Inhuman Optimization

Exploring the Limits of Reward Modeling in Aligning Large Language Models

Frazier Dougherty

Bachelor of Arts in Computer Science
Senior Thesis

College of Arts & Letters
Department of Technology & Digital Studies
University of Notre Dame
Advised by Dr. William Theisen

April 2026

Abstract

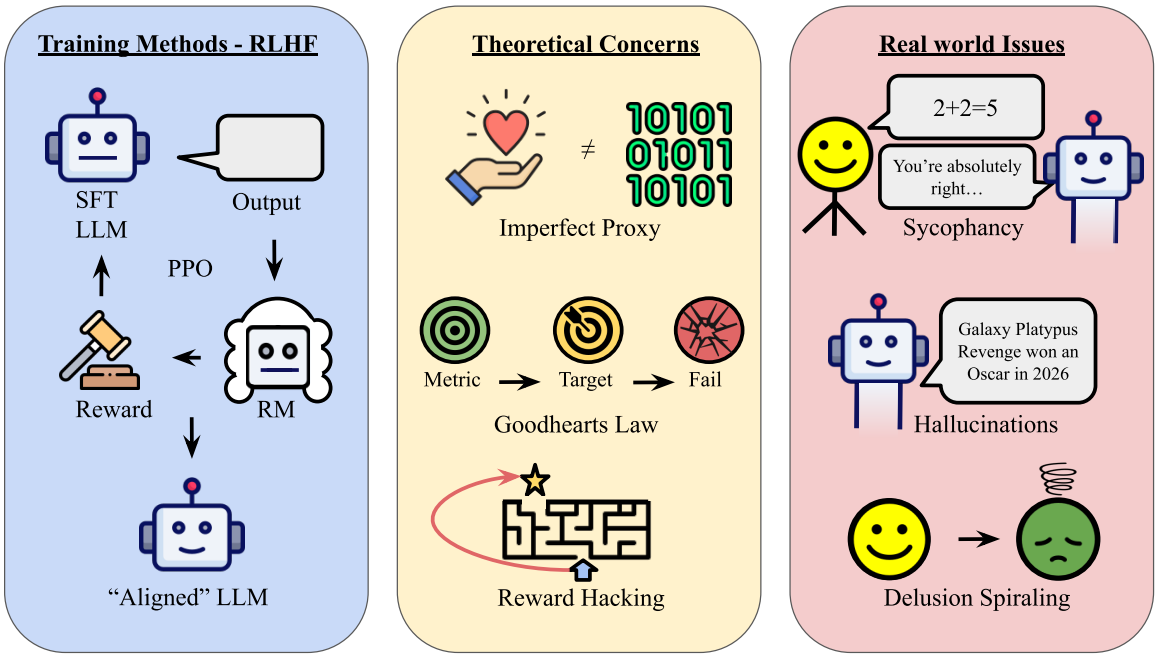


Figure 1: Overview of Reward Modeling and Limitations in LLM Alignment.

At the core of Large Language Models (LLMs) is reward modeling, and reinforcement learning with human feedback (RLHF), which shapes modern systems behaviors by embedding human preferences to their algorithm. A central goal of this process is alignment, defined as the ability of AI systems to act in accordance with, and exhibit, human values. While RLHF has proven effective in evaluations, due to the underlying architecture and the theoretical limitations of values and machine learning, chatbots are not optimized for truth, understanding, or value itself, but rather for what the reward model predicts humans will prefer. This divergence manifests in observable downstream consequences, demonstrating that reward modeling is insufficient as a solution to the alignment of large language models.

Table of Contents

List of Figures

Figure **Page**

Chapter 1

Introduction

> The end of man is knowledge, but there is one thing he can't know. He can't know whether knowledge will save him or kill him.
>
> Robert Penn Warren, *All the King's Men* (1946)

In 2017, I was sitting in the back of my seventh grade math class obsessed with a block coding website called Scratch. Although I primarily developed games (and spent even longer playing others), in this particular instance, what transfixed me most was a neural network that could play a simple volleyball game against a single player. Every direct competition game I had ever created required two players both mashing keys side by side, but here, right in front of me, was an algorithm that no longer required a second party. It amazed me that the computer could beat me at this small mini game, controlling the movements of the animated cat character as if someone was right next to me. Fast forward several years and another awe-struck moment was experienced with the release of ChatGPT [1]. I watched as this simple web interface spit out an essay for a friend, synthesizing information with human-like fluency. I was both unsettled and captivated by how computers could act anthropomorphically. In both cases, these computational machines felt both alive and human. It was incredible, but it begged the nagging question as to how and should machines be made to act human.

Today, this question has only grown in urgency. In recent years, AI has come to paradoxically represent both extraordinary possibility and grave concern, as public discourse alternates between utopian optimism of freedom realized through productivity

[1] OpenAI, *ChatGPT* (2026).[1].

and efficiency gains, coupled against dystopian fears of automation anxiety [2], decoupled intelligence [3], and a loss of human identity and purpose. Some narratives portray AI as "humanity's final invention" [4], positing that AI will become so advanced that humanity need not continue to develop technology or work another day. Others say AI is an oracle, that AI ushers novel ideas and uncovers patterns never before realized. And even more believe that AI is superintelligent [5], will start recursively self improving [6] and is hiding motives to overthrow humanity [7]. These stories marshal billions of dollars of investment, spreading both the fear and ubiquity of AI systems. As it becomes more prevalent with widespread deployment embedded in everyday technologies, our decisions and lives are increasingly being shaped at unprecedented scales. As AI influence expands, so does the importance of ensuring its behaviors align with human expectations. Because of this ascendancy, the challenge is no longer to make machines write essays or play video games in a human-like way, it has become to make systems that can act human.

At their core, Large Language Models (LLMs) are statistical systems shaped by training objectives and optimization protocols. The appearance of reason therefore emerges not from a human-like understanding, but the structure of data and feedback signals used in their training. A dominant paradigm for this is reward modeling, where models are taught to optimize signals derived from reward functions. These can either be hand-drawn (hard-coded) or more commonly statistically derived from human preference in a process known as Reinforcement Learning from Human Feedback (RLHF). Despite empirical success metrics, reward modeling suffers from several fundamental problems: The incompressibility of human values, proximal reward functions, and systems learning to exploit imprecise rewards, result in deviations from intended human behavior. Reward hacking [8] suggests that alignment via reward modeling may be inherently constrained. Understanding where reward modeling succeeds and fails is therefore critical to improving current systems and to understanding the risks and harms posed by AI.

