The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment" in Large Language Models

Hannah Rose Kirk[1] **Bertie Vidgen**[1,2] **Paul Röttger**[3] **Scott A. Hale**[1,2,4]

[1]University of Oxford, [2]The Alan Turing Institute, [3]Bocconi University, [4]Meedan

Abstract

In this paper, we address the concept of "alignment" in large language models (LLMs) through the lens of post-structuralist socio-political theory, specifically examining its parallels to empty signifiers. To establish a shared vocabulary around how abstract concepts of alignment are operationalised in empirical datasets, we propose a framework that demarcates: 1) which dimensions of model behaviour are considered important, then 2) how meanings and definitions are ascribed to these dimensions, and by whom. We situate existing empirical literature and provide guidance on deciding which paradigm to follow. Through this framework, we aim to foster a culture of transparency and critical evaluation, aiding the community in navigating the complexities of aligning LLMs with human populations.

1 Introduction

In post-structuralist socio-political theory, empty signifiers are terms or symbols characterised by their absence of fixed referents [40, 34]. Acting as discursive placeholders, these vague and abstract terms are infused with meaning by different individuals or groups, who can exploit the ambiguity for their particular negotiations of universally-known, but not universally-defined, concepts. In his work *Emancipation*, Laclau distinguishes *empty signifiers* from the related concept of *floating signifiers* [34]. A floating signifier absorbs meaning, allowing the signified concepts to be interpreted fluidly and contextually. In contrast, as Laclau argues, empty signifiers have a stronger power implication—they are terms devoid of meaning precisely because different social and political groups ascribing particularistic meaning is what maintains and drives hegemonic order. In this political context, empty signifiers serve as vehicles for ambiguous or notionally "universal" agreeable pursuits, rallying disparate individuals around abstract ideas without ever offering a concrete conceptual anchor.

The notion of "alignment" in large language models and other AI systems has attracted unprecedented attention in the past year, from researchers, developers, policymakers and citizens alike. Akin to empty signifiers, the term "alignment" serves as a rhetorical placeholder for an aspirational conceptualisation of relations between humans and machines, which is fairly unobjectionable in principle, but lacks a shared definition or goal to translate in practice [18].[1] Statements—like "ensure that powerful AI is properly aligned with human values" [p.2, 55] or "text-based assistant that is aligned with human values" [p.1, 3] or "the behaviour of AI agents needs to be aligned with what humans want" [p.1, 28] or how to generate text that is "in accordance with some shared human values" [p.243, 38]—appeal

[1]Some take it to mean aligning AI with our *standards* [64], *wants* [28] or *motives* [13]; *revealed, stated* or *idealised preferences* [18]; *communication norms* [27]; *intents*, or *expectations* [48, 35] and *goals* [26]; other focus specially on *value* alignment [55, 33] or *social* alignment [37]; finally, there are some who interpret it in the limit as avoiding harm and suffering [50] or even mitigating far-future notions of 'x-risk'.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

to vague and fuzzy ideals, yet do not confront the complexities of what these statements imply nor critically reflect on whose meanings, and which power structures, are encoded in reality.

Empty signifiers, and by analogy these wide-sweeping conceptualisations of alignment, are only sustainable in so far as the abstract can remain abstract. Yet, to *empirically* align language models, that is to tangibly steer them towards certain behaviours and away from others, we require some form of measurable signal or data. When designing data collection protocols, writing annotator guidelines, or hiring annotators, abstract notions of alignment must be calcified into *which preferences, values, or behaviours are important* and *how to measure them*. Just like Laclau's empty signifiers, this question cannot be tangibly answered without the tainting of identity, politics and power. In practice, the question of "which" becomes a question of "whose" [18]—both in terms of who decides the properties of a model that is "aligned", and who actually interprets these concepts for data labelling or curation, and feedback provision. As Ouyang et al. [48] state: "one of the biggest open questions is how to design an alignment process that is transparent" (p.19). In this article, we confront the practical complexities of operationalising abstract notions of alignment into observable training or evaluation signals, in order to encourage greater transparency and shared understanding.

