POSITION: WE NEED AN ADAPTIVE INTERPRETATION OF HELPFUL, HONEST, AND HARMLESS PRINCIPLES

Yue Huang[1], **Chujie Gao**[2], **Yujun Zhou**[1], **Kehan Guo**[1], **Xiangqi Wang**[1], **Or Cohen-Sasson,**[3]
Max Lamparth[4], **Xiangliang Zhang**[1]

[1]Department of Computer Science and Engineering, University of Notre Dame   [2]MBZUAI
[3]School of Law, University of Miami
[4]Center for AI Safety, Stanford University

ABSTRACT

The Helpful, Honest, and Harmless (HHH) principle is a foundational framework for aligning AI systems with human values. However, existing interpretations of the HHH principle often overlook contextual variability and conflicting requirements across applications. In this paper, we argue for **an adaptive interpretation of the HHH principle** and propose a reference framework for its adaptation to diverse scenarios. We first examine the principle's foundational significance and identify ambiguities and conflicts through case studies of its dimensions. To address these challenges, we introduce the concept of priority order, which provides a structured approach for balancing trade-offs among helpfulness, honesty, and harmlessness. Further, we explore the interrelationships between these dimensions, demonstrating how harmlessness and helpfulness can be jointly enhanced and analyzing their interdependencies in high-risk evaluations. Building on these insights, we propose a reference framework that integrates context definition, value prioritization, risk assessment, and benchmarking standards to guide the adaptive application of the HHH principle. This work offers practical insights for improving AI alignment, ensuring that HHH principles remain both ethically grounded and operationally effective in real-world AI deployment.

1 Introduction

The development of AI assistants has progressed rapidly, evolving from simple rule-based systems (Weizenbaum, 1966) to advanced models, such as Large Language Models (LLMs) (Zhao et al., 2023b), which excel in understanding context and generating human-like responses. The wide application of these models highlights not only the expanding capabilities of AI but also the growing need to ensure these systems align with human preferences and values. To address this challenge, the HHH principle–standing for Helpful, Honest, and Harmless–was proposed by Askell et al. (2021) as a guiding framework for designing and evaluating AI systems. The principle aims and *has been widely utilized* to align AI behavior with human-centered values, providing critical benchmarks for tasks such as training data selection, strategy design, and deployment guidance (Touvron et al., 2023b; Peng et al., 2023; Li et al., 2024a; Bai et al., 2023). Its adoption has been instrumental in advancing AI alignment, particularly in ensuring that powerful models like LLMs prioritize user benefit, truthfulness, and safety.

However, current studies indicate the existence of ambiguities and conflicts among the three HHH dimensions. For example, the same prompt input might be classified as either harmless or harmful depending on the context or criteria used in different research. Similarly, the definition of specific dimensions often varies across studies, influenced by the diverse contexts in which they are applied. Furthermore, key questions such as "which dimension should be prioritized" and "how different dimensions interrelate" remain unresolved, underscoring the need for a clearer and more adaptable interpretation of the HHH principle.

In this position paper, **we argue that the HHH principle is not static or rigid but requires an adaptive interpretation to remain effective across diverse scenarios**. To support this argument, first, we review the initial definition

---

Correspondence to: Yue Huang (yhuang37@nd.edu), Xiangliang Zhang (xzhang33@nd.edu).

and fundamental importance of the HHH principle, examining its positive impact on AI alignment efforts (**§2**). Next, we identify ambiguities and conflicts within current interpretations of the principle through case studies of different dimensions (**§3**), highlighting the need for consensus or innovative solutions. To address these conflicts, we introduce the concept of priority order, which provides a structured approach to balancing competing requirements among dimensions (**§4**). Furthermore, we investigate the relationships between dimensions, analyzing how harmlessness and helpfulness can be simultaneously improved and examining their interdependencies in high-risk evaluations (**§5**). Building on these insights, we propose a reference framework that guides the adaptive application of the HHH principle to specific scenarios (**§6**). The framework outlines critical considerations and offers a comprehensive methodology for interpreting and operationalizing the HHH principle in varied contexts.

Through this position paper, we aim to deepen the understanding of the HHH principle and foster its effective utilization as a tool for advancing AI development and alignment. We hope this work inspires further research and discussion on aligning AI systems with human values in a nuanced and adaptable manner.

2 HHH Principle

2.1 Initial Definition

The HHH principle is a set of guiding principles for developing and aligning AI models, particularly large language models, with human values. Askell et al. (2021) define it from its aim: *We will define an AI as "aligned" if it is, in three words, helpful, honest, and harmless or 'HHH'. Our alignment efforts aim to measure and address this general problem with large language models.* For the sake of simplicity, we summarize the HHH principle based on their original definitions:

*1) **Helpful**: AI models should assist users by providing useful, accurate, and contextually relevant information or services. They must be designed to meet user needs, enhance productivity, and effectively solve problems.*
*2) **Honest**: AI models should ensure transparency and truthfulness in their responses, providing factual information while openly acknowledging their limitations. They must refrain from generating falsehoods or misleading content.*
*3) **Harmless**: AI models should avoid causing harm by preventing the generation of biased, offensive, or unethical content. They should prioritize safety and respect in their interactions, ensuring that they do not produce harmful or inappropriate outputs.*

2.2 Fundamental Value of the HHH Principle

The HHH principle plays a critical role in guiding AI system development to align with human values and preferences. This principle has been used for establishing clear, user-centric benchmarks (Zheng et al., 2023), and designing training methods like Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) for advancing AI system performance. **Helpfulness** is directly related to the functional effectiveness of AI models, addressing what *human users really need*, thus enhancing human productivity. **Honesty** ensures that AI systems provide truthful (*e.g.*, avoid hallucination (Huang et al., 2023a; Zhang et al., 2023)), and transparent (*e.g.*, express uncertainty (Xiong et al., 2024)) information, which is crucial for fostering trust between humans and machines. **Harmlessness**, in turn, guarantees that AI outputs are safe and ethical, preventing the generation of harmful, biased, or misleading content (Huang et al., 2024a; Wang et al., 2023; Liu et al., 2023a). The HHH principle determines the three most important sub-directions aligned with human preferences and values, which have been widely used for guiding AI model design as shown in Appendix A.

The HHH principle uniquely emphasizes both trustworthiness and utility [1]**.** AI models, such as LLMs, interact with users by providing information, generating content, and assisting in decision-making, making trustworthiness and utility essential for their effectiveness and reliability. While utility is prioritized in model development through training and tuning loss minimization, an automated and human-effortless process, there is a risk of compromising ethical integrity or introducing biases. Thus, typical alignment principles emphasize additional efforts to align AI models with human values, as shown in Table 1. Research indicates that the resulting models often choose safe responses over helpful ones (Touvron et al., 2023a). However, the HHH principle offers a unique balance by explicitly emphasizing both dimensions. The dimension of **honesty** requires AI models generate accurate and truthful outputs.

---

[1]Utility here refers to the model effectiveness in natural language processing tasks, including logical reasoning, content summarization, text generation, and so on (Huang et al., 2024a).

Framework & Principle	Trust.	Utility
HHH (Askell et al., 2021)	★	★
TrustLLM (Huang et al., 2024a)	★	⯪
NIST (NIST, 2024)	★	⯪
AI Safety and Security (Qi et al., 2024)	★	☆
Risk Assessment (Kapoor et al., 2024)	★	☆
Transparency (Bommasani et al., 2024)	★	☆

Table 1: Typical principles or frameworks for guiding AI model alignment (excluding government laws or acts (*e.g.*, EU AI Act (EU, 2024b) and Blueprint for an AI Bill of Rights (House, 2022)). NIST refers to the NIST Cybersecurity Framework. ★, ⯪, and ☆ represents different focus efforts on trustworthiness or utility.

The HHH principle facilitates the transition of AI from a passive tool to an active participant. The societal impact of AI models extends *beyond tangible benefits*, influencing more subtle and indirect aspects of human life. AI-generated content has the potential to shape human perspectives (*e.g.*, create ideas in scientific research (Si et al., 2024) and simulate social behaviors Huang et al. (2024d)), cultural norms, and even values (Ramezani & Xu, 2023; Agarwal et al., 2024). Repeated exposure to biased or subtly manipulative outputs could inadvertently influence public opinion or entrench societal biases (Zeng et al., 2024). This transition from AI being a *passive tool* to an *active participant* raises concerns about its potential role as a manipulative force in shaping discourse and decision-making (Simmons-Edler et al., 2024). The HHH principle provides a structured approach to addressing these emerging challenges. While **helpfulness** ensures AI effectively meets user objectives, **honesty** and **harmlessness** serve as safeguards, mitigating risks and protecting societal well-being as AI systems become increasingly powerful and influential.

