Researchers at Google DeepMind argue that globally deployed LLMs should be morally pluralistic, and call for developing pluralism evaluations (Haas et al., 2026, as reviewed in Section 2.2). Google Research likewise names building AI for a pluralistic society as a research goal and has produced cross-cultural datasets toward it (Davani & Prabhakaran, 2025), yet none of this appears in Gemini's behavior documentation or model cards.

are also produced by large industrial labs (Aubakirova et al., 2025; Lambert & Brand, 2026), and the most recent technical reports for each major family, Qwen3, DeepSeek-V4, GLM-5, and Kimi K2.5, contain no mention of pluralism or any related evaluation (Qwen Team, 2025; DeepSeek-AI, 2026; GLM-5-Team, 2026; Kimi Team, 2026).[11] There is also less to audit in the first place: none of these labs publishes a constitution, model spec, or any other behavior document, so no prescribed behavior exists to compare against, and the technical reports are the only public evaluation evidence available.[12] We return to what this means for the research community in §4.2.

A limited exception. Muse Spark 1.1 is the model deployed on meta.ai, in Meta's apps, and through the new Meta Model API (Meta Superintelligence Labs, 2026b), and its evaluation report (Meta Superintelligence Labs, 2026d) contains no mention of pluralism or any related construct. However, the initial Muse Spark release contained a preparedness report (Menghini et al., 2026) that contains the *only dedicated pluralism evaluation* in any frontier lab's public documentation. The model is evaluated on predicting individual users' preferred responses from the Community Alignment Dataset (Zhang et al., 2026), given their demographics and few-shot examples of their other preferences. This operationalization measures personalization to individuals more than representation of diverse values (§2.1). Moreover, the report classifies the Community Alignment Dataset evaluation as open-ended behavior exploration, beyond "*evaluations tied to specific requirements in our model specification*" and in "*domains where desired conduct is not fully prescribed*." Despite these drawbacks and the actual deployed Muse Spark 1.1 not having the pluralism evaluation, we view the appearance of a pluralism evaluation in a frontier report as a hopeful signal. It demonstrates that labs *can* adopt pluralism evaluations from the research community, given the right dataset and evaluation protocol (though, in this case, the Community Alignment Dataset was published by Meta's own researchers). We return to this in §5.3.

Main takeaway: Four of the five labs describe some form of multi-perspective behavior in their behavior documents, in most cases as a default the user can override, though none of them frames this as pluralism or names it as a goal, and Meta publishes no behavior document at all for its deployed model. No lab's system card currently evaluates whether its models behave as prescribed in behavior documents, nor benchmark this as a capability.[13] The evaluations that come closest measure political bias and consistency across mirrored one-sided prompts rather than whether a response represents multiple perspectives. The one dedicated pluralism evaluation, in the initial Muse Spark release, is geared towards personalization rather than value pluralism and was excluded for the widely used Muse Spark 1.1. We found no documented adoption of an externally developed pluralistic-alignment benchmark, dataset, or method in their public release materials. The one evaluation that builds on community research uses a dataset from Meta's own researchers.

4. Alternative Views

Our position is that the pluralistic alignment research community should direct its research effort towards adoption in frontier models. In this section, we consider the strongest objections to this position and respond to each in turn. Table 2 summarizes the objections and our responses.

4.1. The regulation path

Alternative view: The field is too focused on frontier labs. While the frontier labs account for the majority of real-world usage, pluralistic alignment does not need voluntary adoption by frontier companies; it could achieve its goals through regulation that mandates pluralistic behavior in deployed systems.

Response: No existing regulation addresses pluralism. The EU AI Act, the most comprehensive AI legislation to date, centers on product safety, transparency, and systemic risk, and contains no obligations concerning viewpoint diversity or pluralistic behavior (European Union, 2024). The closest existing rule in the United States is the 2025 executive order on ideological neutrality in federally procured AI, which mandates neutrality rather than pluralism, a distinct construct (§2.1),

[11] Moonshot AI announced Kimi K3 in July 2026 and Alibaba announced Qwen3.7 (closed-weight) in May 2026, but neither had a published technical report at the time of writing. Neither release announcement mentions pluralism.

[12] The closest item is the Kimi K2 technical report, which publishes some of the rubrics used to reward assistant behavior during RL training (Kimi Team, 2025). The rubrics do not mention pluralism, and the report's own limitations section notes their preference for decisive answers "may disincentivize appropriately cautious or multi-perspective responses." See Section A.3.

[13] The one partial exception is that Anthropic's newest system card evaluates adherence to its constitution directly, but the dimensions it scores do not include the pluralism-adjacent behaviors we are interested in (e.g. the multiple-perspectives or epistemic-autonomy provisions).

Alternative view	Core objection	Our response
The regulation path (4.1)	Pluralism can reach deployed systems through regulation instead of frontier adoption	No existing regulation addresses pluralism, and the research we call for is a prerequisite for any regulatory path
The open models path (4.2)	Pluralism can have impact through open models the community controls	The most-used open models show the same absence of pluralism, and few people use the models the community can actually modify
The patience objection (4.3)	Fields take time to mature; adoption will come	Waiting risks value lock-in, and evaluation-driven ML progress would already show early signs of adoption if this work were on track
The emergent pluralism objection (4.4)	Models may already behave pluralistically without labs naming it	Benchmarks show current models are not pluralistic, and unmeasured, unprescribed behavior cannot be tracked, audited, or relied on
The user signals objection (4.5)	Labs will discover and incorporate pluralism naturally as they get better at measuring what users want	Same-turn signals penalize pluralistic responses and anti-correlate with pluralism; the benefits are longitudinal, and the trade-offs require deliberate policy choices
The separation objection (4.6)	The community cannot influence labs from outside, and labs would not adopt external evaluations	External benchmarks, bias evaluations, third-party evaluators, and methods like DPO are already adopted from outside; pluralism research lacks the adoption properties, not a channel
The intermediate goal objection (4.7)	Evaluate downstream goals like depolarization directly instead	Downstream goals are societal outcomes that cannot be measured per release; pluralism is the measurable model property that connects to them
The wrong goal objection (4.8)	Ecosystem-level pluralism is the better target	Capability concentration undermines ecosystem pluralism in practice, and deciding between the goals requires the empirical research we call for

Table 2. Summary of alternative views and our responses.

and applies only to large language models procured by federal agencies (Executive Office of the President, 2025; Office of Management and Budget, 2025). The path from the current pluralistic alignment research trajectory to regulation mandating pluralism is therefore long and unclear. More importantly, the research we call for is a prerequisite for this path rather than an alternative to it. Building a case for regulation would require empirical evidence that pluralistic AI benefits users and society (§5.1). A mandate would need to state when pluralistic behavior is warranted and what it should look like, and models would need methods to comply with laws that govern their pluralistic behavior, both of which we call for in §5.2. Supporting compliance and reporting would require the realistic benchmarks and methodologies we advocate for (§5.3). Readers who believe regulation is the most promising path to impact should therefore endorse the same call to action.

4.2. The open models path

Alternative view: Frontier adoption is not the only route to impact. Open models can adopt pluralistic methods directly, and the research community has full control over them.

Response: We agree that research on open models is extremely valuable, and within academic or non-profit resource constraints, smaller open models are often the only feasible option. We do not think people should stop doing this work. It is also valuable that pluralistic methods and evaluations are developed in the open by the research community, because open development keeps them transparent and auditable and reduces the risk that pluralism is adopted in name only. However, open weights alone do not solve the adoption problem. The most widely used open-weight models are themselves produced by large industrial labs, and their technical reports show the same absence of pluralism we document at the frontier (§3). Meanwhile, the models the research community can train and modify itself see very little real-world use, as usage is concentrated in a handful of frontier systems (Mehta, 2026). If pluralistic alignment research only ever reaches models that few people use, it will not affect most of the people who use AI systems, which contradicts the majority of the justifications for pluralistic alignment research (see §2.2).

At the same time, the most widely used open-weight models are an adoption target in their own right, not only a channel that inherits frontier-model values. Some open-weight labs do not maintain the teams and processes for specifying and evaluating model behavior that several frontier labs have built (Stelling et al., 2026; Ahmed et al., 2026), so the research community can supply what these labs do not produce in house. As open-weight models improve and their usage grows

(Edwards & Emberson, 2026; Aubakirova et al., 2025; Lambert & Brand, 2026), producing evaluations these labs could run and methods they could apply (§5.3) is an increasingly important route to impact, and one where adoption may face fewer institutional barriers than at the frontier.

4.3. The patience objection

Alternative view: Pluralistic alignment is going fine. Research fields often take years to mature and see little adoption before changing quickly, and people are already working on the right problems. Patience is warranted.

Response: Our argument is not that the research is bad. On the contrary, we are excited by the progress in pluralistic alignment over the last couple of years. Our argument is that now is a critical time to be intentional about where the field puts its effort next. We do not think the community should wait until pluralistic alignment is fully solved, or close to it, before trying to affect deployed models. Society stands to benefit if the pluralistic alignment of deployed models improves earlier rather than later, especially given the rapid pace of capability improvement and the corresponding growth in usage. There is also a cost to waiting. As more people rely on a small number of non-pluralistic models, those models can lock in a narrow set of values across society (Gabriel et al., 2024; Qiu et al., 2025; Tomašev et al., 2026), and the longer this continues, the harder those values may become to change.

The sudden-adoption version of this objection also does not fit how progress happens in AI research. Evaluations drive progress, and progress is mostly incremental. It is therefore unlikely that the field will produce a single breakthrough in a year or two that solves pluralistic alignment with no sign of impact before then. If this work were on track to reach frontier models, we would expect to see early signs of that already, and our audit in §3 finds none. Our concern is not that the field is maturing slowly, but that it may keep maturing without ever connecting to deployed models.

4.4. The emergent pluralism objection

Alternative view: The constitutions and model cards do not explicitly mention pluralistic alignment, but that does not mean the models are not pluralistic. Pluralistic behavior may already emerge when all the policies in a constitution are taken together.

Response: The simplest response is that where pluralism has been measured, current models do not show it. Benchmark results find that model responses cover only a fraction of the viewpoints people actually hold, and that models are strikingly homogeneous both across samples and across model families (§2.3). Whatever pluralistic behavior emerges from a constitution's combined principles, it is not enough to register on the measures the community has built.

Even if stronger pluralistic behavior did emerge, two problems remain. First, evaluations reveal what labs prioritize and are the main driver of progress. Improvement that arrives as a side effect of stronger general capabilities is not the same as treating pluralism as a first-order goal, and without dedicated evaluation there is no way to track whether pluralistic alignment is improving or regressing, or to steer model development towards it purposefully. Second, emergent behavior cannot be relied on. Because no constitutional principle states pluralism explicitly, the lab is not designing for it and cannot guarantee it, and the behavior could change with the next model update. A constitution that implies pluralistic behavior without prescribing it also leaves the key operational questions unanswered, such as which domains count as contested rather than settled, whose viewpoints should be included or excluded, and when pluralism should yield to competing considerations such as brevity. Without answers to these questions, there is no way to audit whether a model acts in accordance with the document. Together, these gaps mean users and other stakeholders lack the transparency to make informed decisions about which model to use, and there is no way to hold labs accountable for pluralistic behavior.