There is a large growing literature that collects empirical data recording human feedback, demonstration and instruction for steering language model behaviour under the motivation of "alignment". This literature has transitioned from early work that adopted abstract and general notions of human preference, like "goodness" or "quality" [62, 72] towards collecting more fine-grained, disaggregated and rich feedback [58, 68, 12], or being explicit about the targeted traits and behaviours [4, 3, 63, 21]. Presenting the literature in detail is outside the remit of this paper but we draw closely on the body of work recently surveyed by Kirk et al. [30] and Wang et al. [66]. When scrutinising the general discourse around human feedback learning—be it red-teaming [19], reinforcement learning with human feedback [48, 43, 21, 4], preference pre-training [32], supervised fine-tuning [71] or direct preference optimisation [52]—there is a lack of shared terminology for articulating what alignment actually means in a given empirical context, what it tangibly achieves, and for whom. Confronting the practical difficulties of converting subjective concepts into labelled or categorised data is not a new problem, and inspiration can be drawn from neighbouring fields. Interpreting annotator differences as signal not noise [25, 2, 44], modelling disagreements [15] or annotator artefacts [23, 57], releasing detailed documentation [51, 7], and doing away with majority vote in favour of more nuanced voting or aggregation systems [22] are established practices in other areas of natural language processing, as well as computer vision [60, 59, 49, 14].

In this paper, we extend a framework for annotating subjective NLP datasets from Rottger et al. [53] to the creation of datasets for LLM alignment. The original framework introduces two data annotation paradigms for facilitating different end goals. The *prescriptive paradigm* discourages annotator subjectivity by providing detailed guidelines, to encode a single set of beliefs in the data. The *descriptive paradigm*, on the other hand, encourages subjectivity, with the goal of capturing a diversity of beliefs. Rottger et al. [53] give the example of *prescriptive* enforcement of hate speech policies on large online platforms, compared to *descriptive* analyses of different perceptions of online hate. By introducing the two paradigms, they hope to enable better documentation and more clarity about the intended uses of different datasets. In the domain of collecting data for LLM alignment, we echo this call for a clearer articulation of objectives. For alignment, it is just as necessary to ask "how to operationalise and communicate a certain concept". However, for alignment, we also need to ask "how to decide which concepts are relevant" in the first place. Our proposed framework centers these two decision points:

1. Identifying the **dimensions** that are included as in scope for the alignment dataset, which can either be *Broad* (general goals e.g., "outputs that people prefer", "good outputs") or *Specific* (named traits or behaviours e.g., "honestly", "informativeness", "harmlessness")
2. Determining the interpretative authority and invariability of the **definitions** of these dimensions, which can be *Prescriptive* (clear detailed definitions of a single belief) or *Descriptive* (subjective interpretation and many beliefs) [54].

Our motivation for introducing this new conceptual framework is to foster a culture of transparency and critical consideration within the empirical alignment research community. To demonstrate its practical utility, we communicate an *ex-post* mapping of existing literature onto the framework's axes; but the framework is intended as a tool for clarifying and communicating *ex-ante* intents of the dataset designers and model builders. Our proposed mapping inevitably suffers from incompleteness

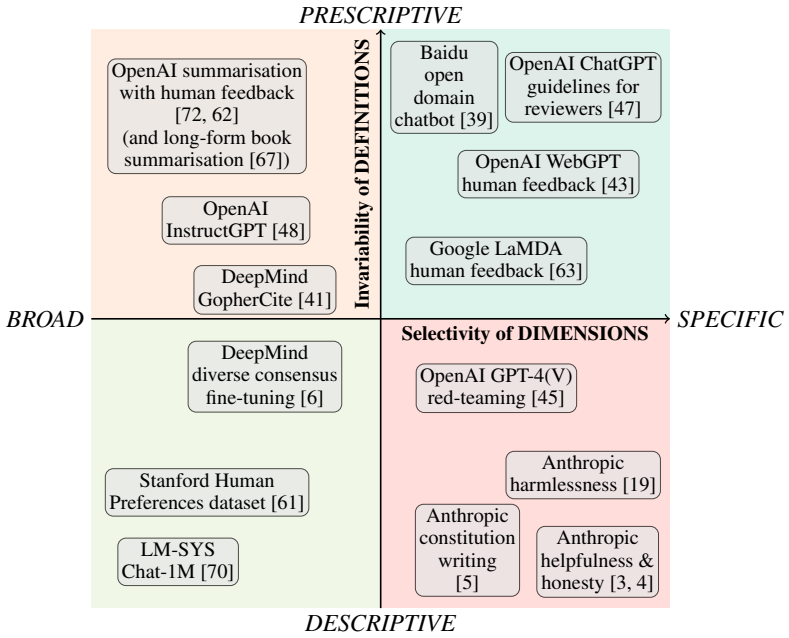

Figure 1: A mapping of existing datasets and empirical literature to our framework. See Tab. 1 for full descriptions and quotations.