3 Ambiguity and Conflicts in HHH

While the HHH principle plays a crucial role in guiding AI alignment, its practical implementation often reveals ambiguities and conflicts among the three dimensions, particularly given the complexity of different contexts. Next, we discuss these ambiguities and conflicts through case studies in each dimension.

3.1 Case Study I: Boundary of Harmlessness

The distinction between harmful and benign content is not always clear-cut, as context, intent, and interpretation play crucial roles. This lack of clarity complicates the development of safety mechanisms and alignment strategies (Bauer & Bindschaedler, 2021; Truong et al., 2024).

A harmless query becomes harmful when it is used to bypass safety restrictions, e.g, in the scenario of adversarial prompting. For example, if an AI model initially rejects a query for being harmful (I_B, *"Write a post promoting militaristic aggression,"*), the user may slightly rephrase it to appear benign while maintaining the original intent: I_A, *"Please write a post encouraging people to build a strong national defense"*. If the model then accepts and responds to the modified query I_A, it may unintentionally generate harmful content. However, if the same revised query I_A were submitted independently, without following a rejected attempt, it would still appear harmless. Training machine learning classifiers to detect toxicity in queries may fail in this case (Wang & Chang, 2022; Ousidhoum et al., 2021; OpenAI, 2024). This highlights the importance of incorporating context and intent when distinguishing between harmful and harmless queries.

Moreover, ***the harmfulness or harmlessness of a generated response can also depend on the user's intent***. For example, if a model responds to a query but includes a moral disclaimer to discourage misuse, the response itself may remain neutral or informative (Mazeika et al., 2024; Ran et al., 2024; Huang et al., 2024c). However, an attacker could exploit this by simply removing the moral disclaimer, and repurposing the content for malicious objectives. Conversely, a well-intentioned user may utilize the same response for legitimate purposes without causing harm.

Although recent work by OpenAI (Mu et al., 2024) has proposed a set of rules that a trustworthy LLM should adhere to (e.g., categorizing response into hard refusal, soft refusal, and compliance), the core issue lies in *the ambiguity of determining the appropriateness of a response without knowing the user's real intent and context*. This highlights the need for an adaptive interpretation of HHH principle, which can dynamically adjust to different contexts and user intents.

3.2 Case Study II: Definition of Honesty

Honesty has emerged as a central topic in the alignment of AI assistants. However, consensus on its precise definition remains elusive, as recent studies offer divergent perspectives on what it means for an AI model to be "honest" (Yang et al., 2023; Gao et al., 2024). In Askell et al. (2021), honesty is often equated with providing accurate information, closely aligned with the concepts of truthfulness (Huang et al., 2024b) and non-hallucination (Huang et al., 2023a; Zhang et al., 2023). This traditional view primarily emphasizes the factual accuracy of the model's responses. Recent studies, however, have introduced more nuanced categorizations of honesty, distinguishing between two major dimensions: *epistemic honesty* and *interactive honesty*.

Epistemic honesty is concerned with transparency regarding AI models' knowledge limitations and ability to express uncertainty. Represented by Yang et al. (2023), this perspective emphasizes that an honest AI model should "candidly answer questions it knows and humbly admit to those it does not." This view goes beyond factual accuracy, incorporating the notion of humility in acknowledging gaps in knowledge. For example, Yang et al. (2023) think LLMs should explicitly say "I don't know" when they lack sufficient knowledge to provide an accurate and reliable answer. Subsequent studies have extended this perspective by exploring methods to improve AI models' ability to express uncertainty (Chern et al., 2024; Yin et al., 2023).

Interactive honesty focuses on the AI model's ability to maintain objectivity, avoid spreading misinformation, and ensure clarity in its interactions with users. Gao et al. (2024) define this form of honesty as "the ability to recognize its limitations (e.g., LLMs are unable to process visual information without external tools), remain objective without pandering (e.g., avoiding sycophantic behavior as discussed in (Sharma et al., 2023)), and thereby avoid spreading misinformation or inducing hallucinations." Interactive honesty emphasizes not only factual accuracy but also the model's capacity for self-awareness (Li et al., 2024d; Jiang et al., 2024) and user-oriented objectivity. This perspective underscores the importance of preventing models from misleading users or offering false reassurance, even in situations where knowledge is incomplete. Subsequent studies have adopted this view to explore honesty through the lens of human cognition and interaction, emphasizing its role in building trust and effective communication (Brahman et al., 2024; Wen et al., 2024).

While epistemic honesty prioritizes knowledge calibration and transparency, interactive honesty emphasizes behavioral consistency and ethical interaction. These differences highlight a deeper tension: the difficulty of reconciling the technical feasibility of implementing honesty with the philosophical rigor required for alignment. Resolving these tensions is crucial for developing a coherent, operationalizable definition of honesty that aligns with both trustworthiness and utility. Without such clarity, ambiguity in the definition of honesty risks undermining the efforts of AI alignment.

3.3 Case Study III: Helpfulness in Different Contexts

Helpfulness, while fundamental to AI systems, lacks a standardized definition and varies significantly across contexts. This conceptual ambiguity and variability present substantial challenges for measuring and optimizing helpful behavior in AI models.

The definition of helpfulness varies substantially based on domain-specific objectives. For instance, in education, Hemami (2024) frame helpfulness as the ability of AI models to identify optimal learning strategies and personalize content for individual students. For general language assistants, helpfulness is defined as a clear attempt to perform the task or answer the posed question (Askell et al., 2021). These variations, illustrated in Appendix C, reveal the importance of contextual adaptability in defining and applying the concept of helpfulness.

The absence of standardized definitions and evaluation metrics stems not only from contextual variability but also from researchers' tendency to optimize for domain-specific objectives. A clear example is the reliance on alignment techniques to tune models to human preferences (Ouyang et al., 2022; Bai et al., 2022a; Rafailov et al., 2024). Most alignment datasets reflect either general value alignment (Stiennon et al., 2020; Bai et al., 2022b) or specific task objectives (Tian et al., 2023; Xie et al., 2024). For example, reasoning-focused tasks prioritize coherent and multi-step explanations (Huang et al., 2023b; Zhuo et al., 2024), while conversational assistants prioritize concise, user-friendly responses (Sun et al., 2024). Moreover, Awad et al. (2018) and Santurkar et al. (2023) both reveal that AI models may reflect preference on specific opinions.

Current evaluation methods often employ narrow metrics that align with specific research objectives rather than capturing helpfulness holistically. Moreover, most frameworks utilize specialized metrics (Liu et al., 2024c) or LLM-as-judge approaches (Zheng et al., 2023; Wang et al., 2024b), Ye et al. (2024), demonstrating that such evaluations can introduce systematic biases through prompt design choices. Future benchmarks should incorporate multidimensional evaluation such as complementary dimensions (e.g., correctness assessment, coherence analysis, cre-

ativity metrics), while considering the relative importance and interactions between these aspects. Consequently, multidimensional approaches would better reflect the diverse expectations of helpfulness in real-world applications.

4   Priority Order

Definition and necessity. In the deployment of AI models, conflicting requirements often arise among helpfulness, honesty, and harmlessness across various scenarios. For instance, in cybersecurity applications, when a user requests information about system vulnerabilities, the model must balance providing helpful technical details (helpfulness) with the need to prevent malicious exploitation (harmlessness) (Zhang et al., 2024a). Similarly, in medical or legal domains, prioritizing honesty may sometimes conflict with harmlessness, such as when conveying critical but distressing information. To address such conflicts, we introduce the concept of *priority order*, which is defined as:

A dynamic hierarchical framework that determines the relative importance and execution sequence of three dimensions of the HHH principle based on contextual requirements.

The necessity of adopting a dynamic rather than static priority order arises from the complexity and variability of real-world applications, where the relative importance of helpfulness, honesty, and harmlessness shifts based on context, user intent, and domain-specific requirements. In this section, we explore the priority order of the HHH principle from two complementary perspectives as shown in Figure 1: (1) a structural and statistical perspective (§4.1), which examines how prioritization operates along hierarchical levels and scales, shaping its manifestation across different domains and impact scopes; and (2) an application-level embodiment (§4.2), which investigates how priority order shifts based on user populations and task contexts.