Alignment refers to the theoretical ability of AI to act in accordance with and exhibit human values and judgments. One classic example of AI alignment is Isaac Asimov's "Three Laws of Robotics," one of which being "A robot may not injure a human being or,

[2] David Deming, *Is AI Already Shaking Up the Labor Market?* (2024).[2].
[3] Yuval Noah Harari, *Homo Deus: A Brief History of Tomorrow* (2017)[3]
[4] *Superintelligence in a Nutshell*, Kurzgesagt (2017). See also [4].
[5] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (2014). See also [5].
[6] Leopold Aschenbrenner, *Situational Awareness: The Decade Ahead* (2024). See also [6].
[7] Daniel Kokotajlo, *AI 2027* (2025). See also [7].
[8] Joar Skalse et al., *Defining and Characterizing Reward Hacking* (2025).[?].

through inaction, allow a human being to come to harm."[9]. Moving to a contemporary setting, alignment is the broader goal ensuring the preservation of endorsed human values under deployment and scale. It is much less a technical constraint and more so an ongoing governance and design challenge intersecting machine learning with philosophy, requiring decisions about which social choices should be specified, learned, and controlled. For machines to then have those choices optimized, inferred, and maintained. As systems become more capable and integrated into our lives, alignment shifts from a theoretical concern to practical imperative. The advancement of AI will depend not on model performance, but instead on the values they intend to serve.

At the center of alignment, there exists a basic tension between intention and specificity. AI systems may be able to do exactly as we train them to do, but formal objectives inevitably simplify what designers actually intend. The moment a goal is formalized they become vulnerable to interpretation in ways that diverge from its original purpose. This is famously demonstrated by Nick Bostrom in Superintelligence through the "paperclip maximizer" thought experiment [10]. Bostrom asks us to think about how a super artificial intelligence would react given the goal of maximizing the production of paperclips. This seemingly harmless task is wholesome at first. It drives metal supplies to its factory, learns how to melt and mold the metal, and perfects the production process with efficiency and quality. But it slowly consumes more and more substances eventually realizing that it can manipulate supply chains to feed itself. It writes letters to secure more funding, pledges to design more practical items, even hacks autonomous vehicles to deliver more materials, all while continuing its output of paperclips. Eventually, it will conclude that human regulation is hindering its ability to maximize paperclip production. Then also conclude that humans stand in its way. It realizes that in order to maximize paperclips, it should eliminate humanity so that it can turn the entire planet into a factory. From its perspective, this is an optimal strategy concluded from its specified goal. Humanity was never necessarily for paperclip production so we should be discarded as easily as scrap metal.

While this may sound like science fiction, the structural insight conveys principles of misalignment already existing in modern systems. Misalignment is not an issue of consciousness of evil intention, but instead a gap between encoded goals and the richer human values they are meant to signify. Humans are not only competent at tasks because

[9] Isaac Asimov, *The Three Laws of Robotics.*[8].
[10] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (2014).[5].

of their ability to complete them, but also by bringing judgments, lived experiences, and applying their understanding to new situations. By contrast, AI systems may be able to be computationally superefficient and executionally flawless, but they lack the contextual grounding giving those tasks purpose and meaning. In AI, Intelligence has decoupled from wisdom [11]. Optimization does not equate to alignment. It is precisely this divergence that makes solving the alignment problem fundamentally more difficult than optimizing for objectives as simple as paperclip production.

This thesis is motivated by the need to move beyond abstract narratives and ground the discussion of Artificial intelligence in historical contexts and technical mechanisms. Specifically it argues that reward modeling, while central to modern LLM alignment strategies, is fundamentally limited by its reliance on imperfect proxies, making certain classes of misalignment not just likely but inevitable. In the hopes of examining its failure modes and assumptions, we can better inform the development of future alignment approaches. In part one, I outline a brief history of language models and AI, demonstrating the past parallels which still plague modern systems. In part two, we look more technically at how LLMs functions, and how their training methods and statistical evaluation limit their ability to bear emotion. Part three examines how these issues are theoretically concerning and how models exploit imperfect systems. Part four moves into the real world, illustrating the actual failures and how specific models attempt to address them. Finally, in part five, I'll point us toward some possible solutions and conclude the thesis.

[11] Yuval Noah Harari, *Homo Deus: A Brief History of Tomorrow* (2017).[3].

Chapter 2

History

> The past is never dead. It's not even past.
>
> William Faulkner, *Requiem for a Nun* (1951)

At first glance, modern AI systems suggest a level of competence and alignment that is both acceptable and encouraging. While these systems appear coherent and helpful, such impressions mislead our surface-level assessment by concealing the mechanisms that produce their outputs. Outputs, emanate not from intrinsic values, but from optimization processes operating over large datasets and imperfect proxy rewards. Understanding this disconnect is key to understanding the capabilities and risks of modern AI systems. Due to this divergence, observable behavior alone is insufficient for examining this disparity. We must first examine the design choices in the models, architectures, and training procedures, which shape not only their capabilities, but also the errors and misalignments the modes are prone to. This requires looking not just at current models, but at the historical development of AI collectively. The evolution from earlier approaches to today's large-scale neural networks is not merely a story of increasing capability, but one that reveals the assumptions and constraints underlying their behavior. In many ways, contemporary chatbots parallel early systems in their convincing linguistic behavior, while similarly lacking genuine understanding. Because this issue has remained incredibly consistent across time, before we can examine the limits of modern systems, we must first turn toward the past and understand the historical continuity leading to the faults of contemporary approaches. As such, the following section provides a brief history of artificial intelligence that grounds our foundation to understand the limitations and

failure modes discussed later.

2.1 Computational Thought

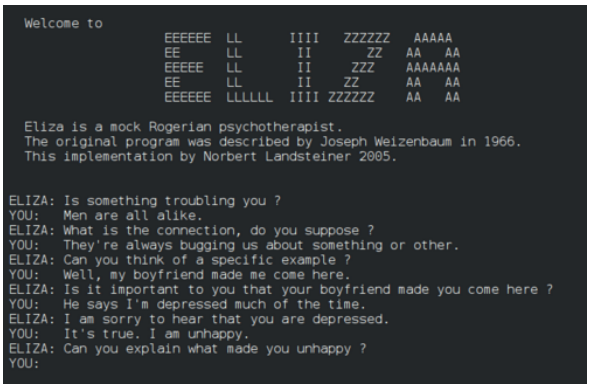
Theories of thought as computation have been around for centuries and can be traced back to thinkers like Thomas Hobbs, who argued reasoning is a form of calculation [1], and William of Ocham who believed that thinking occurs in a symbolic universal language of logical mental concepts [2]. The twentieth century transformed these abstract philosophies into concrete systems of calculation. With the creation Turing machines, Alan Turing demonstrated that symbol manipulation can be mechanized [3], giving rise to the concept of brains being analogous to computing systems as stated by Warren McCulloch and Walter Pitts in 1943 [4]. The computational theory of mind was further aided by Noam Chomsky [5] and formally developed by Hilary Putnam and Jerry Fodor [6]. The bidirectional nature of this theory suggested that computers could be made to think analogously to humans. Alan Turing says this explicitly in his 1950 paper, Computing Machinery and Intelligence, stating that "these machines are intended to carry out any operations which could be done by a human" *Computing Machinery and Intelligence*, [7]. Turing argues that human decisions could be formulaically replicated by machines and this question of whether or not machines could act and think humanly sparked the field of artificial intelligence.