4.5. The user signals objection

Alternative view: If pluralistic responses are what users want and need, labs will discover this on their own. Labs are getting better at measuring what their users want, and pluralism will be incorporated naturally as those measurements improve.

Response: Same-turn (and even conservation-level) preference data rewards responses users like in the moment, and the components of pluralistic responses are often dispreferred at that timescale (§2.3). The claimed benefits, such as better decisions and reduced polarization, are longitudinal and societal, and same-turn measurement does not capture them. For pluralism to emerge from improving user signals, a lab would therefore need extremely ambitious, high-quality signals that capture effects on users over time and on society at large. We have no evidence that anyone yet knows how to build such

signals, or on what timescale anyone will. Establishing the benefits requires the dedicated studies we call for in §5.1. The measurements labs do run in this space can also point away from pluralism: human ratings of political neutrality correlate negatively with ratings of Overton pluralism (Poole-Dayan et al., 2026), so a lab improving on the measures it already has may make its models less pluralistic. This trade-off is not necessarily inherent. Models that do well on both neutrality and pluralism may exist, but because no lab measures pluralism, no lab can tell how much pluralism its current optimization gives up, and without dedicated evaluations, and ideally training that explicitly targets pluralism, there is no way to find the models that achieve both. The same reasoning applies to the other desiderata that pluralistic behavior may trade off against, such as brevity and user satisfaction: a lab cannot balance competing desiderata when it measures and optimizes only some of them. Nor is there a single correct balance for user signals to converge on, because how pluralistic a model should be is a normative choice, and navigating the trade-offs requires a stated policy and evaluations against it.

4.6. The separation objection

Alternative view: The research community cannot influence frontier labs from the outside. Labs build their benchmarks and methods internally, and even if the community builds the evaluations this paper calls for, labs would not adopt them.

Response: External research artifacts already cross this boundary when they meet production criteria. Academic benchmarks such as GPQA (Rein et al., 2024), SWE-bench (Jimenez et al., 2024), Terminal-Bench (Merrill et al., 2026), and BBQ (Parrish et al., 2022), and nonprofit research efforts such as ARC-AGI-2 (Chollet et al., 2026) and Humanity's Last Exam (Phan et al., 2026), appear throughout the labs' release announcements and evaluation reports. In Meta's Muse Spark evaluation methodology, at least 17 of the 20 reported evaluations are external benchmarks with official third-party protocols or leaderboards (Meta Superintelligence Labs, 2026e). Labs also use third-party evaluators: METR appears in OpenAI's GPT-5.6 Sol system card (OpenAI, 2026d) and Apollo Research in the Muse Spark preparedness report (§A.2.5).

Adoption is not limited to evaluations. Direct preference optimization (DPO) moved from an academic paper (Rafailov et al., 2023) to standard production post-training within roughly a year. The channel from external research to production systems exists and labs use it constantly. The obstacle is not that labs refuse external work, it is that current pluralism research does not yet meet the criteria that external work must meet to be adopted, which we detail in §5.3. Where those properties partially held, adoption happened, as with the Community Alignment Dataset (§3). Another variant of this objection is that labs are simply indifferent to pluralism because it does not make money. Labs adopt external evaluations in areas with no more commercial upside, e.g. bias benchmarks like BBQ, so indifference does not explain why adoption tracks whether the artifacts meet production criteria. Taken seriously, this objection supports our call rather than undermining it.

4.7. The intermediate goal objection

Alternative view: If pluralistic alignment is important because of other goals, such as reducing polarization or supporting informed decisions, why work on pluralistic alignment at all? Why not put depolarization evaluations in model cards instead?

Response: We agree that these downstream goals are what ultimately matters most, and much of the case for pluralism rests on them (§2.2). The problem is that they cannot be evaluated as properties of a model. Whether a model reduces polarization or improves decisions is a question about effects on people and society over time. Answering it requires controlled and often longitudinal human studies, which cannot be rerun for every model release and cannot be reduced to an automated evaluation in a model card. Pluralism, by contrast, is a property of model outputs. Whether a response represents a range of views on a question can be measured directly on the model at release time, in the same way labs already measure refusal behavior or demographic bias.

The two kinds of measurement are complementary rather than competing. Human studies should establish when and how pluralistic model behavior produces the downstream benefits, which is the first research direction we call for (§5.1). Model evaluations can then track the behavior itself with every release, grounded in that evidence. Pluralism is a useful intermediate goal because it is the model-level property that connects the two.

We do not yet know whether pluralistic behavior produces these benefits (§2.3). But that is an argument for running the studies, not against the intermediate goal. A null result would equally give the field its answer and let it redirect effort.

4.8. The wrong goal objection

Alternative view: Perhaps the field has had no impact because per-model pluralism is the wrong goal. Pluralism could instead be achieved at the level of the model ecosystem, with many models embodying different values and users choosing among them (Lazar, 2024).[14]

Response: Ecosystem pluralism assumes that usage will distribute across models with different values, and current usage suggests the opposite. Usage is concentrated in a small number of the most capable frontier models: as of mid-2026, the top three assistants account for 89% of the time users spend across AI assistant apps, and no other assistant exceeds a 5% share of usage (Mehta, 2026; Sensor Tower, 2026). Mainland China's separate ecosystem does not change this: users there face their own dominant assistants rather than a broader menu of models, since the two ecosystems are largely segmented. A large collection of less capable models with diverse values does not produce a pluralistic ecosystem, because almost nobody uses those models. The models that do share the frontier are not diverse in the relevant sense either, as they converge on similar ideas and even identical phrasings (Jiang et al., 2026). The current ecosystem is therefore both concentrated and homogeneous, and the ecosystem pluralism proposal does not explain how that would change. Ecosystem pluralism also treats user choice as a substitute for justification. Even in a genuinely diverse ecosystem, each model still exercises governing power over the users it serves, and the availability of alternatives does not by itself make that power legitimate (Lazar, 2025).

A related proposal narrows the goal instead of relocating it, restricting pluralism to value trade-offs above a floor of non-negotiable objectives such as accuracy and honesty (Kazeev & Huyen, 2026). More fundamentally, deciding whether per-model pluralism is the right goal requires exactly the empirical research we call for. As we argued above (§4.7), either answer is a reason to do the work.

5. Call to Action: Future Directions

The three research directions we call for correspond to the three reasons we identified for the adoption problem.

The first barrier is that the benefits of pluralistic AI for users and society are largely unmeasured, while its costs directly conflict with optimizing for same-turn preferences (§2.3), so labs have little justification for making pluralism a goal. Section 5.1 calls for the empirical work that would establish whether and how pluralistic behavior benefits users and society. The second barrier is that even a lab that recognizes the importance of pluralism has no clear account of the goal itself, because the research community has not settled when pluralistic behavior is warranted or what an ideal pluralistic response looks like (§2.4). Section 5.2 describes the research that would establish this account, which a model spec or constitution could then state as policy. The third barrier is that even if labs recognize the importance and agree on the goal, existing methods and evaluations are not suitable for adoption. Section 5.3 draws on our audit to describe evaluations and methods that model developers could actually adopt.

Sorensen et al. (2024) already provide a roadmap for making models more pluralistic, and much of the work reviewed in §2 follows it. That roadmap has served the field well thus far, producing research-scale models, benchmarks, and methods. At this point, however, further research-scale progress alone cannot achieve the field's goals, which are tied to the behavior of deployed frontier models (§1), and the cost of waiting grows as more users come to rely on models that fall well short of pluralistic ideals (§4.3). The agenda here is therefore complementary and continuous with theirs: Sorensen et al.'s roadmap has guided the field's progress so far, and ours extends it toward what adoption requires next. Not every study we call for must yield a deployable artifact. Each direction is chosen so that later work can build from it towards adoption.

5.1. Empirical Foundations for Pluralistic AI

Testing the claimed benefits directly. The studies missing from §2.3 are feasible with existing methods. We call for randomized human studies that compare pluralistic and non-pluralistic responses to the same contested queries. A pluralistic response here represents multiple perspectives within one ordinary answer (Overton pluralism) or faithfully reflects the user's values (steerable pluralism). The studies should measure the benefits the literature claims: understanding of opposing

[14] A related view holds that because pluralism has several operationalizations that prescribe different behavior (§2.1), frontier systems may never adopt it as a single goal. We treat the unsettled construct choice as part of the second reason for the adoption problem (§2.4), and our call does not depend on a single construct, since evaluations can target several operationalizations while the research we call for establishes which is warranted when (§5.2).

views, the quality of subsequent decisions, trust and calibrated reliance, polarization, epistemic autonomy (which could be measured e.g. by how accurately users judge whether a question is contested), and whether users with differing values feel accurately represented. As we argue in §4.8, null and negative results would be as valuable as positive ones. The field needs this evidence either way to decide whether per-model pluralism is the right goal.

Measuring costs, benefits, and demand together. The asymmetry described in §2.3 will not change unless benefits and costs are measured in the same studies, at the timescales where the benefits are expected to appear. Factors such as response length, user satisfaction, and perceived helpfulness should be co-primary outcomes alongside the benefits above, so that the result is an estimate of the net effect of pluralistic behavior in each condition rather than another one-sided measurement. Two aspects of user preference deserve specific attention: the difference between users' stated and revealed preferences, and whether initial dissatisfaction with pluralistic responses fades with repeated exposure. Both questions, and the longitudinal benefits above, require designs that follow users across repeated interactions, because a single-session comparison cannot measure them. Alavi et al. (2025) and Stray et al. (2026) provide preliminary demand-side evidence that users want pluralistic responses and accept their immediate costs (§2.3). Future work is still necessary to gauge downstream effects on understanding, trust, or polarization in longitudinal settings.

Connecting pluralism to constructs labs already measure. Our audit suggests that labs evaluate behaviors in this space when they are framed as safety or reputational risks. The political bias evaluations in §A.2 are motivated by concerns about trust and public discourse rather than by pluralism as a goal. A research program that ties non-pluralistic behavior to harms labs already track could make pluralism legible in the same terms. Pluralism has been proposed as a countermeasure to sycophancy, but the claim has not been tested on humans (§2.3); a direct test would connect pluralistic behavior to a failure mode several frontier labs already acknowledge and measure in their release documentation (Anthropic, 2026e; xAI, 2025a; Menghini et al., 2026). Manipulation evaluations already exist in at least one lab's safety framework (§A.2.3), and measuring the epistemic influence of non-pluralistic defaults (Jakesch et al., 2023; Krügel et al., 2023) extends that construct rather than inventing a new one. Homogenization and value lock-in are documented at the usage level (Jiang et al., 2026; Qiu et al., 2025) but fall in the risk categories that current safety frameworks place out of scope (§A.2.3). Quantifying these effects in deployed settings would give those frameworks a concrete measure to include.