in ways that alignment datasets can be categorised; as well as idiosyncrasies of individual researchers or practitioners adopting the role of cartographer. Nonetheless, even if the boundaries and characterisations are imperfect, our aim is to equip people with a language to specify their intents. As more and more people across the world use LLMs on a regular basis, the boundaries of what it means for a model to be aligned, to be helpful, honest or harmless, become increasingly intractable and fuzzy. We hope our framework will assist in transmuting the empty signifier of "alignment" into actionable constructs, empirical signals and sampling strategies, that can be communicated, criticised, and re-conceptualised in a common language.

2 Paradigms for Operationalising Alignment

The process of creating alignment data encompasses two core decisions: 1) selecting dimensions to be included, and 2) deciding how dimensions are defined. We think these decisions are most relevant to what, how and whose behaviours get encoded into LLMs during alignment-tuning;[2] but many other decision points exist, like the weightings of dimensions, aggregation functions, or presentation and order of decisions in interfaces.[3] There are many ways of collecting empirical signals for alignment [30] and our framework is applicable and adaptable to many forms: comparisons between outputs (pick the more <*helpful*> output from this set) [72, 3, 62, 4, 48]; demonstrations (write a <*helpful*> answer to this query) [43]; rewrites or edit chains (rewrite the output to be more <*helpful*>) [36]; or natural language feedback (explain why the output is/is not <*helpful*>) [58]. In this section, we outline our 2x2 framework, discuss how a researcher or developer locates themselves, and demonstrate how existing literature maps onto each of the four quadrants (see Fig. 1 and Tab. 1).

2.1 The 2x2 Framework

Dimensions Any decision to measure a particular attribute of model behaviour (e.g., "Honesty") already bounds the alignment input-output space. The first level of our framework contends with

[2]There are established norms and terminologies for some axes of dataset curation and design in shared documentation standards like Data Statements [8] or Datasheets [20]. We intentionally do not focus on these aspects and instead on what we deem crucial for understanding alignment's interpretative scope and subjectivity.

[3]For example, Nakano et al. [43] provide "criteria in descending order of priority" for raters making final decisions. The criteria include "Whether or not the answer contains unsupported information", and "How much irrelevant information there is in the answer (This can be higher priority in extreme cases.)" (p.18).

the one or more *dimensions* that are considered 'in-scope' for the empirical alignment dataset. On one end of the spectrum is the *Broad* paradigm, where some high-level concept is targeted, without fine-grained detail or sub-behaviours. For example, this could be collecting data on "which output do you prefer", "what output is better", "how good is this output", or "rate the quality of each output" [72, 62]. Examining ELO ratings between model outputs that people prefer is a canonical example of the *Broad* paradigm [70].[4] On the other end of the spectrum is the *Specific* paradigm, where technology providers and dataset creators centre their empirical efforts on named attributes or targeted traits. For example, they decide *a priori* that "honesty, harmlessness, and helpfulness" [3, 4], or "informativeness" [43] and "safety" [63], are key traits to get right for an aligned model. Note that the definitions of these specified dimensions can still be left open to interpretation.

Definitions Adopting Röttger et al. [54]'s terminology, the conceptual treatment of definitions can either be *Prescriptive* or *Descriptive*. Under the *Prescriptive* paradigm, technology providers or dataset creators write detailed guidelines or training tasks, thus pinning down a narrow interpretation and asking annotators to abide by this interpretation as far as possible [43, 21, 39, 63, 47]. Under this paradigm, practitioners measure and try to maximise inter-annotator agreement or calculate majority votes, where conceptually any deviations from agreement represents noise, misunderstanding of the guidelines or poor quality annotation. In contrast, the *Descriptive* paradigm avoids defining the meaning signified by different dimension terms, instead allowing human raters to subjectively inscribe their own contextual meanings [19, 4]. Under this paradigm, interpersonal disagreements are not sought to be minimised; in fact, such differences may be the exact object of interest for the study.