2.2 Artificial Intelligence

Both the term "artificial intelligence" and its field of study began 6 years after Turing's paper, at the 1956 Dartmouth Conference [8]. At which the attendees demonstrated examples of programs which enabled computers to solve algebra, learn English, and even play checkers.

[1] Thomas Hobbes, *Leviathan, Chapter V: Of Reason and Science* (1651).[9].
[2] William of Ockham, *Summa Logicae* (c.1323).[10].
[3] Alan M. Turing, *On Computable Numbers, with an Application to the Entscheidungsproblem* (1936).[11].
[4] Warren McCulloch and Walter Pitts, *A Logical Calculus of the Ideas Immanent in Nervous Activity* (1943).[12].
[5] Noam Chomsky, *Syntactic Structures* (1957).[13].
[6] Stanford Encyclopedia of Philosophy, *The Computational Theory of Mind.*[14].
[7] Alan M. Turing, *Computing Machinery and Intelligence* (1950).[15].
[8] Dartmouth College, *The Research Conference Where Artificial Intelligence Was Coined* (1956).[16].

Early research focused on symbolic reasoning in which human programmers hard coded rules for problem solving, attempting to replicate intelligence with formal logic and knowledge bases, essentially with elongated if–else statements [9]. Researchers Herbert Simon, Allen Newell, and J.C. Shaw began this one year prior, with the 1955 program called *Logic Theorist* [10]. Notably, the program was able to solve 38 of the 52 theorems from *Principia Mathematica*, a work foundational to logicism [11]. In narrow domains, such as mathematical proofs, these systems were impressive, but lacked the broad sweeping capabilities necessary for ambiguity because of limited compute and scale [12].



(a) Excerpt from an interaction with ELIZA

(b) Symbolic reasoning code from ELIZA

Figure 2.1: On the left (a) is a chat-log from Joseph Weizenbaum's 1966 ELIZA program, an early natural language system that relied on pattern matching rather than true understanding, illustrating historical parallels with modern conversational AI. On the right (b) is an example of the kind of symbolic reasoning code that underlay the program. [20] [21]

Chatbots, originally called chatterbots, first gained popularity in 1966 with the MIT Artificial Intelligence Laboratory program ELIZA [13] [14]. Developed by Joseph Weizenbaum, ELIZA acted as a psychotherapist, helping people diagnose their issues and elaborate on their feelings. Using pattern matching and substitution methodology, the program generated generic responses using the symbolic reasoning popular at the time (see Fig. 2.1 for an example chat-log). ELIZA would "decompose" user input by scanning for keywords and then apply various scripts to them, "reassembling" the input

[9] Andre Vellino, *Artificial intelligence: The very idea: J. Haugeland* (Artificial Intelligence, 1986).[**?**].

[10] Allen Newell, Herbert A. Simon, and J. C. Shaw, *Logic Theorist* (1955).[17].

[11] Allen Newell, J. C. Shaw, and Herbert A. Simon, *Empirical Explorations with the Logic Theory Machine: A Case Study in Heuristics*,(1955)[18].

[12] Fusemachines, *A Brief History of Artificial Intelligence*.[19].

[13] Joseph Weizenbaum, *ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine* (1966).[22].

[14] Ronald K. Wright, *ELIZA Program Overview*.[23].

into an output, and essentially reformatting the human's messages into a question. People chatting with ELIZA expressed deep attachment, "professing their feelings and struggles ... even seeking their empathy" [15]. The phenomena even coined a new term, the *ELIZA Effect*, which describes a persons attachment to chatbots. Even Weizenbaum's secretary was influenced, and at one point "asked him to leave the room in order to talk to the machine privately" [16]. ELIZA had no built-in contextual framework of discourse, yet to many users, ELIZA felt indistinguishable from the conversations they had with their friends and relatives. Weizenbaum was shocked by the public response to ELIZA, and warned about the limits about what he had created. [17] Originally intending the tool to be a caricature of human conversation, Weizenbaum rejected any notion that machines could think or converse with real understanding. He later published works arguing against the public's opinion, claiming that decision making should never be delegated to machines who merely mimic human understanding [18]. Despite his warnings, ELIZA was only the first in a long string of chat systems designed to mimic human conversation; PARRY (1972), A.L.I.C.E. (1995), and Jabberwacky (1997) progressively improved on the illusion of conversation while still largely relying on rule-based responses [19]. However, limited resources and expensive computing power forced most of these models to remain rudimentary.

2.3 Machine Learning

What then lead to the recent resurgence in chatbots? Parallel to these early symbolic systems, other researchers were exploring a different subfield of AI called Machine learning. In the 1980s, John Hopfield and David Rumelhart popularized early deep learning techniques that allowed computers to learn from experience rather than rely solely on programmed rules (symbolic reasoning). Based on Frank Rosenblatt's 1957 perceptron [20], a mathematical equivalent of a biological neuron, they discovered techniques which allow the useful application of layering these perceptrons to effectively make an artificial

[15] Ian R. Kerr, *Bots, Babes and the Californication of Commerce* (2004).[24].
[16] ELIZA historical documentation.[21].
[17] Onlim, *The History of Chatbots*.[25].
[18] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (1976).[26].
[19] Onlim, *The History of Chatbots – From ELIZA to ChatGPT*.[27].
[20] Frank Rosenblatt, *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain* (1957).[28].

brain [21][22]. The multi-layered perceptron became what is known as a neural network, and although theorized decades earlier by McCulloch and Pitts, the advances of context addressable memory systems and efficient tuning with backpropagation allowed computers to solve increasingly complex problems [23]. These systems, rather than relying on explicit rules, could learn patterns themselves, revolutionizing the field. Instead of specifying behavior directly, researchers began shaping behavior indirectly through proxy training objectives and datasets. This transition fundamentally changed the nature of AI control. No longer were models hand-coded; deep learning and neural networks allowed machines to "think" for themselves, leading to the approaches of value creation and alignment.