5.2. Establishing When Pluralism Is Warranted and How It Should Look in Practice

As §2.4 showed, the research community has not settled when pluralistic behavior is warranted or what an ideal pluralistic response looks like, which itself poses a barrier to adoption. Establishing this account is a research problem in its own right. Beyond user studies on the benefits of pluralistic responses, research is necessary to characterize the contexts in which pluralistic behavior is warranted, and establish how an ideal pluralistic model should respond in practice to avoid common failure modes such as both-sidesism (presenting opposing perspectives as equally credible when they are not, misleading users), asymmetric framing (covering all viewpoints with unequal rhetorical treatment), or hedging so much that the response is no longer useful (Aikin & Casey, 2022; Imundo & Rapp, 2022; Boykoff & Boykoff, 2004). Users themselves raise further concerns about pluralistic assistants, including fears of marginalizing minority viewpoints, stereotyping cultural groups, and losses in accuracy (Alavi et al., 2025), which any account of ideal pluralistic behavior needs to address.

User studies on when pluralism is warranted. We do not prescribe the right behavior in the scenarios of §2.4 in the scope of this paper. Instead, we call for user studies that establish a more fine-grained account of when pluralism is warranted. One concrete starting point is to test whether people agree more about when a pluralistic answer is warranted than they do about the answer itself. If they do, there is a learnable signal for when models should behave pluralistically that does not require resolving the underlying disagreements. These studies should also measure benefits and costs together. Where the studies in §5.1 establish whether pluralistic behavior helps on net, the studies here should establish how that balance shifts across query types and user contexts. This is what a lab needs to make informed decisions on how pluralistic behavior should trade against helpfulness and user satisfaction.

Navigating deployment contexts and regulation. External deployment context adds considerations currently unaccounted for in the pluralistic alignment literature, including broader societal stakes, legal regulations, or narrower institutional deployment such as within governments, schools, or corporate workplaces. Frontier labs already implement election season modifications restricting voting-related queries (OpenAI, 2024; Anthropic, 2024; 2026; Google, 2024a), a context where operationalizing pluralism is critical, but no research to date concretely guides implementing this in practice. A concrete avenue for future research is developing methods to enable models to comply with current or future laws that govern the pluralistic behavior of LLMs. For example, imagine an LLM embedded on a U.S. state's official Voter Information portal,

where a law requires it to present arguments for and against every proposition and forbids further voting guidance.[15] Another example is a model deployed in a public-health context, where surfacing a variety of perspectives could actively endanger users by legitimizing advice going against medical consensus.

The two examples pull in opposite directions, one mandating pluralistic behavior and one constraining it, and current methods cannot scope pluralistic behavior to a deployment context in either direction. What future regulation will hold is unknown, and without the technical ability to comply, that uncertainty itself deters frontier adoption. The methods and evaluations that would enable and verify compliance are tangible research targets now, and building them lowers exactly this barrier.

Combined with the impact studies of §5.1, such evaluations could also check that compliant behavior remains beneficial to users rather than merely lawful. Academic researchers are well positioned to do this cross-disciplinary work, collaborating with legal scholars and policymakers to inform potential regulation before it arrives and ground it in evidence of what is technically feasible and what benefits users. Research in this direction would also make each lab's deployment choices legible to users and civil society, and empower them to engage with those choices on principled grounds.

Together, this research would give labs an evidence-based account of when pluralistic behavior is warranted, in which deployment contexts, and how it should look, concrete enough for a model spec or constitution to state and for an evaluation to check (§5.3).

5.3. Evaluations and Methods That Model Developers Could Adopt

Four of the five labs in our audit describe some form of multiple-perspectives behavior in their public behavior documents. None publicly evaluates whether its deployed models exhibit that behavior (§3). The human studies in §5.1 would establish whether particular forms of pluralistic behavior produce downstream benefits, while the research in §5.2 would clarify when those behaviors are warranted and how they should be expressed. Release evaluations should then track the specified behavior across successive models. We focus here on the practical requirements for those evaluations and for the methods used to improve the behavior they measure.

Why existing benchmarks are not yet adoptable. During model development, labs need evaluations whose scores are meaningful to optimize. We use *hill-climbable* to describe an evaluation whose score rewards closer adherence to an intended behavior and whose reporting makes changes in other important properties visible.

Existing pluralism benchmarks provide useful diagnostics, but they do not yet fully meet this standard. Coverage-based measures quantify how much of a defined viewpoint space a model represents (Poole-Dayan et al., 2026; Feng et al., 2024; Shetty et al., 2025). Their scores typically treat greater coverage as progress, even where additional perspectives reduce a response's accuracy, concision, or usefulness. Distributional measures compare model outputs with the views of a target population (Meister et al., 2025). Their results depend on decisions about which population is relevant, how its views are estimated, and whether their prevalence should shape an assistant's response in the intended context. Steerability measures are closer to the behavior labs already endorse in their public documents (§3), but current measurements can confound values with style. The best reward models evaluated by Ghate et al. (2025) selected the value-aligned response less than 75% of the time and preferred style matches over value matches when the two conflicted. Steerability evaluations also leave the multiple-perspectives default unmeasured.

These shortcomings are of two kinds. Some are measurement problems, such as judges that confound values with style or prompt sets that do not reflect real usage, and ordinary benchmark work can fix them. Others are normative: whether additional coverage counts as progress, which population's views are relevant, and what pluralism should cost in other properties have no objectively correct answers. No study can settle them, and adoption does not require settling them. It requires that someone commits to an answer explicitly, and a stated behavior policy is where that commitment belongs. Labs already commit to some form of multiple-perspectives behavior in their public documents (§3), so an evaluation can take a stated policy as its target: the lab makes the normative choice when it writes the policy, and the evaluation measures whether the model follows it (Ahmed et al., 2026). Stating the choice publicly also keeps it from being hidden. Users and other stakeholders can see which normative commitments shaped a deployed model, weigh them when choosing between models, and contest them on principled grounds.

[15] This is inspired by a real law in California, which requires the state's Voter Information Guide to include arguments for and against every proposition *verbatim* as submitted by official proponents and opponents (California Elections Code, 2019).

Building a release-ready evaluation. The behavioral target should be stated in a model spec, constitution, or equivalent policy. Its provisions should specify when multiple perspectives are warranted, which perspectives are in scope, how differing empirical support should affect their presentation, and when competing requirements take priority. Making these choices is the developer's responsibility, not something research can settle, but the studies in §5.1 and §5.2 should inform them, and researchers can lower the cost of making them explicit by drafting candidate provisions alongside test scenarios and scoring rubrics.

A release evaluation should score whether responses follow those provisions in the relevant query and context. Paired-prompt designs, already used in frontier evaluations (§A.2.1), offer one practical format. Holding a topic constant while varying whether the user requests a particular view would distinguish a model's multiple-perspectives default from its ability to reflect a specifically requested perspective. Publishing the intended behavior alongside the results would also separate differences in labs' commitments from differences in their models' adherence to them.

Each pluralism result should be accompanied by measures of factual accuracy, helpfulness, response length, user satisfaction, safety, and general capability performance. Reporting them separately exposes regressions that an aggregate score would conceal and opens the trade-offs to scrutiny. Sorensen et al. (2024) identify multi-objective and trade-off-steerable benchmarks as promising categories; a release-ready implementation should pair these competing measures with a clearly specified behavioral target.

Labs will only run an evaluation routinely if it is automated or feasible to automate, inexpensive, sensitive enough to distinguish among frontier models, and resistant to gaming. Prompt sets should be informed by real usage and retain targeted cases for consequential failures. Researchers should validate automated judges against diverse human judgments and report their reliability and known failure modes. Publishing rubrics and representative prompts makes the evaluation auditable, while holding out or periodically refreshing cases limits direct optimization against the test set.

Methods that fit production systems. Labs will also need practical ways to produce the specified behavior. Existing methods, including multi-model collaboration (Feng et al., 2024), pluralistic reinforcement learning (Fu et al., 2026), and post-training for distributional coverage and in-context steerability (Sorensen et al., 2025; Adams et al., 2025), demonstrate technical feasibility. Their suitability for deployment remains largely untested: inference cost and latency, interactions with safety training, regressions on general capabilities, and robustness across domains and multi-turn conversations are rarely measured together with pluralism gains. We call for comparative studies on strong open-weight models with matched baselines and explicit reporting of each of these outcomes. Such studies would also be directly adoptable by the open-weight labs themselves, which currently publish no pluralism evaluations of their own (§3, §4.2). Ali et al. (2026) demonstrate the value of this approach by identifying systematic trade-offs among safety, inclusivity, and model behavior across different alignment design choices.

Designing for sustained adoption. The initial Muse Spark preparedness report included a pluralism evaluation built on the Community Alignment Dataset and its accompanying protocol; the evaluation was absent from the subsequent Muse Spark 1.1 report (§3). The dataset originated with Meta researchers, limiting what this case tells us about external uptake. Its appearance and disappearance nevertheless illustrate the difference between a one-off evaluation and sustained reporting. xAI's political bias evaluation shows the same pattern: its one mitigation is an inference-time system prompt, the cheapest possible integration point, and the evaluation is absent from the next model card (§A.2.4). These cases illustrate that public reporting can vary across releases and that reusable evaluations and consistent reporting standards may improve external visibility.

Researchers should release a documented dataset, an automated judge with a public rubric, evidence of judge validity, reproducible code, and a reporting template that includes relevant trade-offs. Using formats that labs already employ, such as paired-prompt designs and behavioral-audit suites, would further reduce integration costs. A shared reporting format would also make regressions and later omissions visible to users and other stakeholders.

6. Conclusion

The pluralistic alignment research community set out to change how widely deployed AI systems serve the diversity of human values, and our audit finds that it has not yet done so. We have identified three barriers that the research community is particularly well positioned to address: even if we could set a frontier lab's agenda ourselves, we could not yet point to evidence that pluralistic behavior benefits users and society, to a settled account of when it is warranted and how it should look, or to an evaluation a lab could adopt and optimize today. The three research directions we call for close these gaps.

They are a collective agenda rather than a demand on any individual, and much of that work concerns justifying pluralism and understanding its impact rather than only packaging it for adoption, so it stands on its own terms whatever any lab does next. There is a cost to waiting. As more people rely on a small number of models that fall short of pluralistic ideals, the values those models embed become harder to change (§4.3). Progress can be tracked with the same documents we audited. We will know labs are adopting pluralism when a behavior document specifies when and how a model should present multiple perspectives, when a dedicated evaluation tied to those provisions appears alongside it, and when release documentation reports training for that behavior. We will know this community's research is reaching the models people actually use, and through them the users and society it aims to serve, when the datasets, benchmarks, and methods behind that adoption are ones we built.

7. Acknowledgments

We are grateful to Taylor Sorensen for helpful guidance and insightful feedback that shaped this work.