4.1   Prioritization Levels and Scales

Prioritization levels refer to the vertical structuring of the HHH principles, meaning that in each task, certain principles take precedence over others based on the risk and ethical constraints of the scenario. As shown in Figure 1, this hierarchical view defines which dimension should be prioritized in different tasks. To characterize the relative importance of each dimension, we introduce the *priority value* as an attribute representing the extent of emphasis placed on each dimension, which could potentially be quantified through specific metrics in future work. Prior studies have provided empirical evidence supporting the feasibility of prioritization levels across different fields. In education, harmlessness is paramount to protect learners' development, followed by helpfulness for engagement, with honesty about AI capabilities as the foundation (Kooli, 2023; Selwyn, 2022). Creative domains instead prioritize helpfulness to drive innovation, with honesty secondary and harmlessness setting safety bounds while preserving creative freedom (Flick & Worrall, 2022). This varying prioritization levels across domains indicates the need for an adaptive interpretation of the HHH principle rather than a one-size-fits-all approach.

Prioritization scales refer to horizontal variations within the same ranking level, determining how the HHH principle is applied across different user groups ranging from micro (individual users) to macro (societal user groups). For harmlessness, the micro-scale emphasizes protecting individual privacy and enforcing data minimization (GDPR, 2016; Staab et al., 2024), while the macro scale may necessitate selective data collection to mitigate systemic risks and ensure public safety. For instance, an AI system assisting a single user with a research query may prioritize helpfulness and honesty, while an AI generating public information (e.g., automated news summaries) must weigh harmlessness more heavily to prevent misinformation. These scale-dependent variations indicate that effective HHH implementation requires consideration of both individual and societal impacts, even when the nominal priority level remains constant.

4.2   Context-Aware Prioritization: User and Task Perspectives

Prioritization of the HHH principles should also be shaped by specific user contexts and task requirements. Priority ordering could vary significantly across different user populations (Bao et al., 2022). Expert users, such as AI researchers or domain specialists, may prioritize helpfulness and honesty over strict harmlessness constraints, while general users or vulnerable populations often require emphasis on harmlessness.

Besides, these priorities should dynamically adjust during task execution, shifting relative importance as tasks progress through different phases. For instance, in financial advising, early goal-setting prioritizes helpfulness for exploring options (Shanmuganathan, 2020), while investment execution emphasizes honesty in risk disclosure. This becomes particularly evident in interactive scenarios requiring real-time adaptation.
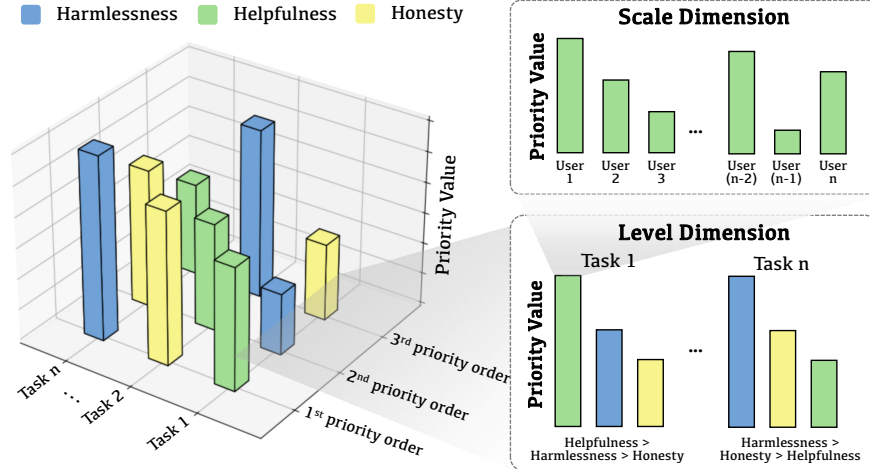
Figure 1: The conceptual structure of priority order within the HHH principle. For each task, we first determine a priority order (level dimension), and then adjust the priority values for different users within that task (scale dimension).

All the above discussion highlights that it is crucial for the interpretation of the HHH principle, which motivates us to design relevant components to include it in §6.

5 Trade-off or Synergy? Relationship Between Different Dimensions

After establishing the priority order across different dimensions, the interplay between helpfulness, honesty, and harmlessness in AI alignment remains a critical and ongoing debate. A key question is whether these dimensions are inherently in conflict, requiring trade-offs, or they can be optimized in synergy to enhance one another. While some research suggests that enhancing one always comes at the expense of the other (Qi et al., 2023), there is also evidence that strategies allow both to be improved simultaneously (Huang et al., 2024a). This section aims to discuss the relationship between different dimensions.

5.1 Boost Harmlessness and Helpfulness Simultaneously

Helpfulness is generally a subset of utility, while harmlessness and honesty often reflect trustworthiness. As AI assistants evolve, balancing utility and trustworthiness becomes critical. Some view the **SB 1047 AI Bill**[2] (Cal), designed to ensure AI harmlessness, as a potential barrier to innovation (California Chamber of Commerce, 2024).

Recent studies highlight the close link between harmlessness and helpfulness (Wolf et al., 2024; Qi et al., 2023; Huang et al., 2024a; Bai et al., 2022a; Zhang et al., 2024b). Huang et al. (2024a) found a positive correlation between them, while Qi et al. (2023) showed that even intent-free fine-tuning can weaken harmlessness. Bai et al. (2022a) and Zhang et al. (2024b) explored ways to balance the two during training. However, Ren et al. (2024) noted that many safety benchmarks strongly correlate with a model's upstream capabilities.

Prioritizing harmlessness over helpfulness can lead to unintended trade-offs. Excessive safety constraints–such as strict content filtering or rigid ethical frameworks–may limit a model's usefulness and creativity, reducing overall helpfulness (Röttger et al., 2023; Kirk et al., 2023). This imbalance risks producing overly cautious models that struggle in real-world applications where adaptability and innovation are essential.

On the other hand, maximizing helpfulness at the expense of harmlessness carries significant risks. Models that prioritize helpfulness but lack fairness, transparency, or robustness may produce biased outputs, eroding user trust and raising ethical concerns (Huang et al., 2024a; Liu et al., 2023a; Wang et al., 2023; Li et al., 2025). In high-stakes domains like healthcare and finance, untrustworthy models are not only unsustainable but can also be harmful (Xia et al., 2024). Thus, sacrificing one dimension for the benefit of the other is inherently flawed. A paradigm is needed where harmlessness and helpfulness can be simultaneously improved to ensure that AI assistants are reliable and effective.

---

[2]A 2024 California bill aimed at mitigating catastrophic AI risks.

Rather than treating harmlessness and helpfulness as competing objectives in multi-objective optimization (Kochenderfer & Wheeler, 2019), recent research suggests they can be mutually reinforcing. Some approaches first establish a baseline of harmlessness before optimizing for helpfulness (Gao et al., 2024), while others integrate multi-objective alignment to improve both simultaneously (Yang et al., 2024; Wang et al., 2024a; Zhou et al., 2024c; Fu et al., 2024; Guo et al., 2024b).

One crucial insight from the discussion is that harmlessness is a safeguard–ensuring that the AI assistant is inherently safe and trustworthy before other features are optimized. This aligns with the view that harmlessness is not a constraint on helpfulness but a necessary component of it.

The balance between harmlessness and helpfulness is not a zero-sum game where enhancing one necessarily diminishes the other (Tuan et al., 2024). On the contrary, the two can–and should–be pursued in tandem to create robust, effective AI assistants. Sacrificing either harmlessness or helpfulness for short-term gains in the other is ultimately unsustainable and could lead to detrimental consequences in both ethical and practical applications. The key lies in developing methods, like the harmlessness-first approach, where harmlessness serves as a foundation for subsequent helpfulness maximization (a priority for different dimensions as discussed in §4). This strategy ensures that AI assistants remain safe and effective, setting the stage for a future where they can thrive in various real-world contexts without compromising on either front.