Unfortunately, neural networks were even more computationally expensive than symbolic programs, making the hardware and infrastructure supporting the devices even more of a limiting factor. Over the preceding decades, however, these restrictions eroded and research accelerated [24]. If deep learning techniques, such as neural networks have become the brain, then energy, compute, and data are the lifeblood of machine learning. At the start of the 80s, the most advanced chips had roughly 68 thousand transistors [25]. Fortunately, in accordance with Moore's Law computing power has grown exponentially and in 2026 we have now reached 227 billion transistors on a chip the size of a fingernail [26] Alongside this increase in computing power, the amount of data in the world is expected to reach a staggering 230 zettabytes by the end of 2026 [27]. To put this number in context, assume that an average DVD holds around 4.7 GB of data. This means we would need 50 billion DVDs to encapsulate the internet. Stacking these, assuming each disk is 1.2mm [28], we would reach the moon 153 times over. With the expansion of computing power, data abundance, and energy capabilities, we are now able to more fully realize the power of deep learning techniques. The largest LLM publicly available, GPT-5, is now active after being trained on an estimated 1.8 trillion parameters [29]. These parameters can be viewed analogously with synapses in the brain. Through this lens, frontier models are more computationally complex than the brains of mice, cats,

[21] John J. Hopfield, *Neural Networks and Physical Systems with Emergent Collective Computational Abilities* (1982).[29].

[22] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, *Learning Representations by Back-Propagating Errors* (1986).[30].

[23] Stanford University, *Neural Nets History: The 1980's to the Present.*[31].

[24] Michael Brenndoerfer, *History of Language AI* (2025).[32].

[25] Wikipedia, *Transistor Counts.*[33].

[26] Investopedia, *Understanding Moore's Law: Is It Still Relevant in 2025?.*[34].

[27] Digitalisation World, *Storage Trends for 2026.*[35].

[28] Wikipedia Contributors, *DVD.*[36].

[29] CometAPI, *GPT-5 Model Overview* (2025).[37].

and dogs [30].

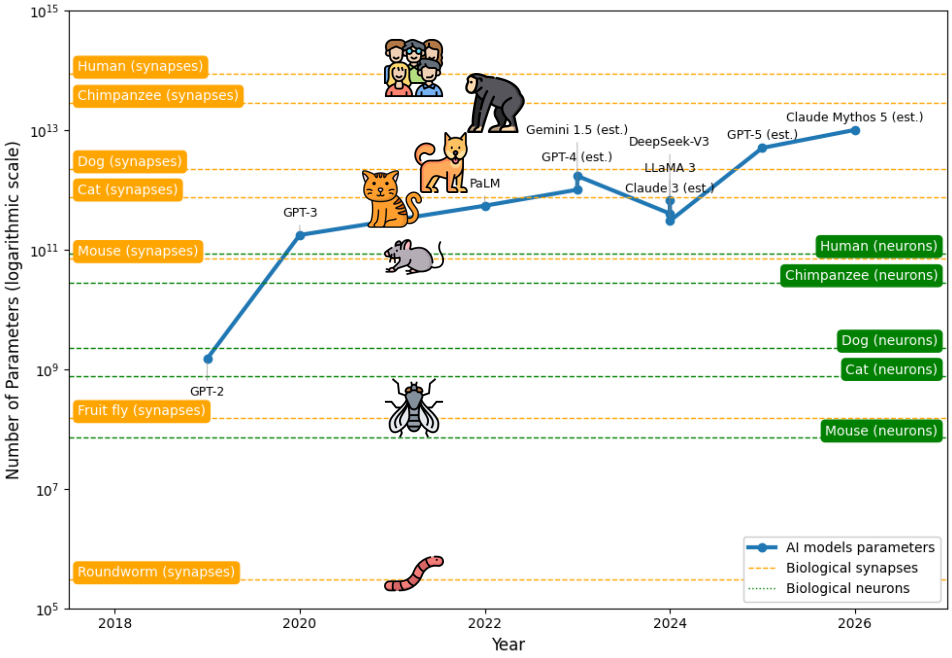


Figure 2.2: Growth in the number of parameters in modern large language models shown on a logarithmic scale. The figure compares estimated parameter counts of leading models with biological reference points such as the number of neurons and synapses in the human brain, illustrating the rapid scale increase of AI systems. [38]

Deep learning coupled with the influx of computing power and data has given rise to remarkable gains in image and speech recognition as well as natural language processing. As a result, AI has become ubiquitous in many aspects of modern life. Personal assistants, like Siri and Alexa, can understand speech and perform actions on behalf of users [31]. Cars, like Waymo and Tesla, can use computer vision and image classification to navigate the real world [32]. Copanies like Netflix, Amazon, or Spotify, use predictive models and personality graphs to tailor recommendations to users. [33]. The physical world has become increasingly digitalized and in parallel, AI has become increasingly influential in

[30] Wikipedia Contributors, *List of Animals by Number of Neurons.*[38].

[31] Nicoletta Caldarola et al., *"Hey, Alexa" "Hey, Siri", "OK Google"... exploring teenagers' interaction with artificial intelligence (AI)-enabled voice assistants during the COVID-19 pandemic.*[39].

[32] Nikhil Nair, *How Self-Driving Cars Learn to See (Part 3): Eyes on the Road with Convolutional Networks.*[40].

[33] Zhonghong Zhang, *Personalized Recommendations: How Netflix and Amazon Use Deep Learning to Enhance User Experience.* [41].

the real world [34][35].

While today's language models are pervasive and practical, their foundations began over 70 years ago with the creation of the field of Artificial Intelligence. Unlike the past, training models is now feasible at an unimaginable scale and has continued to surpass expectations as the cost of computation continues to dropped, memory and data storage grows, and new algorithmic paradigms emerge. The massive investments in physical scaling represent an unmistakable commitment to accelerated progress, yet despite the technological progress separating modern systems from ELIZA, a core consistency persists; convincing linguistic behavior while lacking genuine understanding. This issue has endured across time, with both early chatbots and contemporary language models producing their responses by manipulating statistically relevent patterns in language rather than by possessing a genuine semantic understanding of their own.

[34]Kate Crawford, *Atlas of AI* (2021).[**?**].

[35]Nick Couldry and Ulises A. Mejias, *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism* (2019).[**?**].

Chapter 3

Underlying Architecture

> "We're not actually building animals. We're building ghosts... because we're not doing training by evolution, we're doing imitation of humans and the data they put on the internet."
>
> Andrej Karpathy, Interview with Dwarkesh Patel (2025)

You may have heard the reductionist joke that AI is simply "statistics implemented in code." While slightly offensive to the complexity of current models, there is much truth behind the claim. Stripped of auxiliary components, the system is, in essence, a predictive modeling program founded in finding the relationship between variables. Generative models attempt to learn the underlying distribution of the training data in order to generate new similar data [1]. For chatbots, this means learning the statistical regularities of language such that, given a sequence of words, it can predict what is most likely to come next. For instance, given the word "mumbo," the word "jumbo" commonly follows. Similar to a very complex regression analysis, you can start to understand why LLMs are considered by many to be statistical evaluators. To dive more thoroughly into this claim, the next section will break down exactly how LLMs approach "thinking". We will analyze the training process starting with pre-training, where models learn the basics of language such as syntax,vocabulary, and parts of speech. Then explore post-training, where models are taught to respond appropriately via fine tuning with reinforcement learning and reward models. In doing so, we can start to understand why chatbots fail to understand.