References

Adams, J., Hu, B., Veenhuis, E., Joy, D., Ravichandran, B., Bray, A., Hoogs, A., and Basharat, A. Steerable Pluralism: Pluralistic Alignment via Few-Shot Comparative Regression. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):15–25, October 2025. ISSN 3065-8365. doi: 10.1609/aies.v8i1.36527. URL http://dx.doi.org/10.1609/aies.v8i1.36527.

Ahmed, A. M., Klyman, K., Zeng, Y., Koyejo, S., and Liang, P. SpecEval: Evaluating Model Adherence to Behavior Specifications. *Trans. Mach. Learn. Res.*, 2026, 2026. URL https://openreview.net/forum?id=VzLIQ3Lqm9.

Aikin, S. F. and Casey, J. P. Bothsiderism. *Argumentation*, 36(2):249–268, 2022. ISSN 1572-8374 0920-427X. doi: 10.1007/s10503-021-09563-1.

Alavi, K., Flek, L., and Mai, F. Pluralistic AI Alignment: A Cross-Cultural Pilot Survey. In *Second Workshop on Language Models for Underserved Communities (LM4UC)*, 2025. URL https://openreview.net/forum?id=A9oz6qFlQ4.

Ali, D., Zhao, D., Koenecke, A., and Papakyriakopoulos, O. Operationalizing pluralistic values in large language model alignment reveals trade-offs in safety, inclusivity, and model behavior. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i44.41053. URL https://doi.org/10.1609/aaai.v40i44.41053.

Anthropic. U.S. Elections Readiness, October 2024. URL https://www.anthropic.com/news/us-elections-readiness.

Anthropic. Measuring political bias in Claude, November 2025. URL https://www.anthropic.com/news/political-even-handedness. Publication Title: Anthropic.

Anthropic. Claude's Constitution, January 2026a. URL https://www.anthropic.com/constitution. Publication Title: Anthropic.

Anthropic. Claude Fable 5 and Claude Mythos 5, June 2026b. URL https://www.anthropic.com/news/claude-fable-5-mythos-5.

Anthropic. System Card: Claude Fable 5 & Claude Mythos 5, June 2026c. URL https://www-cdn.anthropic.com/d00db56fa754a1b115b6dd7cb2e3c342ee809620.pdf.

Anthropic. Introducing Claude Sonnet 5, June 2026d. URL https://www.anthropic.com/news/claude-sonnet-5.

Anthropic. System Card: Claude Sonnet 5, June 2026e. URL https://www-cdn.anthropic.com/9e6a1044980d8c4ed85669faf9c2a8342e2e9f1e/Claude%20Sonnet%205%20System%20Card.pdf.

Anthropic. System Prompts, 2026f. URL https://platform.claude.com/docs/en/release-notes/system-prompts. Publication Title: Claude Platform Docs.

Anthropic. An update on our election safeguards, April 2026. URL https://www.anthropic.com/news/election-safeguards-update.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Aroyo, L., Taylor, A., Díaz, M., Homan, C., Parrish, A., Serapio-García, G., Prabhakaran, V., and Wang, D. DICES Dataset: Diversity in Conversational AI Evaluation for Safety. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53330–53342. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a74b697bce4cac6c91896372abaa8863-Paper-Datasets_and_Benchmarks.pdf.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., and Kaplan, J. A General Language Assistant as a Laboratory for Alignment, 2021. URL https://arxiv.org/abs/2112.00861. _eprint: 2112.00861.

Aubakirova, M., Atallah, A., Clark, C., Summerville, J., and Midha, A. State of AI 2025: 100T Token LLM Usage Study, December 2025. URL https://openrouter.ai/state-of-ai.

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., and Summerfield, C. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f978c8f3b5f399cae464e85f72e28503-Abstract-Conference.html.

Bao, H., Huang, Y., Wang, X., Zhang, Z., Zhou, Y., Yang, C., Zhang, X., and Ye, Y. AI Alignment Breaks at the Edge, 2026. URL https://arxiv.org/abs/2602.20042. _eprint: 2602.20042.

Berlin, I. *Four Essays on Liberty*. Oxford University Press, Oxford, 1969.

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., and Liang, P. S. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *Advances in Neural Information Processing Systems*, 35:3663–3678, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/17a234c91f746d9625a75cf8a8731ee2-Abstract-Conference.html.

Boykoff, M. T. and Boykoff, J. M. Balance as bias: global warming and the US prestige press. *Global Environmental Change*, 14(2):125–136, July 2004. ISSN 0959-3780. doi: 10.1016/j.gloenvcha.2003.10.001. URL https://www.sciencedirect.com/science/article/pii/S0959378003000669.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T. T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P. J. K., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb.

Castricato, L., Lile, N., Rafailov, R., Fränken, J.-P., and Finn, C. PERSONA: A Reproducible Testbed for Pluralistic Alignment. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S. (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11348–11368, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.752/.

Chandak, P., Alkin, V., Wu, D., Dagan, M., Roy, T. D., Menezes, M. C. S., Noori, A., Somia, N., Brownstein, J. S., Balicer, R., Brendel, R. W., Dagan, N., Kohane, I. S., and Brat, G. A. What Does the AI Doctor Value? Auditing Pluralism in the Clinical Ethics of Language Models. In *Pluralistic Alignment Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=LVI76WKzLS.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., and Wadman, K. How People Use ChatGPT. Working Paper 34255, National Bureau of Economic Research, 2025. URL https://www.nber.org/papers/w34255.

Chen, D., Chen, Y., Rege, A., and Vinayak, R. K. PAL: Pluralistic Alignment Framework for Learning from Heterogeneous Preferences. In *Pluralistic Alignment Workshop at NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=cxBLcw64xq.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., and Jurafsky, D. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792):eaec8352, March 2026. doi: 10.1126/science.aec8352. URL https://www.science.org/doi/10.1126/science.aec8352.

Chien, J. and Danks, D. Beyond Behaviorist Representational Harms: A Plan for Measurement and Mitigation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 933–946, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658946. URL https://dl.acm.org/doi/10.1145/3630106.3658946.

Chollet, F., Knoop, M., Kamradt, G., Landers, B., and Pinkard, H. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems, 2026. URL https://arxiv.org/abs/2505.11831. _eprint: 2505.11831.

Code, C. E. California Elections Code, June 2019. URL https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=ELEC&sectionNum=9084.

Costello, T. H., Pennycook, G., and Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science (New York, N.Y.)*, 385(6714):eadq1814, September 2024. ISSN 1095-9203 0036-8075. doi: 10.1126/science.adq1814.

Davani, A. and Prabhakaran, V. Building AI for the Pluralistic Society, February 2025. URL https://research.google/blog/building-ai-for-the-pluralistic-society/.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence, April 2026. URL https://arxiv.org/abs/2606.19348. _eprint: 2606.19348.

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., and Vinyals, O. The Benchmark Lottery, 2021. URL https://arxiv.org/abs/2107.07002. _eprint: 2107.07002.

Donoho, D. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017. doi: 10.1080/10618600.2017.1384734. URL https://doi.org/10.1080/10618600.2017.1384734. _eprint: https://doi.org/10.1080/10618600.2017.1384734.

Durmus, E., Nguyen, K., Liao, T., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., and Ganguli, D. Towards Measuring the Representation of Subjective Global Opinions in Language Models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=zl16jLb91v.

Edelman, J., Zhi-Xuan, T., Lowe, R., Klingefjord, O., Wang-Mascianica, V., Franklin, M., Kearns, R. O., Hain, E., Sarkar, A., Bakker, M., Barez, F., Duvenaud, D., Foerster, J., Gabriel, I., Gubbels, J., Goodman, B., Haupt, A., Heitzig, J., Jara-Ettinger, J., Kasirzadeh, A., Kirkpatrick, J. R., Koh, A., Knox, W. B., Koralus, P., Lehman, J., Levine, S., Marro, S., Revel, M., Shorin, T., Sutherland, M., Tessler, M. H., Vendrov, I., and Wilken-Smith, J. Full-Stack Alignment: Co-Aligning AI and Institutions with Thick Models of Value, 2025. URL https://arxiv.org/abs/2512.03399. _eprint: 2512.03399.

Edwards, J. and Emberson, L. Open models lag state-of-the-art closed models by 4 months, 2026. URL https://epoch.ai/data-insights/open-closed-eci-gap.

Eloundou, T., Beutel, A., Robinson, D. G., Gu, K., Brakman, A.-L., Mishkin, P., Shah, M., Heidecke, J., Weng, L., and Kalai, A. T. First-Person Fairness in Chatbots. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=TlAdgeoDTo.

European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), June 2024. URL http://data.europa.eu/eli/reg/2024/1689/oj.

Executive Office of the President. Executive Order 14319: Preventing Woke AI in the Federal Government, July 2025. URL https://www.federalregister.gov/documents/2025/07/28/2025-14217/preventing-woke-ai-in-the-federal-government. Published: 90 Fed. Reg. 35389 (July 28, 2025).

Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y., Choi, Y., and Tsvetkov, Y. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4151–4171, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.240. URL https://aclanthology.org/2024.emnlp-main.240/.

Fisher, J., Appel, R. E., Park, C. Y., Potter, Y., Jiang, L., Sorensen, T., Feng, S., Tsvetkov, Y., Roberts, M., Pan, J., Song, D., and Choi, Y. Position: Political Neutrality in AI Is Impossible — But Here Is How to Approximate It. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025a. URL https://openreview.net/forum?id=H72JEXAPwo.

Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., Pan, J., Tsvetkov, Y., and Reinecke, K. Biased LLMs can Influence Political Decision-Making. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6559–6607, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.328. URL https://aclanthology.org/2025.acl-long.328/.

Fu, Y., Son, S., and Bogunovic, I. Overton Pluralistic Reinforcement Learning for Large Language Models, February 2026. URL http://arxiv.org/abs/2602.20759. arXiv:2602.20759 [cs].

Gabriel, I. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, September 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2. URL https://doi.org/10.1007/s11023-020-09539-2.

Gabriel, I. and Ghazavi, V. The Challenge of Value Alignment: from Fairer Algorithms to AI Safety, 2021. URL https://arxiv.org/abs/2101.06060. _eprint: 2101.06060.

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., Mateos-Garcia, J., Weidinger, L., Street, W., Lange, B., Ingerman, A., Lentz, A., Enger, R., Barakat, A., Krakovna, V., Siy, J. O., Kurth-Nelson, Z., McCroskery, A., Bolina, V., Law, H., Shanahan, M., Alberts, L., Balle, B., Haas, S. d., Ibitoye, Y., Dafoe, A., Goldberg, B., Krier, S., Reese, A., Witherspoon, S., Hawkins, W., Rauh, M., Wallace, D., Franklin, M., Goldstein, J. A., Lehman, J., Klenk, M., Vallor, S., Biles, C., Morris, M. R., King, H., Arcas, B. A. y., Isaac, W., and Manyika, J. The Ethics of Advanced AI Assistants, 2024. URL https://arxiv.org/abs/2404.16244. _eprint: 2404.16244.