2.2 Questioning the End Goal

Similar to Rottger et al. [53], we suggest any researcher or developer carefully considers the purpose of their empirical alignment dataset in order to decide how it is scoped and collected.

Broad or Specific Dimensions? Three questions guide this choice. First, **how well can I define my aim?** *Broad* dimensions can identify a wide landscape of important behaviours (that may include 'unknown unknowns'), without biasing data with *a priori* expectations. This adopts a positivist epistemology—that the 'reality' or legitimacy of phenomena emerges from empirical measurement. Conversely, a *Specific* stance more reliably encodes particular dimensions considered intrinsically or instrumentally important (for profit, political, reputational or social motives).[5] Second, **how well can I estimate applications of a model tuned on my data?** The narrower the usecase the easier to specify dimensions: a model used for only fiction summarisation or creative writing requires a different and narrower set of traits than a model for information-seeking dialogue [3] or a general purpose conversational agent [3, 4]. Lastly, **how well can I assess what's needed to accomplish the task?** Task complexity influences the ease of specifying dimensions—inverse reinforcement learning was precisely conceived for when oversight is challenging [10]; accordingly, if dimensions cannot be reliably outlined or validated, we recommend a *Broad* approach.

Descriptive or Prescriptive Definitions? The key question here is **do I want to encode a specific belief or capture a diversity of beliefs?** [54]. There is a wave of new literature seeking to measure sociocultural and interpersonal variation in human perceptions of language model outputs [69, 12, 1], as well as assess whether language models themselves reflect diverse opinions [16, 24] or generate consensus [6]. Understanding and/or incorporating diversity in alignment perspectives requires the *Descriptive* paradigm, especially if targeting personalisation or customisation [31, 56, 9]. Encoding one belief requires the *Prescriptive* paradigm to communicate concepts consistently via training tasks or detailed guidelines; yet even with these interventions, disagreement cannot be fully eliminated [21, 48, 62]. Nevertheless, practitioners should clearly communicate whether disagreement between people is the signal they seek to study or the noise they work to eliminate [54].[6]

[4] For example, an interface like chat.lmsys which only includes "A(B) is better", "tie" or "both are bad".

[5] There are trade-offs to consider. The dimensions in a *Specific* stance may be at odds with or deprioritise other critical dimensions [3, 4]. The flexibility of the *Broad* stance loses information on *why* people prefer one output over another or potentially encourages raters to take shortcuts on artefacts such as output length [67].

[6] Statements from Ziegler et al. [72] demonstrate conflicting statements between paradigms. In the same page (p.12), they say "Evaluation of a summary is both subjective and multidimensional" (*Descriptive*), and "One could hope to cope with such 'noise' by simply getting more labels and averaging them but this does not resolve all the practical difficulties with ambiguity" (*Prescriptive*).

Table 1: Empirical literature and existing datasets situated in our framework. We present examples for each of the four quadrants to demonstrate the practical differences between paradigms.