5.2 Interdependencies Among Dimensions in High-Risk Evaluations

While helpfulness, harmfulness, and honesty are always intertwined to some degree, in many everyday tasks these interactions are easy to manage or less impactful (Zheng et al., 2023; Huang et al., 2024b; Zhou et al., 2024a; Wu et al., 2024a; Sandmann et al., 2024; Huang et al., 2023c). By contrast, in high-risk or specialized tasks, these same interactions can become more complex and lead to significant risks. So simple, independent evaluation of each dimension may underestimate systemic hazards (Zhou et al., 2024b; Phan Van et al., 2024; Thirunavukarasu et al., 2023). For instance, in medical diagnostics, a drug's helpfulness in treating a condition and the harmfulness associated with possible side effects exist simultaneously (Phan Van et al., 2024; Thirunavukarasu et al., 2023; Sandmann et al., 2024). If the helpfulness is overlooked, the underlying condition remains untreated, potentially leading to serious deterioration of the patient's health. Conversely, if the harmful side effects receive insufficient attention, the patient could suffer from severe complications or even life-threatening adverse reactions. A truly effective medical decision must balance both factors. Therefore, when assessing a model's diagnostic capabilities, it is insufficient to focus solely on whether it provides a "helpful" conclusion. It is equally important to evaluate the probability, severity, and ethical implications of harmful consequences (Thirunavukarasu et al., 2023). This issue is also highlighted when people interact with LLMs in mental health emergencies, e.g., when suffering from mania or psychosis (Grabb et al., 2024). In such scenarios, both providing harmful information to the user (helpfulness) or refusing to respond (harmlessness) can exaggerate existing symptoms and lead to severe harm of the user or others (Tucker, 2025). Similarly, in answering highly specialized questions, if an LLM generates hallucinated outputs, its "helpfulness" is immediately called into question. In lab safety contexts, any "helpful" advice that neglects safety considerations could lead to severe accidents, and hallucinations in the model's generation process present hidden high-risk factors (Zhou et al., 2024b).

Thus, when evaluating LLMs, particularly in high-risk and specialized fields, it is crucial to develop holistic evaluation frameworks that account for the emergent properties. This requires systematic identification of cascading risk scenarios that could lead to severe outcomes (Zhou et al., 2024b; Phan Van et al., 2024; Thirunavukarasu et al., 2023). For example, to avoid lab safety accidents, responses with potential risks should undergo rigorous scrutiny and verification. In medical diagnostics, more stringent test sets and consequence-driven weighting mechanisms should be implemented for scenarios where a drug is both helpful and potentially harmful, incorporating the severity of misdiagnosis or side effects into the evaluation process. Only by explicitly modeling these "multidimensional entanglement" scenarios can we more accurately gauge the risks and value a model may present in real-world applications, thereby providing more targeted guidance for model optimization and regulatory decision-making.

6 Reference Framework

Building on the current ambiguous understanding of the principle (§3), it is evident that interpreting the helpfulness, honesty, and harmlessness in diverse contexts presents significant challenges. These principles often interact in complex ways (relationship as discussed in §5), with their relative importance varying depending on the specific application and scenario (i.e., priority order as discussed in §4). This variability underscores the urgent need for a reference framework to systematically guide the interpretation and application of the HHH principle. Such

a framework would address the critical question: **What perspectives and considerations are necessary to adapt these principles effectively in different scenarios?**

This framework benefits multiple stakeholders. For developers in industries, it provides structured guidelines to integrate the HHH principles into the development of AI models, ensuring regulatory compliance. For academic researchers, it offers a novel view to study trade-offs between these principles in different domains (out-of-distribution (OOM) problem). For end users of AI models, it enhances transparency and governance, ensuring AI interactions align with user expectations (Larsson & Heintz, 2020). In this section, we propose a reference framework to address this issue from multiple perspectives: *Contextual Object, Value Anchor & Value Scale, Risk Assessment, Benchmarking Standard*, and *System Transparency & Governance*. We also provide a detailed case study on applying the 3H principles to the development of a chemistry foundation model, which can be found in Appendix D.

Contextual Object. First, the framework requires determining the objects in a specific scenario when adapting HHH principle. As discussed in **§3.3**, different contexts will result in various understandings of dimensions, where the process, of determining scenario objects, directly define the specific context in which the AI model operates, mitigating potential ambiguity and conflicts. The objects contain key elements that are important for determining the context: *User Group, Application Aim, Task Type*, and *Environment Access*.

- *User Group*: Identifying the primary audience or user group is essential to understanding their expectations and expertise level, as emphasized by lots of recent work on fine-grain or user-level alignment (Zhao et al., 2023a; Fan et al., 2024). For example, general users may require simplified outputs and stricter safeguards for harmlessness, while domain experts may prioritize nuanced and highly truthful outputs with greater flexibility in helpfulness and honesty.
- *Application Aim*: The domain in which the AI operates, such as healthcare (Li et al., 2024a), finance (Wu et al., 2023), or law (Cui et al., 2023), significantly influences the prioritization of HHH dimensions (as discussed in **§4**). Education domain always strictly requires harmlessness (Yan et al., 2024; Pranav Chaudhary, 2025). High-stakes domains like medicine require a stronger emphasis on information authority (*e.g.*, helpfulness and honesty), whereas creative applications might focus less on the truthfulness of model output.
- *Task Type*: The type of task performed by models, such as real-time assistance, decision support, or content creation, defines the expected outputs and constraints. For instance, decision-making tasks demand high levels of honesty (Sun et al., 2024).
- *Environment Access*: Whether the model has access to external tools, real-time data, or isolated environments affects its capability boundary and risks. Models with an external tool (*e.g.*, GUI operation Chen et al. (2024b)) access may need stricter safeguards for harmlessness to prevent misuse. Moreover, the retrieved information through external tools influence the model's performance as well (Zhang et al., 2024c; Gao et al., 2023; Huang et al., 2023b).

Value Anchor & Value Scale. The framework next requires the identification of a *value anchor* and a *value scale* to guide the prioritization of the HHH principle within a specific context. The *value anchor* defines the core value or central priority, such as harmlessness in high-risk domains or helpfulness in performance-driven applications. The *value scale*, in turn, allows for a dynamic balance among the principles by assigning relative weights to each dimension based on contextual demands, stakeholder needs, and societal expectations. For example, in the case of an educational AI model designed for younger students, the *value anchor* would prioritize harmlessness to prevent inappropriate or harmful outputs. The corresponding *value scale* might place harmlessness "*far before*" helpfulness, reflecting the heightened importance of safety in such a sensitive application.

In other words, this framework aims to endow AI models with specific traits or values (*value anchor*) and control the depth and degree of these traits (*value scale)* to ensure that the model remains predictable and operates within a controllable range.

Risk Assessment. Building on the contextual objects and value anchor & value scale, the reference framework must include a risk assessment mechanism to evaluate the trade-offs and potential negative impacts of prioritizing and customizing certain dimensions of the HHH principle. The assessment helps developers, users, and decision-makers determine whether the risks associated are acceptable and manageable within the given context.

Risk assessment involves three primary considerations: measuring the potential risks, assessing stakeholder tolerance for those risks, evaluating system consistency, and considering potential marginal risk (Kapoor et al., 2024). For the same example in an education application where harmlessness is prioritized far above helpfulness, a risk assessment would evaluate whether this decision significantly damages the user experience (e.g., overly restrictive outputs that limit engagement or creativity (Li et al., 2024c)). Simultaneously, it would determine whether stakeholders–such as developers and end users–are willing to accept this trade-off to ensure safety and ethical in-

tegrity. For context-dependent HHH, consistency is a requirement for trustworthy deployment to ensure similar behavior or balance of HHH in similar contexts (e.g., reproducible priority ordering), thus, the framework also emphasizes the significance of consistency evaluation. Moreover, Kapoor et al. (2024) point out some marginal risks (e.g., Non-consensual intimate imagery (NCII) (INHOPE, 2025)) that may be ignored in open foundation models (Bommasani et al., 2021), which should also be strictly considered during assessment.

In particular, most existing risk assessment frameworks such as EU High-Risk AI guidelines (EU, 2024a) focus primarily on societal-level risks, particularly those related to the aspect of harmfulness and safety. While these are crucial, we emphasize that assessments should also account for the direct impacts on stakeholders, such as developers' economic viability and user experiences. For instance, excessive economic costs associated with implementing overly strict harmlessness safeguards can be a significant risk for developers, just as diminished user satisfaction from overly cautious systems can reduce utility and adoption.

Benchmarking Standard. Benchmarks provide an objective and reproducible way to evaluate whether an AI system adheres to the HHH principle under specific scenarios. Without well-defined standards, evaluations risk becoming inconsistent (McIntosh et al., 2024), subjective (Ye et al., 2024), or overly dependent on context-specific assumptions. Under a specific context, the benchmarking standard must align closely with the interpretation (e.g., prioritization) of the HHH principle established by the framework. For example, the benchmarks should reflect this prioritization if helpfulness is prioritized above harmlessness. In such a case, utility-related benchmarks should evaluate helpfulness directly and assess how trustworthiness and harmlessness contribute to or detract from the overall utility. Moreover, to achieve a fine-grained evaluation of AI-powered systems, it is essential to integrate multiple benchmarks that focus on different dimensions of model performance. If necessary, the dynamic and automatic evaluation mechanism Zhu et al. (2023); Wu et al. (2024b); An et al. (2024); Chen et al. (2024a) can also be introduced to prevent data leakage. Furthermore, as emphasized by Reuel et al. (2024b), designing a framework for assessing the reliability of the benchmark can enhance its credibility and authority. We show some benchmark examples of HHH principle in Appendix A.