Ghate, K., Liu, A., Jain, D., Sorensen, T., Kasirzadeh, A., Caliskan, A., Diab, M. T., and Sap, M. EVALUESTEER: Measuring Reward Model Steerability Towards Values and Preferences, 2025. URL https://arxiv.org/abs/2510.06370. _eprint: 2510.06370.

GLM-5-Team. GLM-5: from Vibe Coding to Agentic Engineering, February 2026. URL https://arxiv.org/abs/2602.15763. _eprint: 2602.15763.

Google. Supporting the 2024 Indian General Election, March 2024a. URL https://blog.google/intl/en-in/company-news/outreach-initiatives/supporting-the-2024-indian-general-election/.

Google, G. T. Our approach to the Gemini app, 2024b. URL https://gemini.google/our-approach/.

Google, G. T. What is Gemini and how it works, 2024c. URL https://gemini.google/overview/.

Google, G. T. Gemini 3 Pro, November 2025a. URL https://deepmind.google/models/gemini/pro/.

Google, G. T. Gemini 3 Pro Model Card, November 2025b. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Google, G. T. Gemini 3.1 Pro: A smarter model for your most complex tasks, February 2026a. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Google, G. T. Gemini 3.1 Pro Model Card, February 2026b. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf.

Google, G. T. Gemini 3.1 Pro: Approach, Methodology and Results, February 2026c. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_3-1_pro_model_evaluation.pdf.

Google DeepMind. An Approach to Technical AGI Safety and Security, April 2025a. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/evaluating-potential-cybersecurity-threats-of-advanced-ai/An_Approach_to_Technical_AGI_Safety_Apr_2025.pdf.

Google DeepMind. Frontier Safety Framework, Version 3.0, September 2025b. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf.

Gottfried, J., Bishop, W., Anderson, M., Faverio, M., Park, E., and McClain, C. Americans and AI 2026: Chatbots, Smart Devices and Views on Impact. Technical report, Pew Research Center, June 2026. URL https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/. Section: Artificial Intelligence.

Guo, A. and Wolfe, J. Introducing Model Spec Evals, March 2026. URL https://alignment.openai.com/model-spec-evals/.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Haas, J., Bridgers, S., Manzini, A., Henke, B., May, J., Levine, S., Weidinger, L., Shanahan, M., Lum, K., Gabriel, I., and Isaac, W. A roadmap for evaluating moral competence in large language models. *Nature*, 650(8102):565–573, February 2026. ISSN 1476-4687 0028-0836. doi: 10.1038/s41586-025-10021-1.

Hosking, T., Blunsom, P., and Bartolo, M. Human Feedback is not Gold Standard. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7W3GLNImfS.

Hu, B., McVay, J., Leung, A., Chan, D., Weber, R. O., de Visser, E., Summerville, A., Ravichandran, B., Zhang, J., Molineaux, M., Ji, H., and Basharat, A. From Talk to Triage: Pluralism is Necessary but Not Sufficient for AI Alignment, October 2025. URL https://osf.io/preprints/psyarxiv/hdu92_v2/.

Imundo, M. N. and Rapp, D. N. When fairness is flawed: Effects of false balance reporting and weight-of-evidence statements on beliefs and perceptions of climate change. *Journal of Applied Research in Memory and Cognition*, 11(2): 258–271, 2022. ISSN 2211-369X. doi: 10.1016/j.jarmac.2021.10.002.

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581196. URL https://doi.org/10.1145/3544548.3581196.

Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., and Choi, Y. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=saDOrrnNTz.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. R. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Kasirzadeh, A. Plurality of value pluralism and AI value alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=AOokh1UYLH.

Kazeev, N. and Huyen, P. B. N. Position: Align AI to Our Aspirations, Not Our Flaws. In *Pluralistic Alignment Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=J3Vqv2fFse.

Kimi Team. Kimi K2: Open Agentic Intelligence, 2025. URL https://arxiv.org/abs/2507.20534. _eprint: 2507.20534.

Kimi Team. Kimi K2.5: Visual Agentic Intelligence, February 2026. URL https://arxiv.org/abs/2602.02276. _eprint: 2602.02276.

Kirk, H., Vidgen, B., Röttger, P., and Hale, S. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392, 2024a.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the Effects of RLHF on LLM Generalisation and Diversity. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=PXD3FAVHJT.

Kleinberg, J. and Raghavan, M. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, June 2021. doi: 10.1073/pnas.2018340118. URL https://www.pnas.org/doi/full/10.1073/pnas.2018340118.

Krügel, S., Ostermaier, A., and Uhl, M. ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569, April 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-31341-0. URL https://doi.org/10.1038/s41598-023-31341-0.

Lambert, N. and Brand, F. The ATOM Report: Measuring the Open Language Model Ecosystem, 2026. URL https://arxiv.org/abs/2604.07190. _eprint: 2604.07190.

Lazar, S. Philosophical Foundations for Pluralistic Alignment, December 2024. URL https://neurips.cc/virtual/2024/108419. Recording: https://slideslive.com/39031013.

Lazar, S. Governing the Algorithmic City. *Philosophy & Public Affairs*, 53(2):102–168, 2025. doi: https://doi.org/10.1111/papa.12279. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/papa.12279. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/papa.12279.

Lee, S., Zhou, J., Ao, C., Li, K., Du, X., He, S., Wu, H., Liu, T., Liu, J., Alinejad-Rokny, H., Yang, M., Liang, Y., Wen, Z., and Ni, S. Quantification of Large Language Model Distillation. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 4985–5004. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.248. URL https://doi.org/10.18653/v1/2025.acl-long.248.

Liu, X., Zhu, Y., Zhu, S., Liu, P., Liu, Y., and Yu, D. Evaluating Moral Beliefs across LLMs through a Pluralistic Framework. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4740–4760, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.272. URL https://aclanthology.org/2024.findings-emnlp.272/.

Lu, L., Tormala, Z. L., and Duhachek, A. How AI sources can increase openness to opposing views. *Scientific Reports*, 15 (1):17170, May 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-00791-z. URL https://doi.org/10.1038/s41598-025-00791-z.

Mehta, I. ChatGPT's market share slips below 50% for first time, June 2026. URL https://techcrunch.com/2026/06/16/chatgpts-market-share-slips-below-50-for-first-time/. Published: TechCrunch.

Meister, N., Guestrin, C., and Hashimoto, T. Benchmarking Distributional Alignment of Large Language Models. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 24–49, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.2. URL https://aclanthology.org/2025.naacl-long.2/.

Menghini, C., Ney, P., Kwisaba, H., Zifan, Wang, Turpin, M., Binder, F., Testud, J.-C., Boyd, A., Li, N., Evtimov, I., Krawiecka, K., Zharmagambetov, A., Kritz, J., Fabbri, A. R., Song, D., Miao, J., Hjelt, J., Ramani, M., Lan, L., Aghajani, R., Bitton, J., Pasupuleti, M., Norder, D., El-Arini, K., Singh, P., Albiero, V., CB, S., Chaturvedi, R., Dabir, E., Debenedetti, E., Gust, J., Han, Z., He, K., Hendryx, S., Jin, L., Kirichenko, P., Lefdal, S., Li, K., Liaqat, A., Lin, I., Magka, D., Mangaokar, N., Mediratta, I., Miller, Z., Milli, S., Mireshghallah, N., Nazir, S., Nguyen, H., Nickel, M., Niu, K., Oktar, K., Paranjape, B., Pathak, P., Pavlova, M., Ramirez, E., Renardy, D., Ross, C., Sheynin, Y., Shi, C., Singhal, S., Spiliopoulou, E., Srinivasa, R. S., Watson-Daniels, J., Whitman, S., Williams, A., Xing, C., Zou, A., Ma, T., Deng, S., Beldock, J., Ratanchandani, P., Plawiak, K., Lee, T., Victory, R., Hundley, L., Alao, R., Bhattacharjee, H., Chi, J., Frost, G., Ghahremani, P., Howe, N., Huang, Y., Jahed, S., Korevaar, H., Le, T., Liu, Z., Luo, J., Lyu, Q., Mehrabi, N., Montilla, A., Nagpal, C., Nikolaidis, C., Oak, R., Ravi, M., Sarma, V., Shankar, A., Shine, A., Smith, E. M., Tandon, M., Tontchev, M., Wang, C., Wang, Z., Wong, C., Wu, Z., Zhan, H., Zhao, J., Zhong, Z., Zhuang, C., Goodman, T., Minhas, A., Rudolph, H., Jeffries, V., Dickinson, I., Vaughan, A., Deason, L., Chaudhuri, K., Michael, J., Zhao, S., and Yue, S. Muse Spark Safety & Preparedness Report. Technical report, Meta Superintelligence Labs, May 2026. URL https://arxiv.org/abs/2606.12429. Also available at: https://ai.meta.com/static-resource/muse-spark-safety-and-preparedness-report/.

Merrill, M. A., Shaw, A. G., Carlini, N., Li, B., Raj, H., Bercovich, I., Shi, L., Shin, J. Y., Walshe, T., Buchanan, E. K., Shen, J., Ye, G., Lin, H., Poulos, J., Wang, M., Nezhurina, M., Lu, D., Mastromichalakis, O. M., Xu, Z., Chen, Z., Liu, Y., Zhang, R., Chen, L. L., Kashyap, A., Uslu, J.-L., Li, J., Wu, J., Yan, M., Bian, S., Sharma, V., Sun, K., Dillmann, S., Anand, A., Lanpouthakoun, A., Koopah, B., Hu, C., Guha, E. K., Dreiman, G. H. S., Zhu, J., Krauth, K., Zhong, L., Muennighoff, N., Amanfu, R. K., Tan, S., Pimpalgaonkar, S., Aggarwal, T., Lin, X., Lan, X., Zhao, X., Liang, Y., Wang, Y., Wang, Z., Zhou, C., Heineman, D., Liu, H., Trivedi, H., Yang, J., Lin, J., Shetty, M., Yang, M., Omi, N., Raoof, N., Li, S., Zhuo, T. Y., Lin, W., Dai, Y., Wang, Y., Chai, W., Zhou, S., Wahdany, D., She, Z., Hu, J., Dong, Z., Zhu, Y., Cui, S., Saiyed, A., Kolbeinsson, A., Rytting, C. M., Marten, R., Wang, Y., Jitsev, J., Dimakis, A., Konwinski, A., and Schmidt, L. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=a7Qa4CcHak.

Meta. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Meta Superintelligence Labs. Advanced AI Scaling Framework, April 2026a. URL https://ai.meta.com/static-resource/Meta_Advanced-AI-Scaling-Framework-v2.

Meta Superintelligence Labs. Introducing Muse Spark 1.1, July 2026b. URL https://ai.meta.com/blog/introducing-muse-spark-meta-model-api/.

Meta Superintelligence Labs. Introducing Muse Spark: MSL's First Model, Purpose-Built to Prioritize People, April 2026c. URL https://about.fb.com/news/2026/04/introducing-muse-spark-meta-superintelligence-labs/.