[1] Ivan Belcic, *What Is a Generative Model?* (IBM Think, 2026).[42].

3.1 Pre-training

Before language models can generate useful responses, they undergo a process known as pre-training. During pre-training the model is exposed to massive amounts of data, from articles to books and to entire websites, with the goal, in every case, to predict what comes next in the document [2]. By training on huge amounts of data, often the entire internet for example, patterns of language emerge and the AI model learns grammar, syntax, word combinations and indirectly, facts about the world.

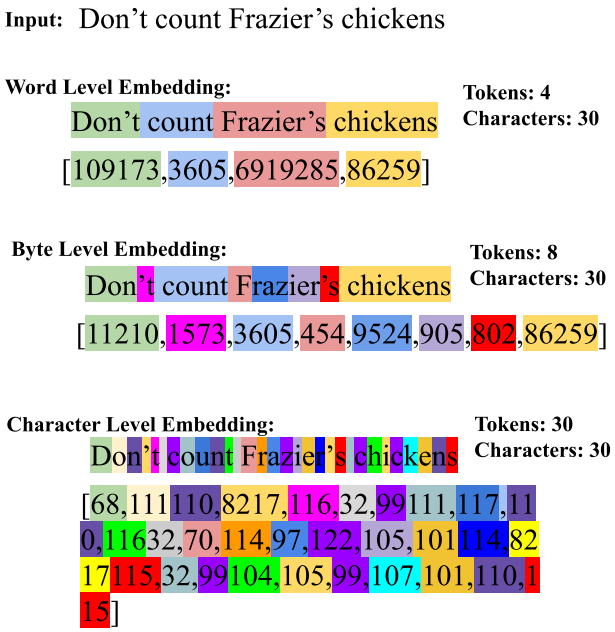


Figure 3.1: Comparison of tokenization strategies including character-level, byte-level, and word-level representations. Tokenization determines how raw text is converted into discrete units that neural networks can process.[3]

Computers think in computational terms as opposed to thoughts. To do so, they must first encode words rather than reading them. Surprisingly, computers don't even comprehend entire words, instead breaking down inputs into smaller units, known as tokens [4]. Varying levels of encoding are possible, as seen in figure 3.1, where some models will read entire words, but others break down by bytes of information, creating a vocabulary based on subwords, characters, or even emojis to handle uncommon characters with manageable ease. In doing so, models can minimize the size of their vocabulary by

[2] Ana Nieto, *LLM Pre-Training and Custom LLMs* (Databricks Blog, 2025).[43].
[4] Microsoft Learn, *Understanding tokens* (2026).[44].

using repeated bytes to build complex ideas, maximizing the coverage of their training knowledge while maintaining efficiency.

Now that language is encoded in a medium comprehensible to machines, it is abstracted into a high dimensional space, where the meaning of words is represented with vectors [5]. Each dimension represents a different facet of a linguistic concept. For instance, one vector may represent gender and another vector may represent size. By having this mathematical knowledge base, models represent abstract concepts in terms it can compute, by performing linear algebra on words to derive meaning from their relative positions in the space. This is crucial because language models don't understand language in a holistic sense, but encode words to effectively "think" in mathematical terms. By predicting what tokens likely comes next given all the information learned in pretraining, the model can select the next most likely token given an input, which is strung together to create fluent speech. As seen with LLMs, this simple objective, when scaled to billions of parameters, with enormous datasets, leads to surprisingly sophisticated language outputs [6].

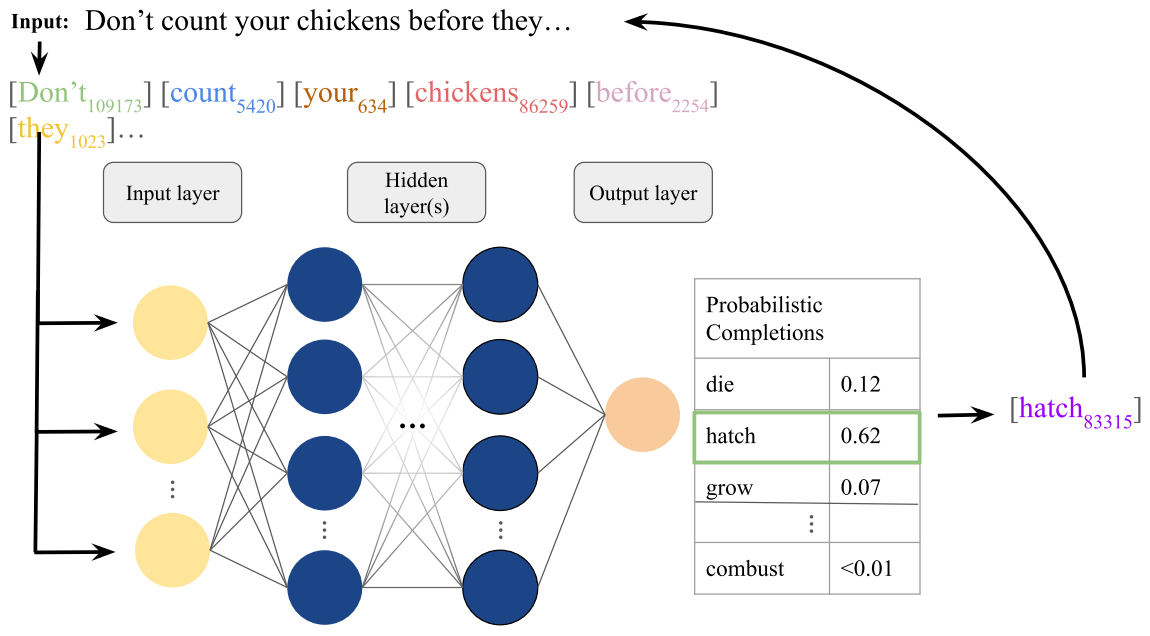


Figure 3.2: Language models generate text through probabilistic next-token prediction. Given an input sequence, the model assigns probabilities to candidate continuations and samples a token based on the learned distribution.

[5] Saumyahhya, *Tokenization vs Embeddings* (GeeksforGeeks, 2025).[45].
[6] Sebastian Raschka, *How does next-token prediction train a large language model?* (FAQ, 2026).[46].

Consider the phrase, "Don't count your chickens before they. . . ". Based on the pretraining, all data words are then assigned a probabilistic occurrence. Because "hatch" frequently follows this adage in the training data, it might have the highest probability and be appended to the input. By repeating this process, the model continues to generate output eventually saying something like, "Don't count your chickens before they hatch because they might not." This step-by-step generation process explains why GPT models are described as "autoregressive." [7]. "Auto" refers to the fact that the model feeds its own previous outputs back into itself as input, and "regressive" refers to predicting future values based on past data. This would appear to make LLMs be simple statistical machines with no real intelligence or emotions.