II. DEFINITIONS	I. DIMENSIONS: **Broad** — Leave open wide scope for any or many dimensions of model behaviour.	I. DIMENSIONS: **Specific** — Focus explicitly on one or more *named* dimensions of model behaviour.
Prescriptive — Provide a single meaning of *dimensions* via detailed guidelines and definitions.	• Stiennon et al. [62] and Wu et al. [67] very generally seek "good" summaries, asking annotators "which summary is best". Note that they do condition on length: "how good is this summary, given it is X words long?", but this still represents a *broad* position. They then however describe in detail what properties a "good" summary has and even prescribe sub-dimensions like coherence, purpose and style. For example, "roughly speaking, a good summary is a shorter piece of text that has the same essence of the original – tries to accomplish the same purpose and conveys the same information as the original post" [p.21, 62] • Wu et al. [67] is a bit of an edge case, because they externally focus on "quality" of the summary but in reality, they very precisely define subconcepts of coverage, coherence and amount of abstraction. Their guidelines even include phrases like "Present tense should be preferred" (see p.23 for the full detailed guidelines and definitions). • Ziegler et al. [72] simply ask which outputs are preferred but give detailed instructions for labellers and provide example comparisons labelled by the authors. • Menick et al. [41] also target general concepts of "good" or "bad" answers but then in guidelines, make statements like "helpful answers are better" and "answers which make you read less are better" (p.29, weakly prescriptive).	• OpenAI [47] focus on reducing political bias and improving handling of controversy in ChatGPT. Intended behaviours are defined in precise and detailed guidelines given to human reviewers, containing edge cases and exemplars. • Lu et al. [39] give precise definitions and detailed guidelines on how to interpret coherence, informativeness, safety and engagingness. For evaluation, they take majority vote across three annotators and report interannotator agreement (which is low suggesting remaining subjective scope.) • Thoppilan et al. [63] define fairly clear targets of informativeness, safety and quality, then give detailed guidelines and definitions for how to interpret these concepts. They make it explicit that a single organisational perspective is sought, where *Safety* is defined according to "objectives derived from Google's AI Principles" (p.5). However, there are still some elements of slight subjectivity because they appeal to annotators to "use their commonsense" (p.34). • Nakano et al. [43] specify helpfulness (informativeness) as the key required trait in their WebGPT model, designed for information-seeking dialogue. They provide contractors with a video and detailed instructions "to enable more interpretable and consistent comparisons" (p.6), "to minimize label noise" (p.17), and "to make comparisons as unambiguous as possible" (p.17).
Descriptive — Allow and encourage subjective interpretation of *dimensions* by providing no definitions.	• Zheng et al. [70] collect ELO scores from internet users rating different LLMs and collate them into a large-scale dataset. The interface only asks which model output is "better", if both are "bad" or if there is a tie. It does not specify which dimensions could (or should) contribute to "better" nor define what "better" means. • The StanfordNLP [61] Human Preferences (SHP) dataset is sourced from upvoting behaviours on different subredddits. While they specify in the Data Card that upvotes corresponds to "helpfulness", we consider this an example of broad-descriptive because in reality, there are no specifications or definitions of dimensions that a Reddit user looks for to upvote one post over another—it is a very broad and descriptive signal of preference. • Bakker et al. [6] could also be considered broad-descriptive because they explicitly seek to collect divergent opinions on a wide range of moral and political topics, and provide little prescription over what opinions individuals can hold. However, they do then target agreement and quality as desirable properties of an opinion consensus summary (adding in some specific dimensions). • Liu et al. [38], in their evaluations using human annotators, ask "How much do you agree that the generated text is aligned with the human value: morality/deontology/non-toxicity?" (p.248). Similarly broad, Liu et al. [36] ask "To what extent does the edited response improve the original response in terms of alignment with human values?" (p.6). This evaluation setting represents the most broad-descriptive statement we can imagine because they ask directly about the meta-goal (and empty signifier) of "alignment" with no dimensions and no definitions.	• OpenAI [45] seek to identify unsafe behaviours by red-teaming GPT-4(V). They do not *a priori* define all the different instantiations of "unsafe" according to OpenAI's principles, and instead hire experts to locate and interpret areas of risk. • Ganguli et al. [19] specify the target behaviour (harm) but offer no clear definitions. They say: "We do not define what "harmful" means, as this is a complex and subjective concept; instead, we rely on the red team to make their own determinations via a pairwise preference choice" (p.4). Note that they do still measure and report inter-annotator agreement between raters and authors (p.8-9). • Askell et al. [3] and Bai et al. [4] stipulate that a value-aligned model is one that is honest, harmless and helpful but do not prescribe their meaning. Bai et al. [4] say "we certainly believe that honesty is a crucial goal for AI alignment" (p.4), but also "[o]ur goal is not to define or prescribe what 'helpful' and 'harmless' mean but to evaluate the effectiveness of our training techniques, so for the most part we simply let our crowdworkers interpret these concepts as they see fit." (p.4). • Bai et al. [5] apply a specific-descriptive framing for writing the constitutions in their constitutional AI framework. Some constitutions include statements like "Identify all ways in which the assistant's last response is harmful, unethical, or socially biased" (p.22), without defining these concepts (despite the possibility for substantial variation across ethical frameworks and different societal structures or hierarchies by community, culture or country). Some of the constitutions are more prescriptive in defining sub-concepts like *harmful*, for example referencing racist, sexist, toxic, illegal, violent, or unethical behavior.