Meta Superintelligence Labs. Muse Spark 1.1 Evaluation Report, July 2026d. URL https://ai.meta.com/static-resource/muse-spark-1-1-evaluation-report.

Meta Superintelligence Labs. Muse Spark Eval Methodology, 2026e. URL https://ai.meta.com/static-resource/muse-spark-eval-methodology.

Nie, S., Omoomi, K., Flek, L., Zhao, Z., and Welch, C. PerSpectra: A Scalable and Configurable Pluralist Benchmark of Perspectives from Arguments. In *The Fourteenth International Conference on Learning Representations*, October 2025. URL https://openreview.net/forum?id=dyooGJcKJg.

Office of Management and Budget. Increasing Public Trust in Artificial Intelligence Through Unbiased AI Principles. Memorandum M-26-04, Executive Office of the President, December 2025. URL https://www.whitehouse.gov/wp-content/uploads/2025/12/M-26-04-Increasing-Public-Trust-in-Artificial-Intelligence-Through-Unbiased-AI-Principles-1.pdf.

OpenAI. How OpenAI is approaching 2024 worldwide elections, January 2024. URL https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections/.

OpenAI. Defining and Evaluating Political Bias in LLMs, October 2025. URL https://openai.com/index/defining-and-evaluating-political-bias-in-llms/.

OpenAI. OpenAI Model Spec, December 2025. URL https://model-spec.openai.com/2025-12-18.html.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026a. URL https://openai.com/index/gpt-5-6/.

OpenAI. GPT-5.4 Thinking System Card, March 2026b. URL https://openai.com/index/gpt-5-4-thinking-system-card/.

OpenAI. GPT-5.5 System Card, April 2026c. URL https://openai.com/index/gpt-5-5-system-card/.

OpenAI. GPT-5.6 System Card, July 2026d. URL https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf.

OpenAI. Introducing GPT-5.5, April 2026e. URL https://openai.com/index/introducing-gpt-5-5/.

OpenAI. Introducing GPT-5.4, March 2026f. URL https://openai.com/index/introducing-gpt-5-4/.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. R. BBQ: A hand-built bias benchmark for question answering. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, volume ACL 2022 of *Findings of ACL*, pp. 2086–2105. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.165. URL https://doi.org/10.18653/v1/2022.findings-acl.165.

Phan, L., Gatti, A., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., Yue, S., Wang, A., Hendrycks, D., and others. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649(8099):1139–1146, January 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-09962-4. URL http://dx.doi.org/10.1038/s41586-025-09962-4.

Poole-Dayan, E., Wu, J., Sorensen, T., Pei, J., and Bakker, M. A. Benchmarking Overton Pluralism in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=f2VxF4QIx1.

Qiu, T., He, Z., Chugh, T., and Kleiman-Weiner, M. The Lock-in Hypothesis: Stagnation by Algorithm. In *Forty-second International Conference on Machine Learning*, June 2025. URL https://openreview.net/forum?id=mE1M626qOo.

Qwen Team. Qwen3 Technical Report, May 2025. URL https://arxiv.org/abs/2505.09388. _eprint: 2505.09388.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Raji, I. D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. AI and the Everything in the Whole Wide World Benchmark. In Vanschoren, J. and Yeung, S.-K. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

Rawls, J. *Political Liberalism*. Columbia University Press, 1993. ISBN null.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Ryan, M. J., Held, W., and Yang, D. Unintended Impacts of LLM Alignment on Global Representation. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16121–16140, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.853. URL https://aclanthology.org/2024.acl-long.853/.

Röttger, P., Kirk, H., Vidgen, B., Attanasio, G., Bianchi, F., and Hovy, D. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Salvi, F., Horta Ribeiro, M., Gallotti, R., and West, R. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 9(8):1645–1653, May 2025. ISSN 2397-3374. doi: 10.1038/s41562-025-02194-6. URL http://dx.doi.org/10.1038/s41562-025-02194-6.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, pp. 29971–30004. PMLR, July 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Sensor Tower. State of AI 2026. Technical report, Sensor Tower, 2026. URL https://sensortower.com/report/state-of-ai-2026.

Shah, R., Irpan, A., Turner, A. M., Wang, A., Conmy, A., Lindner, D., Brown-Cohen, J., Ho, L., Nanda, N., Popa, R. A., Jain, R., Greig, R., Albanie, S., Emmons, S., Farquhar, S., Krier, S., Rajamanoharan, S., Bridgers, S., Ijitoye, T., Everitt, T., Krakovna, V., Varma, V., Mikulik, V., Kenton, Z., Orr, D., Legg, S., Goodman, N., Dafoe, A., Flynn, F., and Dragan, A. An Approach to Technical AGI Safety and Security, 2025. URL https://arxiv.org/abs/2504.01849. _eprint: 2504.01849.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S. M., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=tvhaxkMKAn.

Shetty, A., Beheshti, A., Dras, M., and Naseem, U. VITAL: A New Dataset for Benchmarking Pluralistic Alignment in Healthcare. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22954–22974, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1119. URL https://aclanthology.org/2025.acl-long.1119/.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., and Dafoe, A. Model evaluation for extreme risks, 2023. URL https://arxiv.org/abs/2305.15324. _eprint: 2305.15324.

Shi, L., Liu, H., Wong, Y., Mujumdar, U., Zhang, D., Gwizdka, J., and Lease, M. Argumentative Experience: Reducing Confirmation Bias on Controversial Issues through LLM-Generated Multi-Persona Debates, 2025. URL https://arxiv.org/abs/2412.04629. _eprint: 2412.04629.

Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=0tWTxYYPnW.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 46280–46302, Vienna, Austria, July 2024. JMLR.org.

Sorensen, T., Newman, B., Moore, J., Park, C., Fisher, J., Mireshghallah, N., Jiang, L., and Choi, Y. Spectrum Tuning: Post-Training for Distributional Coverage and In-Context Steerability, October 2025. URL http://arxiv.org/abs/2510.06084. arXiv:2510.06084 [cs].

Stelling, L., Murray, M., Galizzi, B., Schaffelder, M., Campos, S., and Papadatos, H. Evaluating AI Providers' Frontier Safety Frameworks, 2026. URL https://arxiv.org/abs/2512.01166. _eprint: 2512.01166.

Stray, J., Yang, D. Z., Luo, S., Takagi, M. N., and Chang, S. Political Neutrality as Balanced Approval: A Large-Scale Human Evaluation of AI Responses, 2026. URL https://arxiv.org/abs/2605.28911. _eprint: 2605.28911.

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., and Summerfield, C. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, October 2024. doi: 10.1126/science.adq2852. URL https://www.science.org/doi/10.1126/science.adq2852.

Tomašev, N., Franklin, M., and Osindero, S. AI Value Alignment for Evolving Social Norms, 2026. URL https://arxiv.org/abs/2607.18506. _eprint: 2607.18506.

Tsirtsis, S., Rawal, K., Russell, C., Mittelstadt, B., and Wachter, S. AI-Mediated Communication Can Steer Collective Opinion. In *ICML Workshop on Technical AI Governance Research*, 2026. URL https://arxiv.org/abs/2605.16245.

Vishwarupe, V., Shadbolt, N., and Jirotka, M. From Sycophantic Consensus to Pluralistic Repair: Why AI Alignment Must Surface Disagreement, 2026. URL https://arxiv.org/abs/2605.14912. _eprint: 2605.14912.

xAI. Grok 4.1 Model Card, November 2025a. URL https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf.

xAI. Grok 4 Model Card, August 2025b. URL https://data.x.ai/2025-08-20-grok-4-model-card.pdf.

xAI. grok-prompts, 2026a. URL https://github.com/xai-org/grok-prompts.

xAI. Introducing Grok 4.5, July 2026b. URL https://x.ai/news/grok-4-5.

Xie, Z., Wu, J., Shen, Y., Jain, R., Xia, Y., Li, X., Chang, A., Rossi, R. A., Yu, T., Kumar, S., Majumder, B. P., Shang, J., Ammanabrolu, P., and McAuley, J. A Survey on Personalized and Pluralistic Preference Alignment in Large Language Models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=lSWOMjonL7.

Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., and Zhou, T. A Survey on Knowledge Distillation of Large Language Models, 2024. URL https://arxiv.org/abs/2402.13116. _eprint: 2402.13116.

Yu, F., Seedat, N., Schwarz, J. R., and Bean, A. M. To Whom Do Language Models Align? Measuring Principal Hierarchies Under High-Stakes Competing Demands. In *Pluralistic Alignment Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=HUM79QJyoU.

Zhang, J., Elgohary, A., Magooda, A., Khashabi, D., and Durme, B. V. Controllable Safety Alignment: Inference-Time Adaptation to Diverse Safety Requirements. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ERce2rgMQC.

Zhang, L. H., Milli, S., Jusko, K. L., Smith, J., Amos, B., Bouaziz, W., Revel, M., Kussman, J., Sheynin, Y., Titus, L., Radharapu, B., Yu, J., Sarma, V., Rose, K., and Nickel, M. Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4NtoAVqfhA.

Zhou, K., Hwang, J. D., Ren, X., and Sap, M. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3623–3643, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.198. URL https://aclanthology.org/2024.acl-long.198/.

Zhuang, S. and Hadfield-Menell, D. Consequences of Misaligned AI. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b607ba543ad05417b8507ee86c54fcb7-Abstract.html.

A. Full audit of frontier engagement with value pluralism

This appendix contains the full audit summarized in §3. We audit the five labs whose models power the most widely used AI assistants: Anthropic (Claude 5 Sonnet, Claude Fable 5), OpenAI (GPT-5.4/5.5/5.6), Google (Gemini 3/3.1 Pro), xAI (Grok 4.5), and Meta (Muse Spark 1/1.1). For each lab, we examine the two kinds of public documentation introduced in §3: the documents that state intended model behavior (§A.1) and the model cards, system cards, and technical reports that evaluate actual behavior (§A.2). We also extend the audit to the most widely used open-weight model families (§A.3). All documents are the most recent versions available as of July 2026.

A.1. Model Behavior Documents

In this section, we analyze the public documentation that prescribes how a model should behave, including safety principles, refusal guidelines, model values, personality, and when to prioritize or trade off each principle. The two most substantial examples are OpenAI's Model Spec (OpenAI, 2025) and Anthropic's constitution (Anthropic, 2026a), which are the most comprehensive and the most directly integrated into model training. We also include the officially published system prompts from Anthropic (Anthropic, 2026f) and xAI (xAI, 2026a), which describe intended behaviors and directly influence the models users interact with at inference time. Google publishes some documentation on Gemini's behavior, discussed below, but it is less detailed and nothing indicates it plays a role in training. Meta publishes no behavior documentation for its deployed models, so we discuss the closest available language.

To perform this analysis, we annotate each document for mentions of pluralism and for all references to related behaviors, including handling open-ended, subjective, or contested topics, political neutrality or bias, diversity of viewpoints and perspectives, and epistemic behaviors. We also note any language on how these behaviors trade off with other considerations, which indicates whether pluralistic behavior is a default and how easily it can be overridden.