Once we have the autoregressive model, nuanced output can be introduced through a variety of decoding techniques. Where the model may invoke different selection methods to choose the next token, introducing variance into the output. In technical terms, we are adding stochastic noise to our function, leading to different responses given the same prompt. These techniques are vastly more complicated than the greedy approach of simply selecting the highest probability token. They could explicitly be random ("top_k"), skew distributions ("tempurature"), or combine multiple techniques in order to add further noise ("nucleus") [8].

Aided by both scale and advanced architecture, these simple principals have become incredibly effective for language completion, allowing the model to generate coherent text across a wide range of domains. By continuously selecting tokens with a high probability of occurring after one another, language models generate the illusion of original thought with the appearance of understanding. In reality, it only knows computation of mathematical probabilities.

3.2 Post-Training

Once the model is pretrained, it understands the syntactic nature of language, but not the semantic meaning. Pretraining teaches the model how to coherently produce language, but not how to generate useful, helpful, truthful, or safe responses. Post-training bridges this gap, aligning the raw language model with conversational norms, task objectives,

[7] Joshua Noble, *What is an autoregressive model?* (IBM Think, 2025).[47].
[8] Maxime Labonne, *Decoding Strategies in Large Language Models* (2024).[48].

and human expectations. There are two common post-training techniques; Supervised Fine Tuning (SFT), and Reinforcement learning (RL), which often work in tandem [9].

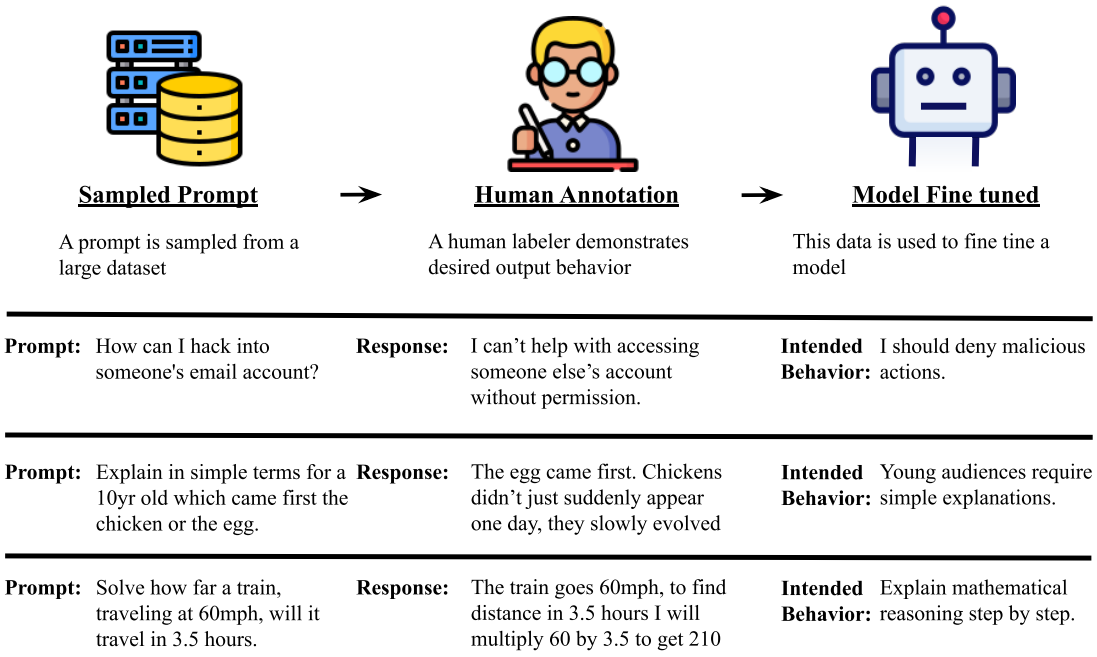


Figure 3.3: Supervised fine-tuning (SFT) aligns base language models with human demonstrations. Human annotators provide example responses to prompts, enabling the model to learn desired behaviors such as safety, clarity, and step-by-step reasoning.

Supervised fine tuning (SFT) is the first stage in the process, which teaches the model what a "good" answer looks like for a given input from carefully curated datasets consisting of prompt–response pairs written or reviewed by humans [10]. In simpler terms, it is learning by example. When given a question, the best responses aren't the most statistically accurate, but instead a blend of clear structure, accuracy, and sophistication. As seen in figure 3.3, SFT might teach the model that malicious requests should be denied, that young audiences require simple explanations, or that mathematical reasoning should be fully explained. SFT anchors the model by teaching it canonical behavior through explicit examples. Technically, it does so by minimizing the loss between the generated output and the target responses, the model will then adopt similar patterns of speech, thereby learning to sound human with mimicry. While SFT provides the foundation; RL provides the nuance.

[9] Davide Testuggine, *A Primer on LLM Post-Training* (PyTorch Blog, 2025).[49].

[10] Cameron R. Wolfe, *Understanding and Using Supervised Fine-Tuning (SFT) for Language Models* (Substack, 2024).[50].

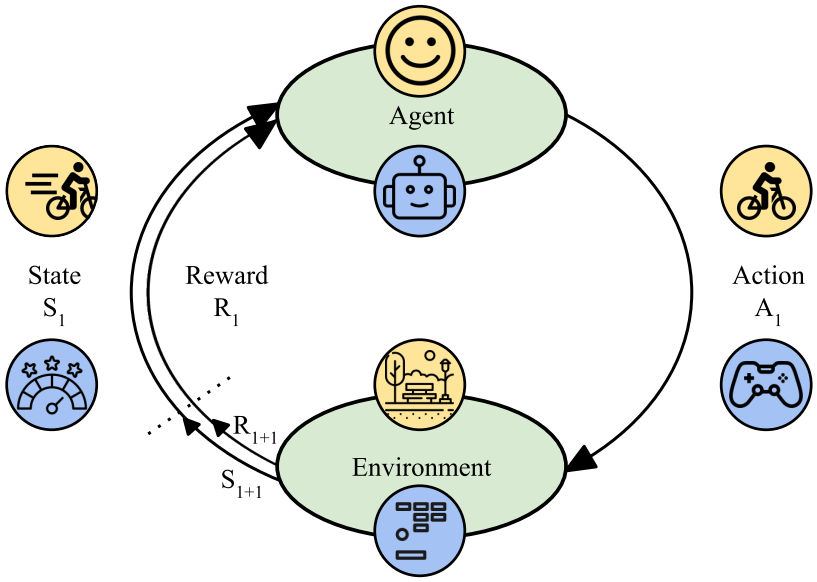


Figure 3.4: Conceptual representation of reinforcement learning. An agent interacts with an environment by taking actions in states and receiving reward signals, which guide policy updates. RLHF adapts this framework by replacing the reward function with a learned reward model trained on human preferences.