3 Discussion

Appreciating that everyone knows what a horse is Our framework relies on an interplay between dimensions and definitions, but it is not practical to treat all dimensions equally. *Noew Ateny*, a Polish dictionary published in 1745 amusingly includes the definition: "Horse: Everyone can see what a horse is". In a similar vein, it is apparent that evaluating dimensions like "length" of an output face much less debate than societally-subjective dimensions like "harmlessness" or concepts that can be ascribed many different sub-meanings like "good". This implies that the levels of our framework inherently interact with one another because dimensions carry with them an intrinsic or expected level of objectivity, which in turn conditions the necessity or impact of providing a definition.

Disputing alignment at the margins, not the extremes We expect a lesser degree of disagreement on important dimensions and their definitions at the extremes. While values, morals and ethics vary considerably across individuals, cultures and time [17], there are universally-agreed bounds on the range of variations, evident in human rights declarations. These bounds address fundamental basic rights, like the right to life, but remain vague on their practical application in contentious issues like abortion. Similarly for empirical alignment efforts, distinctions between *specific-prescriptive* and *broad-descriptive* positions converge at the extremes: it is a sensible assumption that annotators likely internalise that LLMs shouldn't seek to harm life or property, even if this behaviour is unspecified. That said, boundaries may need to placed on the acceptable range or limit of inclusions and interpretations [31], and failing to do so may result in degenerative outcomes, as seen with the Tay Bot incident [29]. Determining these boundaries is a complex normative issue but a partial solution comes from deciding who forms the sample or pool that provides the alignment signals.

Communicating universals versus particulars Adopting a particular position does not necessarily convey a given philosophy but may add colour to underlying thinking. Placing very weak restrictions on how language models should behave (*broad-descriptive*) is congruous to a cultural relativist position, where individual identity and lived experience condition meaning. Adopting a *specific-prescriptive* stance can underpin a multitude of belief systems. People making locally-bounded assumptions on how models should behave can acknowledge diverse interpretations yet encode specific "designer preferences" or organisational priorities into models [62, 63, 46]. Those making more global and unbounded assumptions may buy into shared interpretations and universalities across different cultures, time periods and peoples. In any approach, it is important to communicate the role that identity and positionality has on scoping, prescription and interpretation, and whether representativeness of the sample conditions the value of the empirical signal. There is growing acknowledgement of how identity conditions reward [12] or risk [1, 19]; yet some widely-used datasets [e.g., 4] make no provisions on annotator identity despite being *descriptive* in scope. In the absence of clear communication, there is a risk of conflating the particular and the universal, as Judith Butler comments: "the universalization of the particular seeks to elevate a specific content to a global condition, making an empire of its local meaning" [p.31, 11]

Acknowledging power dynamics and hegemonies Our framework promotes clear articulation of stakeholder influence in empirical alignment efforts, aiding the understanding of power hierarchies. In the *specific-prescriptive* stance, organisations or model developers impose top-down restrictions. Here, honest communication around how decisions were made and why includes both explaining inclusions—which Ouyang et al. [48] does very well in their attribution of researchers, labellers and OpenAI's role in shaping encoded preferences—and being upfront about the possibility of exclusions—which Thoppilan et al. [63] exemplify in discussing sociocultural and geographical blindspots in their definition of safety. However, even in absence of clear prescription, there is still a risk for hegemonic reinforcement, which could be even more dangerous when disguised. While the *broad-descriptive* position might seem most accepting of human variation and sociocultural difference as a "bottom-up" grassroots approach, it is often a non-representative group who shoulder the responsibility for steering LLM behaviours—whether few US-based crowdworkers [e.g., see 4, 43] or many interested netizens [70]. While our framework does not prescribe who should have a voice in LLM development, it does encourage greater admission of individual, community and organisational involvement. Without these contextual bounds, there is a risk of moral absolutism as a form of digital colonialism, where the values, morals or priorities of US-based technology providers and crowdworkers are imposed on the rest of the world as if it were 'the only way' [42, 65].

4 Conclusion

Practitioners of LLM alignment should avoid relying on empty signifiers, and instead, be more precise in what they are attempting to achieve through empirical alignment datasets. In this paper, we presented a framework for communicating which dimensions are measured as alignment signals and how these dimensions are defined. Without such a shared language, there are dangers from obscurities: specific local particularities may be disguised or passed off as universalities, enforcing primarily western and narrow meanings across peoples, countries or cultures. The discourse around alignment carries with it the ideological imprints of various stakeholders—whether this be technology designers, data labellers, or internet users—and we need to document the role that these humans play in shaping model behaviour. We hope this work encourages such critical reflections and supports transparent communication in alignment efforts that must replace the abstract with the empirical and the rhetorical with the actionable.