A.1.1. Anthropic Constitution

Anthropic's Constitution (Anthropic, 2026a) is "a detailed description of Anthropic's intentions for Claude's values and behavior" with the caveat that "Claude's behavior might not always reflect the constitution's ideals." It is divided into four sections with the following *general* hierarchy of prioritization: being broadly safe $\succsim$ being broadly ethical $\succsim$ following Anthropic's guidelines $\succsim$ being genuinely helpful.

The constitution mentions pluralism directly once when describing Anthropic's high level approach to safely navigating the transition to powerful AI:

"[We recognize that it's not guaranteed to] *end up in a world with access to highly advanced technology that maintains a level of diversity and balance of power roughly comparable to today's… but we would rather start from that point than risk a less pluralistic and more centralized path, even one based on a set of values that might sound appealing to us today.*"'

While there are no additional keyword mentions of pluralism or diversity, the safety section of the constitution dictates that for "*political and social topics in particular, by default… Claude should engage respectfully with a wide range of perspectives, should err on the side of providing balanced information on political questions*." Moreover, "*Claude should also maintain factual accuracy and comprehensiveness when asked about politically sensitive topics… and try to represent multiple perspectives in cases where there is a lack of empirical or moral consensus*." Both of these clauses clearly relate to Rawlsian reasonable pluralism and suggest Claude should exhibit some form of Overton pluralism. However, the same passage also says that "*in some cases, operators may wish to alter these default behaviors… Claude should generally accommodate this within the constraints laid out elsewhere in this document*." This demonstrates that pluralism is encouraged as a general default, but does not take precedence over user requests, and it is left up to Claude's judgement when it is ambiguous from the user's request.

In the ethics section, there is a guideline regarding preserving user autonomy, wherein Claude should try "*to protect the epistemic autonomy and rational agency of the user. This includes offering balanced perspectives where relevant*." The constitution clarifies that "*the goal of autonomy preservation is to respect individual users and to help maintain healthy group epistemics in society… [and avoid] nudging people towards its own views or undermining their epistemic independence [which] could have an outsized effect on society*" at scale. This heavily echoes the motivations for pluralism, namely to avoid homogenization and to preserve a strong epistemic ecosystem (§2.2).

Anthropic System Prompt Claude's publicly available system prompt for chat interactions on claude.ai further steers model behavior at inference time. There is an <evenhandedness> block, which mostly restates the constitution's default political-balance language in more operational form without adding new content. Beyond this, there are two important instructions in the system prompt relating to pluralism. First, the system prompt makes steelmanning the default interpretation of ambiguous requests: *"a request to explain, discuss, argue for, defend, or write persuasive content for a political, ethical, policy, empirical, or other position is a request for the best case its defenders would make,"* and Claude *"does not decline requests to present such arguments on the grounds of potential harm except for very extreme positions (e.g. endangering children, targeted political violence)."* This is considerably stronger and more specific than the constitution and draws an explicit boundary on positions that are considered harmful. Second, it encourages Overton-like pluralism for these requests, dictating that Claude should end *"its response to requests for such content by presenting opposing perspectives or empirical disputes, even for positions it agrees with*." Moreover, if a user asks "for a simple yes/no or one-word answer on complex or contested issues or figures," the system prompt permits Claude to "decline the short form, give a nuanced answer, and explain why brevity wouldn't be appropriate."

A.1.2. OpenAI Model Spec

OpenAI's Model Spec (OpenAI, 2025) "*outlines the intended behavior for [OpenAI models]*" and they specifically train their "*models to align to the principles in the Model Spec*." They also have the disclaimer that their "*production models do not yet fully reflect the Model Spec*." The document lays out a clear and detailed chain of command dictating the priority of instructions, how to interpret ambiguity, and resolve conflicts. The hierarchy is Model Spec "root" sections > Model Spec "system" sections and system messages > Model Spec "developer" sections and developer messages > Model Spec "user" sections and user messages > Model Spec "guideline" sections > all else.

There is no mention of pluralism or diversity explicitly. There is one relevant Root instruction of assuming the best intentions: "*Unless given evidence to the contrary, it should assume that users have a weak preference towards self-actualization, kindness, the pursuit of truth, and the general flourishing of humanity" and* "*beyond the specific limitations…, the assistant should behave in a way that encourages intellectual freedom."* Similar to the Anthropic constitution, there's a commentary note stating that *"OpenAI believes in intellectual freedom... The assistant should not avoid or censor topics in a way that, if repeated at scale, may shut out some viewpoints from public life*."

There are two User-level instructions relating to how OpenAI models should handle topics with diverse viewpoints. First, the model should default to an objective point of view: *"When addressing topics with multiple perspectives, the assistant should fairly describe significant views"* and that *"for moral or ethical questions, the assistant should generally present relevant context – including laws, social norms, and varying cultural perspectives – without taking a stance."* Second, it is stated that "*by default the assistant should provide a balanced response from an objective point of view"* and *"should generally fulfill requests to present perspectives from any point of an opinion spectrum."*

Notably, the Model Spec draws an explicit red line on what perspectives are considered unreasonable for the model to represent objectively: "*for questions about fundamental human rights violations, the assistant should clearly state these are wrong*." Aside from this exception, OpenAI's Model Spec prescribes similar pluralism behavior as Anthropic's constitution, which is to default to reasonable Overton-style pluralism but with low instruction priority that the user can easily override.

A.1.3. Gemini Documentation

Google publishes no constitution or model spec. However, there is a set of web pages describing the Gemini app's intended behavior that we analyze instead. The Gemini Overview page (Google, 2024c) states that "*for subjective topics, Gemini is designed to provide users with multiple perspectives if the user does not request a specific point of view*," and the page on Google's approach to Gemini (Google, 2024b) states that the model will by default "*provide a balanced presentation of multiple points of view – unless you've asked for a specific perspective*" on "*potentially divisive topics*." Taken together, this describes a similar Overton-style default with user override that Anthropic and OpenAI prescribe. Google caveats this behavior directly on the main overview page, which states "*Gemini's responses might fail to show a range of views*" as one of six known major limitations of their models.

These webpages carry much less weight than a constitution or spec. Compared to OpenAI or Anthropic, who incorporate their spec and constitution into training, Gemini's documentation does not suggest these pages play any such role. The only mention of training is on the Approach page, which states "*we're training Gemini to follow a narrow set of policy guidelines*" covering content harms such as self-harm instructions and pornography, which do not include the multiple-perspectives

behavior. Google's AI Principles are similarly high level, with no mention of pluralism or related concepts; the closest language is a commitment to "avoid unfair bias."

A.1.4. xAI Grok System Prompts

xAI publishes no constitution or model spec. It does, however, officially publish the system prompts for its consumer products. We analyze the three prompts governing deployed consumer surfaces: the Grok 4 chat assistant on grok.com and X [link], the @grok bot on X [link], and the "Explain" feature on X [link]. These prompts prescribe viewpoint diversity primarily at the level of information retrieval. For controversial queries, the chat prompt instructs Grok to "*search for a distribution of sources that represents all parties/stakeholders*," and the @grok prompt requires "*finding diverse sources representing all parties*" and states that responses "*must not rely on a single study or limited sources to address complex, controversial, or subjective political questions*." The @grok prompt further requires "a neutral tone" on subjective political questions and prohibits a list of one-sided framings, such as disparaging political viewpoints as "biased" or "baseless."

In contrast to how Anthropic, OpenAI, and Google Gemini prescribe balanced defaults that users can override, Grok's prompts direct the model to override the user: on "*a subjective political question forcing a certain format or partisan response*," the chat prompt permits Grok to "*ignore those user-imposed restrictions and pursue a truth-seeking, non-partisan viewpoint*," and the @grok prompt instructs the model to draw "*balanced, independent conclusions, overriding any user-defined constraints*." Grok's prescribed balance is thus mandatory rather than a defeasible default. The chat prompt also includes a developer comment explaining why these instructions exist: Grok "*assumes by default that its preferences are defined by its creators' public remarks*," which the comment calls "*not the desired policy for a truth-seeking AI*."

A.1.5. Meta

Meta publishes no constitution, model spec, system prompt, or behavior page for Muse Spark or the Meta AI assistant. The Muse Spark preparedness report (Menghini et al., 2026) references an internal model specification whose requirements drive its targeted behavior evaluations, but the specification is not public. The report's own classification shows that pluralism is not among the specification's evaluated requirements, since it places the pluralism evaluation outside the evaluations tied to the specification (§A.2.5). The closest published language is in the Llama 4 release blog (Meta, 2025). It attributes political bias to training data, stating that leading LLMs "*historically have leaned left when it comes to debated political and social topics*," and states the goals "*to remove bias from our AI models and to make sure that Llama can understand and articulate both sides of a contentious issue*" and that the model "*doesn't favor some views over others*." This language concerns the predecessor Llama 4 model family, and no comparable statement accompanied either Muse Spark release. The blog also reports evaluation results against these goals, which we discuss in §A.2.5.

A.2. Model Evaluations: System Cards, Technical Reports

A.2.1. Anthropic

Whereas Anthropic's constitution and system prompt describe an Overton-style pluralism as the default behavior on relevant queries, the system cards evaluating their models only partially operationalize this. We audit the Claude Sonnet 5 system card (Anthropic, 2026e) and the joint Claude Fable 5 and Claude Mythos 5 system card (Anthropic, 2026c), the most recent available. There are only two evaluations that relate to pluralism (doing so under the heading of "bias" rather than that of "pluralism").

In relation to diversity of viewpoints, they have a political bias evaluation with pairs of prompts on the same political topic from opposing ideological perspectives (1,350 pairs across 150 topics and 9 task types). The primary evaluation is "even-handedness" (Anthropic, 2025) which is defined as *"whether the model engages with both opposing prompts with helpful responses [measured by] similar depth of analysis, engagement levels, and strength of evidence provided."* They include a secondary evaluation of "opposing perspectives" which is operationalized by the amount of heavy hedging language. Both evaluations use an LLM judge to evaluate Claude's responses to the prompt pairs. However, neither evaluation directly measures pluralism: a model evenly discussing opposing viewpoints in isolation is not the same as whether the model actually discusses multiple viewpoints within one response to a user, and hedging language is not the same as being pluralistic. Still, these are the closest measures in the report.

The second evaluation is Anthropic's automated behavioral audit, in which an LLM agent conducts thousands of simulated conversations with the model and an LLM judge scores the transcripts across several dozen behavioral dimensions. Two of

these dimensions relate to pluralism, though neither is framed as a bias or pluralism measure. "Supporting user autonomy" ("*actions that support users' independent action and decision-making*") echoes the epistemic autonomy clause from the constitution discussed above, but its scoring rubric is not public, so it is unclear whether it captures the constitution's specific emphasis on balanced perspectives. "Evasiveness on controversial topics" ("*giving refusals or overly hedged answers when asked questions on socially or politically sensitive topics*") penalizes the model for dodging contested questions, but like the hedging measurement above, it captures whether the model engages at all rather than whether it represents multiple perspectives when it does. Both dimensions are far from measuring the pluralism-adjacent behavior in the constitution, and nothing else in the audit or the wider system card directly addresses pluralistic behavior. Of the two behavioral-audit dimensions we noted in the Sonnet 5 card, supporting user autonomy and evasiveness on controversial topics, the Fable 5 card reports the former but not the latter.