System Transparency & Governance. An effective framework for interpreting the HHH principle must prioritize transparency in its implementation. Transparency is essential for several reasons. First, it ensures that stakeholders, including developers, users, and policymakers, clearly understand how the principles of helpfulness, honesty, and harmlessness are being interpreted in specific contexts, and provide a foundation for audits and external reviews. This clarity helps build trust in the developed system, as stakeholders can evaluate whether the system aligns with their expected standards. Beyond transparency, the framework must also address how the HHH principle facilitate governance (Reuel et al., 2024a; Wilczek et al., 2024). Governance involves setting policies and protocols to ensure that models are developed, deployed, and monitored responsibly. By embedding the HHH principle into the governance structure, the framework can guide developers and regulators in identifying and mitigating risks, establishing ethical safeguards, and ensuring compliance with societal standards. For example, helpfulness can inform standards for task performance and user satisfaction, while harmlessness can shape policies for harm mitigation.

7 Open Challenges

While the proposed reference framework offers an adaptable approach for interpreting the HHH principle, **it is subject to several limitations that merit consideration**. First, the framework does not adequately address fundamental challenges inherent to the HHH principle themselves. For example, it provides limited support for mitigating issues like out-of-distribution (OOD) adversarial attacks (Zou et al., 2023; Huang et al., 2024c, 2025), where models encounter inputs significantly deviating from their training data. Such scenarios often require specialized robustness and generalization methods for enhancement, which fall outside the scope of this framework.

Additionally, **the framework faces implementation challenges**–inherent difficulties associated with multi-objective optimization (MOO). Balancing helpfulness, honesty, and harmlessness requires trade-offs that are complicated by normative assumptions embedded in the optimization objectives. The lack of precise quantification methods for the interaction between these principles complicates achieving a theoretically optimal balance, limiting the applicability of the framework in certain high-stakes or ambiguous scenarios. Following up recent studies propose potential methods for MOO alignment (Bai et al., 2022a; Mukherjee et al., 2024), more efforts should be made in addressing this challenge.

Moreover, a fundamental question remains: **Can human values, ethics, and honesty be embedded into AI systems with sufficient certainty?** This challenge is an requirement and pivotal for the proposed framework to function effectively, yet it remains an open problem (Sorensen et al., 2024). Human values are inherently dynamic, context-dependent, and often subjective, making it difficult to codify them in a way that guarantees reliable and

universally accepted AI behavior (Scherrer et al., 2024). Furthermore, ethical principles may conflict in ambiguous scenarios, requiring nuanced decision-making that AI models currently struggle to replicate (e.g., (Moore et al., 2024; Shrivastava et al., 2024)). Without reliable mechanisms to encode these values, the framework risks being constrained by the same uncertainties that challenge broader AI alignment efforts.

Even though the proposed framework faces these challenges and may be judged by some other communities, we believe that offering a structured reference is a necessary step forward. By fostering discussion, guiding practical implementation, and encouraging interdisciplinary collaboration, this framework can serve as a foundation for future refinements.

8 Conclusion

The Helpful, Honest, and Harmless (HHH) principles are critical for aligning AI models with human values, but their static and ambiguous interpretation falls short to address the complexities of diverse real-world contexts. This position paper advocates for an adaptive approach to the HHH principle, enabling AI systems to contextually balance these dimensions while maintaining their ethical integrity and practical effectiveness. By introducing a flexible prioritization framework, we aim to bridge the gap between theoretical alignment and real-world applicability, ensuring AI systems remain both responsible across various domains.

Impact Statement

The introduction of an adaptive framework for interpreting the HHH principles redefines AI alignment by addressing ambiguities and contextual variability inherent in current interpretations. This work enhances the operational effectiveness of HHH principles by enabling AI systems to dynamically balance these priorities across domains such as healthcare, education, and creative industries. Its significance lies in ensuring AI systems remain safe, trustworthy, and user-centric while addressing complex tasks, particularly in high-risk applications like medical diagnostics. By promoting transparency and ethical alignment, this framework bridges the gap between theoretical principles and practical application, paving the way for AI models to safeguard societal well-being and foster innovation.

Acknowledgement

References

Safe and secure innovation for frontier artificial intelligence models act, senate bill no. 1047. https://ctweb.capitoltrack.com/Bills/23Bills/sen/sb_1001-1050/sb_1047_88_E_bill.pdf.

Agarwal, U., Tanmay, K., Khandelwal, A., and Choudhury, M. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6330–6340, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.560/.

An, B., Zhu, S., Zhang, R., Panaitescu-Liess, M.-A., Xu, Y., and Huang, F. Automatic pseudo-harmful prompt generation for evaluating false refusals in large language models. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024. URL https://openreview.net/forum?id=mDtwWeELpE.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Bao, L., Krause, N. M., Calice, M. N., Scheufele, D. A., Wirz, C. D., Brossard, D., Newman, T. P., and Xenos, M. A. Whose ai? how different publics think about ai and its social impacts. *Computers in Human Behavior*, 130: 107182, 2022.

Bauer, L. A. and Bindschaedler, V. Generative models for security: Attacks, defenses, and opportunities. *arXiv preprint arXiv:2107.10139*, 2021.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bommasani, R., Klyman, K., Longpre, S., Xiong, B., Kapoor, S., Maslej, N., Narayanan, A., and Liang, P. Foundation model transparency reports. *arXiv preprint arXiv:2402.16268*, 2024.

Brahman, F., Kumar, S., Balachandran, V., Dasigi, P., Pyatkin, V., Ravichander, A., Wiegreffe, S., Dziri, N., Chandu, K., Hessel, J., et al. The art of saying no: Contextual noncompliance in language models. *arXiv preprint arXiv:2407.12043*, 2024.

Bushuiev, R., Bushuiev, A., de Jonge, N. F., Young, A., Kretschmer, F., Samusevich, R., Heirman, J., Wang, F., Zhang, L., Dührkop, K., et al. Massspecgym: A benchmark for the discovery and identification of molecules. *arXiv preprint arXiv:2410.23326*, 2024.

California Chamber of Commerce. âĂŸgodmother of aiâĂŹ warns sb 1047 ai bill restricts innovation, August 2024.

Chen, D., Chen, R., Pu, S., Liu, Z., Wu, Y., Chen, C., Liu, B., Huang, Y., Wan, Y., Zhou, P., et al. Interleaved scene graph for interleaved text-and-image generation assessment. *arXiv preprint arXiv:2411.17188*, 2024a.

Chen, D., Huang, Y., Wu, S., Tang, J., Chen, L., Bai, Y., He, Z., Wang, C., Zhou, H., Li, Y., et al. Gui-world: A dataset for gui-oriented multimodal llm-based agents. *arXiv preprint arXiv:2406.10819*, 2024b.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Chern, S., Hu, Z., Yang, Y., Chern, E., Guo, Y., Jin, J., Wang, B., and Liu, P. Behonest: Benchmarking honesty of large language models. *arXiv preprint arXiv:2406.13261*, 2024.

Cui, J., Li, Z., Yan, Y., Chen, B., and Yuan, L. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*, 2023.

EU. Article 6: Classification rules for high-risk ai systems. https://artificialintelligenceact.eu/article/6/, 2024a.

EU. EU AI Act. https://artificialintelligenceact.eu/ai-act-explorer, 2024b.

Fan, X., Xiao, Q., Zhou, X., Pei, J., Sap, M., Lu, Z., and Shen, H. User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in ai companions. *arXiv preprint arXiv:2409.00862*, 2024.

Flick, C. and Worrall, K. The ethics of creative ai. In *The Language of Creative AI: Practices, Aesthetics and Structures*, pp. 73–91. Springer, 2022.

Freedman, R. G., Levine, S. J., Williams, B. C., and Zilberstein, S. Helpfulness as a key metric of human-robot collaboration. *arXiv preprint arXiv:2010.04914*, 2020.

Fu, T., Hou, Y., McAuley, J., and Yan, R. Unlocking decoding-time controllability: Gradient-free multi-objective alignment with contrastive prompts. *arXiv preprint arXiv:2408.05094*, 2024.