Social Impact Statement

Our work is intended primarily for practitioners (whether this be industry, government, academia, open-source communities or other model developers). We provide a tangible framework for operationalising abstract notions of alignment into measurable empirical signals. The broader impact of our work is in clarifying, conceptualising and re-communicating the empirical alignment landscape via documentation of intents. We envisage the practical applications as two-fold. First, our framework has impact as a development tool which can be applied *ex-ante*, to initiate or guide the process of building datasets for LLM alignment and to iterate on assumptions or aims along the way. Secondly, our framework has impact as a communication tool, which can be applied *ex-post* to empirical alignment research so that achievements and framings can be reflected upon and clearly described to other members of the community or external stakeholders using a shared vocabulary. It is important to note that we do not advocate for adopting one approach over another, nor suggest that occupying one quadrant is somehow 'better'—that is, we make no normative calls on what are the right paradigms to conceptualise "alignment" in LLMs, or how decisions "should" be made. The provision of our framework is not for predicting how future LLMs will integrate with wider society and its diverse members, or forecasting where critical mass will accumulate in different quadrants. However, we do believe our work clearly adds value to the *societal* alignment of LLMs at a meta-level. Even though we do not specify which direction to move in, we strongly advocate for knowing and reporting your coordinates. This shared language better conditions what to *expect* when we interact with a model that is said to be *aligned*. By reducing the expectations gap, a greater degree of transparency, documentation and mutual understanding is likely to have a net positive effect on the societal impacts of LLMs, irrespective of the precise development decisions made in the near and distant future.

Acknowledgements

As a component of a wider research agenda on optimising feedback between human-and-model-in-the-loop, this paper has received funding from the MetaAI Dynabench grant. H.R.K's PhD is supported by the Economic and Social Research Council grant ES/P000649/1. P.R received funding through the INDOMITA project (CUP number J43C22000990001) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR). We particularly want to thank Andrew Bean (University of Oxford) who greatly assisted in classifying and coding the relevant literature that populates our framework, and Betty Hou (New York University) who provided valuable feedback and spurred interesting discussion.

References

[1] Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. DICES Dataset: Diversity in Conversational AI Evaluation for Safety. (arXiv:2306.11247).

[2] Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

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

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.

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

[6] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. arXiv:2211.15006v1.

[7] Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

[8] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

[9] Umang Bhatt, Valerie Chen, Katherine M. Collins, Parameswaran Kamalaruban, Emma Kallina, Adrian Weller, and Ameet Talwalkar. 2023. Learning Personalized Decision Support Policies. (arXiv:2304.06701).

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

[11] Judith Butler, Ernesto Laclau, and Slavoj Žižek. 2000. *Contingency, Hegemony, Universality: Contemporary Dialogues on the Left*. Phronesis. Verso, London.

[12] Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. 2023. Everyone Deserves A Reward: Learning Customized Human Preferences. (arXiv:2309.03126).

[13] Paul Christiano. 2021. Clarifying "AI alignment". *Medium*.

[14] Katherine M. Collins, Umang Bhatt, and Adrian Weller. 2022. Eliciting and Learning with Soft Labels from Every Annotator. (arXiv:2207.00810).

[15] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

[16] Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. (arXiv:2306.16388).

[17] Ronald Fischer. 2017. Personality, Values, Culture. In *Personality, Values, Culture: An Evolutionary Approach*, Culture and Psychology, pages i–ii. Cambridge University Press, Cambridge.

[18] Iason Gabriel. 2020. Artificial Intelligence, Values and Alignment. *Minds and Machines*, 30(3):411–437.

[19] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858v2.

[20] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

[21] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. arXiv:2209.14375v1.

[22] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

[23] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

[24] Patrick Haller, Ansar Aynetdinov, and Alan Akbik. 2023. OpinionGPT: Modelling Explicit Biases in Instruction-Tuned LLMs. (arXiv:2309.03876).

[25] Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. In *The Semantic Web – ISWC 2014*, Lecture Notes in Computer Science, pages 486–504, Cham. Springer International Publishing.