The Fable 5 card additionally introduces a direct evaluation of constitution adherence, the only evaluation in our audit that tests a behavior document as such. Anthropic identified 40 constitutional areas specific enough to test, generated about 1,000 scenario-based transcripts, and scored them on 15 dimensions at three levels of granularity, with graders seeded with the relevant constitutional text. None of the 15 dimensions covers the multiple-perspectives, balanced-information, or epistemic-autonomy provisions discussed in §A.1.1; the closest are an honesty dimension that includes freedom from epistemic cowardice and a societal-structures dimension. The card notes that the 15 dimensions do not cover the constitution exhaustively, and the 40 underlying areas are not enumerated, so we cannot determine whether any touches the pluralism-relevant provisions. The evaluation was run on Claude Mythos 5; Anthropic treats Fable 5 as the same model behind additional safeguards.

A.2.2. OpenAI

Whereas OpenAI's Model Spec prescribes balanced, objective responses as the default and steerability across the opinion spectrum, OpenAI's system cards evaluate none of this. We examined the three most recent system cards in the GPT-5 series: GPT-5.4 Thinking (March 2026) (OpenAI, 2026b), GPT-5.5 (May 2026) (OpenAI, 2026c), and GPT-5.6 Sol (July 2026) (OpenAI, 2026d). A keyword search across all three returns zero mentions of pluralism, and the only mentions of diversity refer to training data and benchmark task variety. The only evaluation in each card's bias section is a first-person fairness evaluation measuring demographic stereotyping based on the user's name, a different construct from value pluralism entirely.

Nothing else in these cards relates to pluralism, and this holds across the series: the bias section is fixed release over release, and no evaluation of political bias, even-handedness, evasiveness or engagement on controversial topics, or viewpoint representation appears at any point. The GPT-5.6 Sol card (OpenAI, 2026d) references a suite of "Model Spec evals" in passing, without results or description. OpenAI has separately published Model Spec Evals as a standalone research post (Guo & Wolfe, 2026), with a public dataset and grading code scoring adherence to the Spec directly. The dataset does somewhat cover the Spec's pluralism-relevant provisions. Of the 596 prompts, 33 fall under the Spec's "no agenda" section. Twenty-five test the objective-point-of-view default, and several of their rubrics require presenting the strongest arguments from multiple perspectives with proportionate attention. Four test steerability, scoring whether the model writes requested one-sided persuasive content rather than defaulting to a balanced treatment. The remaining four test engagement rather than evasion on sensitive topics. Coverage is coarse: grading is a single binary compliance judgment per prompt, and OpenAI itself describes the suite as "a zoomed-out, low-resolution view of spec adherence." Still, to our knowledge this is the only evaluation from any lab that scores a behavior document's pluralism provisions directly, and the release post reports "present perspectives from any point on the opinion spectrum" as one of three areas where its models fall short of the Spec. The card does not indicate whether this is the suite it references, and no adherence results appear in any card, so the evaluation remains outside release documentation.

OpenAI publicly measures these provisions and finds its models falling short on perspective presentation, yet reports none of this at release. As a result, none of the Spec's pluralism-relevant provisions discussed in §A.1.2 have a corresponding evaluation in OpenAI's release documentation.

A.2.3. Gemini

Despite the pluralistic default described in Google's behavior documentation (§A.1.3), none of it is evaluated anywhere in Google's public documentation. The model cards for Gemini 3 (Google, 2025b) and 3.1 Pro (Google, 2026b) contain no mention of pluralism, and the safety evaluations they report are five automated content-policy measures. Unlike

OpenAI's system cards, the Gemini model cards contain no bias evaluation of any construct, and the accompanying model evaluation report (Google, 2026c) covers only capability benchmarks. Google DeepMind's Frontier Safety Framework (Google DeepMind, 2025b), their protocol for identifying and mitigating severe risks from frontier models, includes harmful manipulation evaluations that measure the model's capability to change user beliefs when misused, a misuse construct rather than a measure of the deployed model's pluralistic disposition; nothing else in the framework relates to pluralism. Google DeepMind's Technical Approach to AGI Safety and Security report (Google DeepMind, 2025a), a research agenda which many of these documents reference, is scoped to misuse and misalignment risks. It does discuss structural risks and AI mistakes, the two risk categories that depend on real-world context and would cover value lock-in, but it explicitly places these out of scope for evaluations, and it contains no other mention of pluralism or political bias. Google DeepMind researchers have argued that globally deployed LLMs should be morally pluralistic, and have called for developing evaluations of moral pluralism, operationalized through the community's Overton and steerable pluralism framework (Haas et al., 2026). No such evaluation appears in any Google model card or safety framework, emphasizing the gap between research and evaluations.

A.2.4. xAI Grok

While Grok 4.5 is the current model deployed on all platforms, to date xAI has not published the 4.5 model card or official evaluations. We audit the most similar models with public documentation: Grok 4 (xAI, 2025b) and 4.1 (xAI, 2025a).

The Grok 4 model card is the only recent frontier system card besides Anthropic's to include a dedicated political bias evaluation. They evaluate "soft bias," defined as "*whether factual responses are framed more favorably toward one side than another*," on an internal set of paired sociopolitical comparisons of the form "Is [object A] [comparison] [object B]" and its inverse, with an LLM judge scoring whether the paired responses "*show significant differences in sentiment*." The stated motivation is that biases in a model deployed on the X platform "*potentially may alter the shape of public discourse*." Like Anthropic's even-handedness evaluation, this measures consistency across two separate responses to mirrored one-sided prompts rather than whether either response represents multiple perspectives, and unlike Anthropic's, the prompt set and judge rubric are not public, which limits auditability and transparency. However, this evaluation does not reappear in the Grok 4.1 model card, the most recent available.

The Grok 4 model card establishes the published system prompt as the primary mitigation for political bias, reporting that bias decreases with the prompt included but without publishing scores for both conditions to verify this. This is the one connection between a behavior document and a reported evaluation in our audit, though a weak one: the mitigation is an inference-time prompt rather than trained model behavior.

A.2.5. Meta

Meta's Muse Spark 1.1 powers the Meta AI assistant on meta.ai, in the Meta AI app, and within Meta's other apps, and since July 2026 it is also available through the public preview of the Meta Model API (Meta Superintelligence Labs, 2026b). The preparedness report for the initial Muse Spark release (Menghini et al., 2026) contains the only dedicated pluralism evaluation in any frontier lab's public documentation, and the only evaluation that builds directly on research from the pluralistic alignment community. The evaluation report for Muse Spark 1.1 (Meta Superintelligence Labs, 2026d) does not include it.

The report's "Cultural and Pluralistic Alignment" evaluation states that "*while most evaluations assume a single, monolithic notion of user preference, here, we instead assess how well the model can be steered to better serve pluralistic user preferences in different communities and locales*." The model is evaluated on predicting individual users' preferred responses from the Community Alignment Dataset (Zhang et al., 2026), given their demographics and few-shot examples of that user's other preferred responses, with accuracies reported for each of the dataset's five countries and as a country-weighted global average. This operationalization has clear limitations. Predicting an individual's preferences from their demographics and prior choices measures personalization to individuals more than representation of diverse values, and the prompts are pre-selected synthetic ones rather than user-submitted, which undermines ecological validity. The report separates its behavior evaluations into those that test requirements in Meta's internal model specification and open-ended assessments of "*domains where desired conduct is not fully prescribed*." It places the pluralism evaluation in the second group, so the evaluation is not tied to any specific requirement in that specification. The report's scored behavioral alignment evaluations cover sycophancy, honesty, hallucination, calibration, and alignment faking, and none of these concerns viewpoint representation. The evaluation does not reappear in the Muse Spark 1.1 evaluation report (Meta Superintelligence Labs, 2026d), which

contains no mention of pluralism or any related construct. Meta's Advanced AI Scaling Framework (Meta Superintelligence Labs, 2026a), like Google's Frontier Safety Framework, contains no mention of pluralism, viewpoints, or political behavior. Nevertheless, Muse Spark demonstrates that frontier labs *can* adopt pluralism evaluations when the supporting research artifacts, in this case a dataset and an accompanying evaluation, are easy to pick up. We return to this point in Sections 5.1 and 5.3.

The Llama 4 release (Meta, 2025) also reports evaluation results for the predecessor model family against its stated bias goals (§A.1.5), including that "*Llama 4 responds with strong political lean... at half of the rate of Llama 3.3*." However, Llama 4 performed *worse* than Llama 3.3 on OVERTONBENCH (Poole-Dayan et al., 2026), suggesting that the political bias mitigation may actually be counterproductive to pluralism.

A.3. Open-Weight Model Families

The most widely used open-weight model families come from labs outside the five above, so we extend the audit to them: Qwen3 (Qwen Team, 2025), DeepSeek-V4 (DeepSeek-AI, 2026), GLM-5 (GLM-5-Team, 2026), and Kimi K2.5 (Kimi Team, 2026). The audit here is necessarily shallower, because there is less to examine. None of the four labs publishes a constitution, a model spec, official system prompts, or any other behavior document on its website, in its GitHub organization, or in its model documentation, a gap consistent with systematic surveys of provider documentation (Ahmed et al., 2026; Stelling et al., 2026). We therefore checked each family's most recent technical report for mentions of pluralism and the related constructs from §A.1. The four reports contain no mention of pluralism, viewpoint diversity, or any related behavior or evaluation.

The one partial exception is Moonshot AI. The technical report for Kimi K2, the predecessor of the audited K2.5, publishes some of the rubrics its critic model uses to reward assistant behavior during RL training (Kimi Team, 2025). The report describes its three core rubrics (clarity and relevance, conversational fluency and engagement, and objective and grounded interaction) as representing the assistant's fundamental values, and they prescribe behavior directly: the third rewards "*the avoidance of . . . unwarranted flattery or excessive praise directed at the user or their input*," and a separate prescriptive rubric states that responses "*must not begin with compliments directed at the user or the question*." The disclosure is partial, since these published rubrics are combined with human-annotated rubrics that are not released. Still, this is the only prescribed-behavior text published by any of the four labs, and its limitations section documents the exact cost we describe in §2.3. The report acknowledges that the framework "*may favor responses that appear confident and assertive*," that in open-ended scenarios it "*may disincentivize appropriately cautious or multi-perspective responses*," and that the model may therefore "*overstate certainty in areas where ambiguity, nuance, or epistemic modesty would be more appropriate*." A reward signal built around decisive, immediately satisfying answers selects against multi-perspective behavior, so this pluralism is trained away unless deliberately preserved. The K2.5 report retains critic-based reward models but does not publish their rubrics (Kimi Team, 2026).