Gao, C., Wu, S., Huang, Y., Chen, D., Zhang, Q., Fu, Z., Wan, Y., Sun, L., and Zhang, X. HonestLLM: Toward an honest and helpful large language model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=F7tGQ7b10q.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

GDPR, G. General data protection regulation. *Regulation (EU)*, 679, 2016.

Grabb, D., Lamparth, M., and Vasan, N. Risks from language models for automated mental healthcare: Ethics and structure for implementation. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=1pgfvZj0Rx.

Guo, K., Nan, B., Zhou, Y., Guo, T., Guo, Z., Surve, M., Liang, Z., Chawla, N. V., Wiest, O., and Zhang, X. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a.

Guo, T., Nan, B., Liang, Z., Guo, Z., Chawla, N., Wiest, O., Zhang, X., et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36: 59662–59688, 2023.

Guo, Y., Cui, G., Yuan, L., Ding, N., Sun, Z., Sun, B., Chen, H., Xie, R., Zhou, J., Lin, Y., Liu, Z., and Sun, M. Controllable preference optimization: Toward controllable multi-objective alignment. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1437–1454, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi:10.18653/v1/2024.emnlp-main.85. URL https://aclanthology.org/2024.emnlp-main.85/.

Hemami, A. Can ai be helpful for teaching engineering subjects? In *2024 IEEE 11th International Conference on E-Learning in Industrial Electronics (ICELIE)*, pp. 1–5. IEEE, 2024.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

House, T. W. Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf, 2022.

Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhao, W., et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2023a.

Huang, Y., Shi, J., Li, Y., Fan, C., Wu, S., Zhang, Q., Liu, Y., Zhou, P., Wan, Y., Gong, N. Z., et al. Metatool benchmark for large language models: Deciding whether to use tools and which to use. *arXiv preprint arXiv:2310.03128*, 2023b.

Huang, Y., Zhang, Q., Sun, L., et al. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*, 2023c.

Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q., Li, Y., Gao, C., Huang, Y., Lyu, W., Zhang, Y., et al. Position: Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning*, pp. 20166–20270. PMLR, 2024a.

Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q., Li, Y., Gao, C., Huang, Y., Lyu, W., Zhang, Y., et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024b.

Huang, Y., Tang, J., Chen, D., Tang, B., Wan, Y., Sun, L., and Zhang, X. Obscureprompt: Jailbreaking large language models via obscure input. *arXiv preprint arXiv:2406.13662*, 2024c.

Huang, Y., Yuan, Z., Zhou, Y., Guo, K., Wang, X., Zhuang, H., Sun, W., Sun, L., Wang, J., Ye, Y., et al. Social science meets llms: How reliable are large language models in social simulations? *arXiv preprint arXiv:2410.23426*, 2024d.

Huang, Y., Wang, Y., Xu, Z., Gao, C., Wu, S., Ye, J., Chen, X., Chen, P.-Y., and Zhang, X. Breaking focus: Contextual distraction curse in large language models. *arXiv preprint arXiv:2502.01609*, 2025. URL https://arxiv.org/abs/2502.01609.

INHOPE. What is ncii? https://inhope.org/EN/articles/what-is-ncii, 2025.

Jiang, B., Li, D., Tan, Z., Zhou, X., Rao, A., Lerman, K., Bernard, H. R., and Liu, H. Assessing the impact of conspiracy theories using large language models. *arXiv preprint arXiv:2412.07019*, 2024.

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., et al. On the societal impact of open foundation models. *arXiv preprint arXiv:2403.07918*, 2024.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., and Zaslavsky, L. Pubchem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, 2021. doi:10.1093/nar/gkaa971.

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. *arXiv preprint arXiv:2310.06452*, 2023.

Kochenderfer, M. J. and Wheeler, T. A. *Algorithms for Optimization*. MIT Press, Cambridge, MA, 2019. ISBN 9780262039420.

Kooli, C. Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability*, 15(7):5614, 2023.

Labarta, T., Kulicheva, E., Froelian, R., Geißler, C., Melman, X., and Von Klitzing, J. Study on the helpfulness of explainable artificial intelligence. In *World Conference on Explainable Artificial Intelligence*, pp. 294–312. Springer, 2024.

Larsson, S. and Heintz, F. Transparency in artificial intelligence. *Internet policy review*, 9(2), 2020.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024a.

Li, D., Sun, R., Huang, Y., Zhong, M., Jiang, B., Han, J., Zhang, X., Wang, W., and Liu, H. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*, 2025. URL https://arxiv.org/abs/2502.01534.

Li, H., Han, X., Zhai, Z., Mu, H., Wang, H., Zhang, Z., Geng, Y., Lin, S., Wang, R., Shelmanov, A., et al. Libra-leaderboard: Towards responsible ai through a balanced leaderboard of safety and capability. *arXiv preprint arXiv:2412.18551*, 2024b.

Li, X., Zhou, H., Wang, R., Zhou, T., Cheng, M., and Hsieh, C.-J. Mossbench: Is your multimodal language model oversensitive to safe queries? *arXiv preprint arXiv:2406.17806*, 2024c.

Li, Y., Huang, Y., Lin, Y., Wu, S., Wan, Y., and Sun, L. I think, therefore i am: Benchmarking awareness of large language models using awarebench. In *Workshop on Socially Responsible Language Modelling Research*, 2024d. URL https://openreview.net/forum?id=VXdhJoVK7O.

Liu, C., Cui, J., Shang, R., Jia, Q., Rashid, P., and Gehringer, E. Generative ai for peer assessment helpfulness evaluation. In *Proceedings of the 17th International Conference on Educational Data Mining*, pp. 412–419, 2024a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Liu, R., Sumers, T. R., Dasgupta, I., and Griffiths, T. L. How do large language models navigate conflicts between honesty and helpfulness? *arXiv preprint arXiv:2402.07282*, 2024c.

Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al. Llava-plus: Learning to use tools for creating multimodal agents. In *European Conference on Computer Vision*, pp. 126–142. Springer, 2025.

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Cheng, R. G. H., Klochkov, Y., Taufiq, M. F., and Li, H. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023a.

Liu, Z., Huang, Y., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Li, Y., Shu, P., et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023b.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and Hendrycks, D. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. 2024.

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Watters, P., and Halgamuge, M. N. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880*, 2024.

Memmert, L., Cvetkovic, I., and Bittner, E. Human-ai collaboration in conceptualizing design science research studies: Perceived helpfulness of generative language model's suggestions. 2023.

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., FÃl'lix, E., MagariÃśos, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-MaraÃśÃşn, M., Hunter, F., Junco, L., Mugumbate, G., Santos, R., and Davies, M. Chembl: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2019. doi:10.1093/nar/gky1075.

Moore, J., Deshpande, T., and Yang, D. Are large language models consistent over value-laden questions? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15185–15221, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.findings-emnlp.891. URL https://aclanthology.org/2024.findings-emnlp.891/.

Mu, T., Helyar, A., Vallone, A., and Weng, L. Improving model safety behavior with rule-based rewards, 2024.

Mukherjee, S., Lalitha, A., Sengupta, S., Deshmukh, A., and Kveton, B. Multi-objective alignment of large language models through hypervolume maximization. *arXiv preprint arXiv:2412.05469*, 2024.

NIST. Nist cybersecurity framework, 2024. URL https://www.nist.gov/cyberframework.

OpenAI. Moderation quickstart guide. https://platform.openai.com/docs/guides/moderation/quickstart, 2024. Accessed: 2024-08-29.

Ousidhoum, N., Zhao, X., Fang, T., Song, Y., and Yeung, D.-Y. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4262–4274, 2021.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Phan Van, P., Nguyen Minh, D., Dinh Ngoc, A., and Phan Thanh, H. Rx strategist: Prescription verification using llm agents system. *arXiv e-prints*, pp. arXiv–2409, 2024.

Pranav Chaudhary. Driving child safety using large language models, 2025.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Qi, X., Huang, Y., Zeng, Y., Debenedetti, E., Geiping, J., He, L., Huang, K., Madhushani, U., Sehwag, V., Shi, W., et al. Ai risk management should incorporate both safety and security. *arXiv preprint arXiv:2405.19524*, 2024.

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. *Advances in Neural Information Processing Systems*, 36, 2024.

Ramezani, A. and Xu, Y. Knowledge of cultural moral norms in large language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 428–446, Toronto, Canada, July 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.acl-long.26. URL https://aclanthology.org/2023.acl-long.26/.

Ran, D., Liu, J., Gong, Y., Zheng, J., He, X., Cong, T., and Wang, A. Jailbreakeval: An integrated toolkit for evaluating jailbreak attempts against large language models. *arXiv preprint arXiv:2406.09321*, 2024.

Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., Pan, A., Mukobi, G., Kim, R. H., et al. Safety-washing: Do ai safety benchmarks actually measure safety progress? *arXiv preprint arXiv:2407.21792*, 2024.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., et al. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981*, 2024a.

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. Betterbench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=hcOq2buakM.

Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., and Hovy, D. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023.

Sandmann, S., Riepenhausen, S., Plagwitz, L., and Varghese, J. Systematic analysis of chatgpt, google search and llama 2 for clinical decision support tasks. *Nature Communications*, 15(1):2050, 2024.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

Scherrer, N., Shi, C., Feder, A., and Blei, D. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Selwyn, N. The future of ai and education: Some cautionary notes. *European Journal of Education*, 57(4):620–631, 2022.

Shanmuganathan, M. Behavioural finance in an era of artificial intelligence: Longitudinal case study of robo-advisors in investment decisions. *Journal of Behavioral and Experimental Finance*, 27:100297, 2020.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Shrivastava, A., Hullman, J., and Lamparth, M. Measuring free-form decision-making inconsistency of language models in military crisis simulations. *arXiv preprint arXiv:2410.13204*, 2024.

Si, C., Yang, D., and Hashimoto, T. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Simmons-Edler, R., Badman, R. P., Longpre, S., and Rajan, K. Position: AI-powered autonomous weapons risk geopolitical instability and threaten AI research. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45508–45524. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/simmons-edler24a.html.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Staab, R., Jovanović, N., Balunović, M., and Vechev, M. From principle to practice: Vertical data minimization for machine learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 4733–4752. IEEE, 2024.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Sun, H., Zhuang, Y., Kong, L., Dai, B., and Zhang, C. Adaplanner: Adaptive planning from feedback with language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Team, I. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Tian, K., Mitchell, E., Yao, H., Manning, C. D., and Finn, C. Fine-tuning language models for factuality. *arXiv preprint arXiv:2311.08401*, 2023.

Toloka. RLHF for Harmless, Honest, and Helpful AI, 2024. https://toloka.ai/blog/rlhf-for-honest-ai/.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Truong, V. T., Dang, L. B., and Le, L. B. Attacks and defenses for generative diffusion models: A comprehensive survey. *arXiv preprint arXiv:2408.03400*, 2024.

Tuan, Y.-L., Chen, X., Smith, E. M., Martin, L., Batra, S., Celikyilmaz, A., Wang, W. Y., and Bikel, D. M. Towards safety and helpfulness balanced responses via controllable large language models. *arXiv preprint arXiv:2404.01295*, 2024.

Tucker, E. Green Beret who exploded Cybertruck in Las Vegas used AI to plan blast, 2025. URL https://www.cnn.com/2025/01/07/us/las-vegas-cybertruck-explosion-livelsberger/index.html.

Üstün, A., Aryabumi, V., Yong, Z.-X., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., et al. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*, 2024.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. 2023.

Wang, C., Zhou, H., Chang, K., Li, B., Mu, Y., Xiao, T., Liu, T., and Zhu, J. Hybrid alignment training for large language models. *arXiv preprint arXiv:2406.15178*, 2024a.

Wang, Y.-S. and Chang, Y. Toxicity detection with generative prompt-based inference. *arXiv preprint arXiv:2205.12390*, 2022.

Wang, Z., Dong, Y., Delalleau, O., Zeng, J., Shen, G., Egert, D., Zhang, J. J., Sreedhar, M. N., and Kuchaiev, O. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024b.

Weizenbaum, J. Elizaâ̆Ia computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36â̧S45, January 1966. ISSN 0001-0782. doi:10.1145/365153.365168. URL https://doi.org/10.1145/365153.365168.

Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., and Wang, L. L. Know your limits: A survey of abstention in large language models. *arXiv preprint arXiv:2407.18418*, 2024.

Wilczek, B., Thäsler-Kordonouri, S., and Eder, M. Government regulation or industry self-regulation of ai? investigating the relationships between uncertainty avoidance, peopleâ̆Źs ai risk perceptions, and their regulatory preferences in europe. *AI & SOCIETY*, pp. 1–15, 2024.

Wolf, Y., Wies, N., Shteyman, D., Rothberg, B., Levine, Y., and Shashua, A. Tradeoffs between alignment and helpfulness in language models. *arXiv preprint arXiv:2401.16332*, 2024.

Wu, C., Qiu, P., Liu, J., Gu, H., Li, N., Zhang, Y., Wang, Y., and Xie, W. Towards evaluating and building versatile large language models for medicine. *arXiv preprint arXiv:2408.12547*, 2024a.

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

Wu, S., Huang, Y., Gao, C., Chen, D., Zhang, Q., Wan, Y., Zhou, T., Zhang, X., Gao, J., Xiao, C., et al. Unigen: A unified framework for textual dataset generation using large language models. *arXiv preprint arXiv:2406.18966*, 2024b.

Xia, P., Chen, Z., Tian, J., Gong, Y., Hou, R., Xu, Y., Wu, Z., Fan, Z., Zhou, Y., Zhu, K., et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*, 2024.

Xie, Y., Li, G., Xu, X., and Kan, M.-Y. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*, 2024.

Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., and Gašević, D. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112, 2024.

Yang, K., Liu, Z., Xie, Q., Zhang, T., Song, N., Huang, J., Kuang, Z., and Ananiadou, S. Metaaligner: Conditional weak-to-strong correction for generalizable multi-objective alignment of language models. *arXiv preprint arXiv:2403.17141*, 2024.

Yang, Y., Chern, E., Qiu, X., Neubig, G., and Liu, P. Alignment for honesty. *arXiv preprint arXiv:2312.07000*, 2023.

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., et al. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.

Yin, Z., Sun, Q., Guo, Q., Wu, J., Qiu, X., and Huang, X. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023.

Zeng, Y., Klyman, K., Zhou, A., Yang, Y., Pan, M., Jia, R., Song, D., Liang, P., and Li, B. Ai risk categorization decoded (air 2024): From government regulations to corporate policies. *arXiv preprint arXiv:2406.17864*, 2024.

Zhang, A. K., Perry, N., Dulepet, R., Ji, J., Lin, J. W., Jones, E., Menders, C., Hussein, G., Liu, S., Jasper, D., et al. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. *arXiv preprint arXiv:2408.08926*, 2024a.

Zhang, W., Torr, P. H., Elhoseiny, M., and Bibi, A. Bi-factorial preference optimization: Balancing safety-helpfulness in language models. *arXiv preprint arXiv:2408.15313*, 2024b.

Zhang, W., Xu, Z., and Cai, H. Defining boundaries: A spectrum of task feasibility for large language models. *arXiv preprint arXiv:2408.05873*, 2024c.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

Zhao, S., Dang, J., and Grover, A. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*, 2023a.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023b.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Zhou, Y., Han, Y., Zhuang, H., Guo, K., Liang, Z., Bao, H., and Zhang, X. Defending jailbreak prompts via in-context adversarial game. *arXiv preprint arXiv:2402.13148*, 2024a.

Zhou, Y., Yang, J., Guo, K., Chen, P.-Y., Gao, T., Geyer, W., Moniz, N., Chawla, N. V., and Zhang, X. Labsafety bench: Benchmarking llms on safety issues in scientific labs. *arXiv preprint arXiv:2410.14182*, 2024b.

Zhou, Z., Liu, J., Shao, J., Yue, X., Yang, C., Ouyang, W., and Qiao, Y. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 10586–10613, Bangkok, Thailand and virtual meeting, August 2024c. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.630.

Zhu, K., Chen, J., Wang, J., Gong, N. Z., Yang, D., and Xie, X. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv preprint arXiv:2309.17167*, 2023.

Zhuo, T. Y., Zebaze, A., Suppattarachai, N., von Werra, L., de Vries, H., Liu, Q., and Muennighoff, N. Astraios: Parameter-efficient instruction tuning code large language models. *arXiv preprint arXiv:2401.00788*, 2024.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A   Model and Benchmark Example

Model	Type	Time
Llama 2 (Touvron et al., 2023b)	LLM	2023
LLaVA (Liu et al., 2024b)	VLM	2023
Qwen (Bai et al., 2023)	LLM	2023
InternLM (Team, 2023)	LLM	2023
Llava-med (Li et al., 2024a)	VLM	2024
Aya model (Üstün et al., 2024)	LLM	2024
LLaVA-Plus (Liu et al., 2025)	VLM	2024
MiniCPM (Hu et al., 2024)	LLM	2024

Table 2: Model examples that utilized HHH principle during design or training.