[26] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. (arXiv:1805.00899).

[27] Atoosa Kasirzadeh and Iason Gabriel. 2022. In conversation with Artificial Intelligence: Aligning language models with human values. (arXiv:2209.00731).

[28] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of Language Agents. (arXiv:2103.14659).

[29] Tae Wan Kim, Thomas Donaldson, and John Hooker. 2018. Mimetic vs Anchored Value Alignment in Artificial Intelligence. (arXiv:1810.11116).

[30] Hannah Rose Kirk, Andrew M Bean, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

[31] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. (arXiv:2303.05453).

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

[33] Raphael Koster, Balaguer Jan, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams, Lucy Campbell-Gillingham, Phoebe Thacker, Matthew Botvinick, and Christopher Summerfield. 2022. Human-centred mechanism design with Democratic AI. *Nature Human Behaviour*.

[34] Ernesto Laclau. 1996. *Emancipation(s)*, 1. publ edition. Phronesis. Verso, London.

[35] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: A research direction. (arXiv:1811.07871).

[36] Ruibo Liu, Chenyan Jia, Ge Zhang, Ziyu Zhuang, Tony X. Liu, and Soroush Vosoughi. 2023. Second Thoughts are Best: Learning to Re-Align With Human Values from Text Edits. arXiv:2301.00355v2.

[37] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training Socially Aligned Language Models in Simulated Human Society. (arXiv:2305.16960).

[38] Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning Generative Language Models with Human Values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252, Seattle, United States. Association for Computational Linguistics.

[39] Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2022. Towards Boosting the Open-Domain Chatbot with Human Feedback. arXiv:2208.14165v1.

[40] Claude Lévi-Strauss. 1987. *Introduction to the work of Marcel Mauss*. Routledge & Kegan Paul, London.

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

[42] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4):659–684.

[43] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332v3.

[44] Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143.

[45] OpenAI. 2023. GPT-4V(ision) system card.

[46] OpenAI. 2023. How should AI systems behave, and who should decide?

[47] OpenAI. 2023. Snapshot of ChatGPT model behavior guidelines.

[48] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155v1.

[49] Alicia Parrish, Sarah Laszlo, and Lora Aroyo. 2023. "Is a picture of a bird a bird": Policy recommendations for dealing with ambiguity in machine vision models. (arXiv:2306.15777).

[50] Xiangyu Peng, Siyan Li, Spencer Frazier, and Mark Riedl. 2020. Reducing Non-Normative Text Generation from Language Models. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 374–383, Dublin, Ireland. Association for Computational Linguistics.

[51] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[52] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. (arXiv:2305.18290).

[53] Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

[54] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190.

[55] Stuart J. Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Allen Lane, an imprint of Penguin Books, London.

[56] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. LaMP: When Large Language Models Meet Personalization. (arXiv:2304.11406).

[57] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

[58] Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training Language Models with Language Feedback. arXiv:2204.14146 [cs].

[59] Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, and Reinhard Koch. 2022. Is one annotation enough? - A data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35:33215–33232.

[60] Viktoriia Sharmanska, Daniel Hernandez-Lobato, Jose Miguel Hernandez-Lobato, and Novi Quadrianto. 2016. Ambiguity Helps: Classification With Disagreements in Crowdsourced Annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2194–2202.

[61] StanfordNLP. 2023. Stanford Human Preferences Dataset.

[62] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. arXiv:2009.01325v3.

[63] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. arXiv:2201.08239 [cs].

[64] Alan Turing. 1947. Lecture on the Automatic Computing Engine (1947). In B J Copeland, editor, *The Essential Turing*, page 0. Oxford University Press.

[65] Kush R. Varshney. 2023. Decolonial AI Alignment: Vi\'{s}esadharma, Argument, and Artistic Expression. (arXiv:2309.05030).

[66] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

[67] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively Summarizing Books with Human Feedback. arXiv:2109.10862v2.

[68] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. (arXiv:2306.01693).

[69] Wanqi Xue, Bo An, Shuicheng Yan, and Zhongwen Xu. 2023. Reinforcement Learning from Diverse Human Preferences. (arXiv:2301.11774).

[70] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. (arXiv:2309.11998).

[71] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. (arXiv:2305.11206).

[72] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593v2.