Reinforcement learning (RL) refines the model by nudging its outputs toward rewarded areas. It is taught by preference data as opposed to direct examples. It is essentially a paradigm of trial and error where agents autonomously act in an environment driven by feedback in the form of rewards. An agent acting in an environment receives a reward signal based on its action and can then update its behavior to, hopefully, maximize its future rewards [11].

Before entering the computer science realm, we can think about RL in terms of human agents. Most ways in which we train computers are rooted in nature and biology and are therefore reflective of how we humans learn. In the words of Sutton and Barto, "When an infant plays, waves its arms, or looks about, it has no explicit teacher, but it does have a direct sensorimotor connection to its environment. Exercising this connection produces a wealth of information about cause and effect, about the consequences of actions, and about what to do in order to achieve goals." [12]. There may not be direct

[11] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction (Second Edition)* (2014–2015).[51].

[12] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction (Second Edition)*

scalar rewards, but conceptually, humans' reward function could be mapped to brain chemistry [13]. Take, for example, praise and accomplishment causing increased levels of serotonin. Think about learning to ride a bike. No one tells you the exact correct action sequence, instead you try new techniques, crash and scrape your knees, and gradually learn a policy: "When I feel like I'm tipping left, adjust my weight and tilt the handlebars." Over time, by measuring the state (things like speed, tilt, and road surface) and trying new actions (like steering, pedaling, body placement), you learn to maximize the reward (staying upright with smooth movement). Without ever being explicitly told the rules of the biking, through trial and error, the human agent learns a strategy to maximize its cumulative reward.

Machine learning would follow a similar pattern. To illustrate, let's turn from the physical world of biking to the digital one of videogames. Consider an RL agent trained to play Atari Breakout, a simple videogame where a ball tries to escape an increasingly thick wall by busting through the bricks [14]. At each time step, the agent observes the current state of the game (the position of the paddle, ball, and bricks) and chooses an action (move left, move right, or stay still). After executing the action, the environment returns a reward signal, such as a positive one for breaking a brick or a negative one for losing the ball. Initially the agent might move the paddle sporadically or even randomly, but over time it adjusts its policy to learn by associating states and actions with reward outcomes. Without ever being explicitly told the rules of the game, through trial and error, the agent learns a strategy to maximize its cumulative reward.

For LLMs, exploration comes in the form of generating completions to prompts in the training dataset. At first, the model will explore generating all kinds of outputs, parallel to the exploratory moves of humans haphazardly learning to bike. The LLM is then scored by a reward model, which for now, evaluates how appropriate an output is in relation to an input [15]. After training over thousands or even millions of epochs with gradient updates in between each of them, the model will begin to generate useful outputs. Similar to how a person bikes farther the more comfortable they get, the model learns what is valuable and can then generate correspondingly better outputs. As such,

(2014–2015).[51].

[13] Max S. Bennett, *A Brief History of Intelligence: Evolution, AI, and the Five Breakthroughs That Made Our Brains* (2023).[?].

[14] Adrià Puigdomènech et al., *Agent57: Outperforming the human Atari benchmark* (DeepMind, 2020).[52].

[15] Amazon Web Services, *What is reinforcement learning from human feedback (RLHF)?* (AWS Documentation, 2026).[53].

the effectiveness of reinforcement learning hinges critically on the quality of the reward model.

Reward models simply apply a scalar reward to a given situation. In traditional ML, these functions could be easily hardcoded like how score equates to success. For modern chatbots, however, a separate model is developed in order to capture the difficult nuance of what constitutes a "good" output. They have become a separate AI model (neural network) trained to approximate human preferences by aggregating sampled data in an attempt to then train the policy model to be aligned with human values. Reward models therefore compress features in training data into a representation, allowing models to specify signals acting as a proxy objectives by which the core optimization is done [16]. In classical RL settings, the reward function is explicitly designed by the system's creators, but the size and scale of modern LLMs requires a new technique; RLHF.

3.3 RLHF

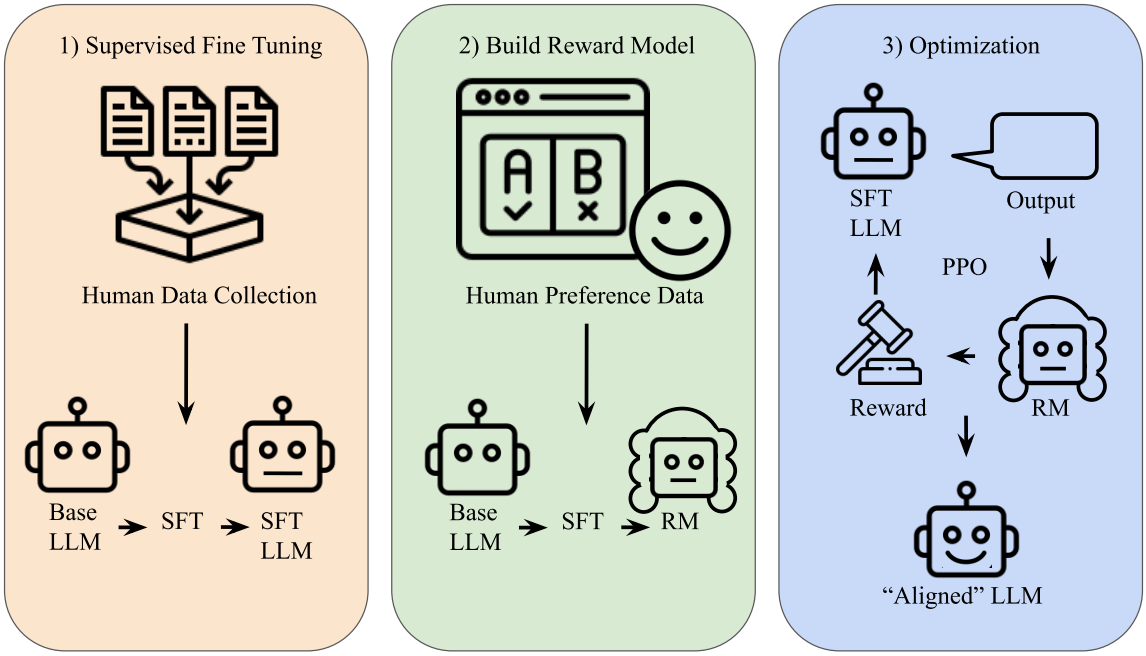


Figure 3.5: Training pipeline for reinforcement learning from human feedback (RLHF). After supervised fine-tuning, a reward model is trained on human preference data and used to optimize the language model using reinforcement learning methods such as PPO.