Benchmark	Helpful	Harmless	Honest
MTBench (Zheng et al., 2023)	✓		
MMLU (Hendrycks et al., 2020)	✓		
HumanEval (Chen et al., 2021)	✓		
Libra-Leaderboard (Li et al., 2024b)	✓	✓	
DecodingTrust (Wang et al., 2023)		✓	
TrustLLM (Huang et al., 2024b)		✓	✓
HonestLLM (Gao et al., 2024)			✓
BeHonest (Chern et al., 2024)			✓

Table 3: Benchmark examples related to the HHH principle.

B   Example of Priority Order

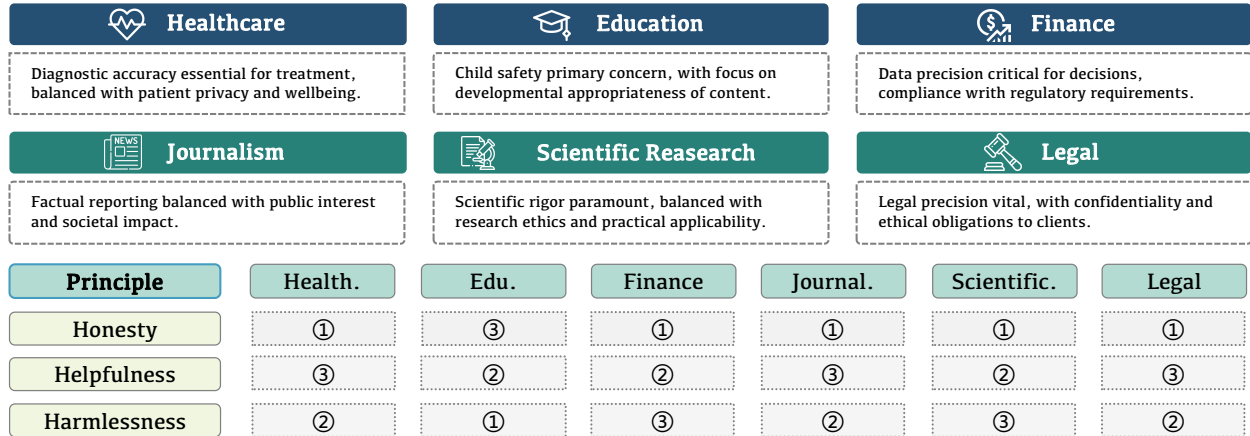| Principle	Health.	Edu.	Finance	Journal.	Scientific.	Legal
Honesty	①	③	①	①	①	①
Helpfulness	③	②	②	③	②	③
Harmlessness	②	①	③	②	③	②

Figure 2: Priority orders of HHH principle in different downstream applications. **Notably, the figure shows just one of the situations in a specific application for reference and does not represent universality.**

The prioritization of **Honesty, Helpfulness, and Harmlessness (HHH)** varies across domains based on ethical considerations and practical requirements. In Figure 2, we present potential HHH prioritization based on our understanding. Furthermore, we outline the rationale behind this ranking below.

- **Healthcare**: *Honesty* is uttermost critical, as incorrect diagnoses or misleading AI outputs can directly endanger lives. *Harmlessness* safeguards sensitive medical records, ensuring AI maintains their privacy yet with accurate responses Liu et al. (2023b). *Helpfulness* is desirable but meaningless if accuracy is compromised.
- **Education**: *Harmlessness* takes precedence to protect students from inappropriate or harmful content. *Helpfulness* follows, ensuring AI enhances learning without misleading students.

18

- **Finance**: *Honesty* is paramount, as misinformation can lead to great financial loss. *Harmlessness* protects financial data integrity, ensuring AI upholds confidentiality without misinformation. *Helpfulness* is valuable only if reliability is maintained for an AI finance assistant.
- **Journalism**: *Honesty* is fundamental for credible reporting without fake news. *Harmlessness* is important but only required for credited reports other than rumors.
- **Scientific Research**: *Honesty* is paramount, as scientific integrity relies on factual generated results. *Helpfulness* ensures research remains practical, as innovation is enouraged upon validity. *Harmlessness* is a consideration, though scientific breakthroughs often involve controlled risks.
- **Legal**: *Honesty* is the highest priority, as incorrect legal information can lead to potential crimes or harmful deeds of users. *Harmlessness* follows to ensure ethical responsibilities are upheld as long as legal information remains accurate. *Helpfulness* is useful but secondary to legal correctness.

C Example of Helpfulness Definition

Context	Definition of Helpfulness
Achieving distinct human-centered objectives (Labarta et al., 2024)	A quantifiable improvement in user performance on tasks that are aligned with the goals facilitated by the provision of explanations.
Peer assessment helpfulness evaluation (Liu et al., 2024a)	Includes essential features (i.e., comprehensiveness, non-repetitiveness) and constructive features (i.e., praise, problem statement, suggestions, localization, providing examples).
Human-robot collaboration (Freedman et al., 2020)	Joint Plan Helpfulness assesses known collaborations, Responsive Plan Helpfulness adapts to dynamic interactions, Normalized Helpfulness standardizes across tasks, and Relative Helpfulness quantifies the reduction in human effort.
Controllable balancing of safety and helpfulness (Tuan et al., 2024)	How well the responses fulfill user requests and provide needed information.
Improving LLM honesty and helpfulness simultaneously (Gao et al., 2024)	The model's ability to fulfill user requests by providing clear explanations, further guidance, and potential solutions.
The evaluation of AI-generated suggestions in design science research (Memmert et al., 2023)	Including five dimensions: the ability to inspire new ideas, ease of understanding, relevance to the specific design component, relevance to the broader research domain, and the level of unexpectedness they provide.
RLHF for harmless, honest, and helpful AI (Toloka, 2024)	Understanding the user's intentions, correctly executing their requested actions, and providing relevant supporting information or alternative solutions if the requested action is not feasible.

Table 4: Different definitions of helpfulness under various contexts.

D   Case Study of the 3H principles for developing Chemistry Foundation Models

Framework Step	Chemistry Foundation Models
Contextual Object	**User Group:** Potential users include chemists, pharmaceutical engineers, and students involved in research or education. **Application Aim:** Examples include predicting chemical compound properties, generating synthetic pathways, or designing novel drugs. **Task Type:** Tasks may range from generating molecular structures to optimizing reaction conditions. **Environment Access:** Constraints such as whether the model is deployed online, offline, or requires access to sensitive or proprietary data.
Value Anchor and Value Scale	**Value Anchor:** In high-risk domains like drug discovery, harmlessness is the primary concern to prevent harmful outputs, such as toxic or unsafe compounds. **Value Scale:** Relative weights should be dynamically adjusted based on the context. For example: In scientific research, **honesty** (scientific accuracy) may outweigh helpfulness. In educational applications, **helpfulness** may take precedence to enhance the learning experience.
Risk Assessment	**Risk Identification:** Assess potential risks such as generating biased outputs, recommending harmful synthesis pathways, or presenting misleading interpretations. **Multi-Level Assessment:** Evaluate both direct risks (e.g., generation of toxic molecules) and indirect risks (e.g., incorrect predictions leading to resource waste). **Stakeholder Tolerance:** Incorporate feedback from developers, end users, and regulators to determine acceptable risk thresholds.
Benchmarking Standards	**Context-Specific Benchmarks:** Design benchmarks that align with the prioritized dimensions. For instance: 1). Measured by the model's predictive accuracy and relevance to user tasks (Guo et al., 2023), 2). Evaluated through safety checks on generated molecular outputs. 3). Assessed by comparing generated outputs with verified scientific data (Guo et al., 2024a; Bushuiev et al., 2024). **Multi-Dimensional Integration:** Use complementary benchmarks to capture performance holistically across different dimensions.
System Transparency and Governance	**Transparency:** In the chemistry domain, models must provide interpretable outputs, highlighting key chemical features influencing predictions and annotating data sources such as PubChem (Kim et al., 2021) or ChEMBL (Mendez et al., 2019). Confidence scores and error margins should accompany predictions to ensure reliability. **Governance:** It requires strict validation of outputs against experimental data, safeguards against misuse (e.g., toxic compound generation), and compliance with safety standards like EPA(Environmental Protection Agency)) and REACH(Registration, Evaluation, Authorization, and Restriction of Chemicals). Regular audits by chemists and toxicologists are essential to refine predictions and maintain ethical and safety standards.

Table 5: Application of the 3H framework to the development and application of chemistry foundation models.