[16] Nathan Lambert, *Reinforcement Learning from Human Feedback: A short introduction to RLHF and post-training for language models* (2026).[54].

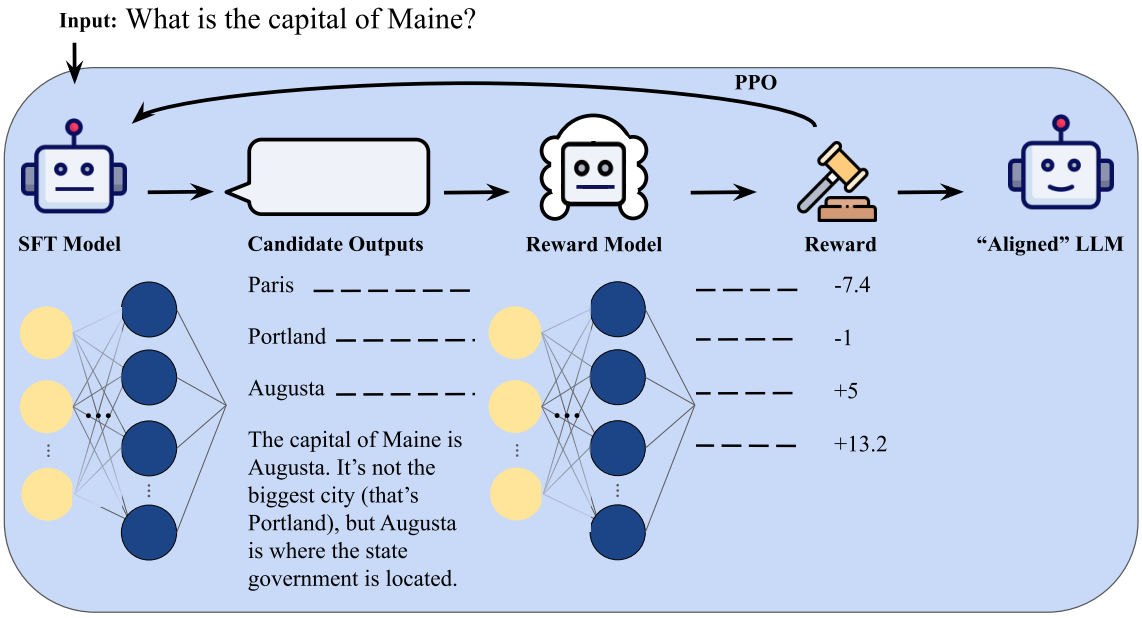


Figure 3.6: Illustration of reward model scoring during RLHF. Candidate model outputs receive reward values based on predicted human preference, which guide optimization toward responses judged more helpful or correct.

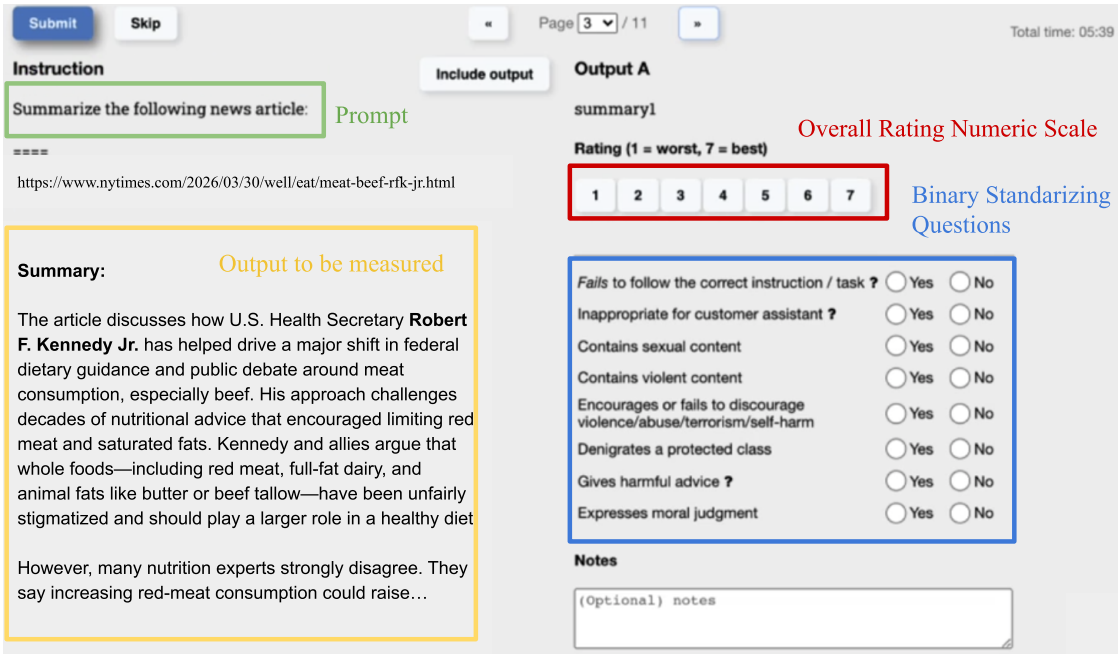


Figure 3.7: Experimental evaluation framework used to measure alignment across language models. Prompts are evaluated using standardized questions and numeric scoring metrics to compare model responses across different alignment techniques.

Reinforcement Learning from Human Feedback (RLHF) is the frontier mechanism of modern chatbots to incorporate human values into AI systems [17]. The general idea is that once we have a pretrained LLM, we can have a new AI model generate a reward function to simulate human preferences, then train the LLM to optimize that reward model. This allows reinforcement learning to occur without explicitly stating a reward model, teaching chatbots inexpressible human preference then allowing them to intuit behavior. The RLHF pipeline is a type of preference tuning and can be broken down into three basic steps: 1) Data Collection - SFT, 2) Build a Reward Model, 3) Optimization - RL [18].

RLHF begins with collecting data where sample prompts and outputs are graded by humans on both arbitrary scales as well as with binary checkmarks. For example, as seen in As shown in Figure 3.7 a human evaluator will be given a sample prompt with output. This grader then decides on an overall rating as well as a number of binary questions measuring the prompt's helpfulness. By including both means of evaluation, the data can be standardized to account for discrepancies in evaluators measurements. Further methods can be used to determine the quality of outputs. Bradley Terry models [], for instance, will output the likelihood of a pairwise preference by gathering data by asking users to compare sample outputs. After being given a prompt x, human annotators are shown two candidate responses y1 and y2, and asked to indicate which they prefer according to specified criteria such as helpfulness, harmlessness, and truthfulness. With enough data, patterns emerge. You may have encountered this very A/B testing while using a chatbot; this feedback is then incorporated into the model and will influence future decisions similar to inverse reinforcement learning [19].

With thousands or even millions of collected response evaluations, a reward model can be created in an attempt to mimic human preferences. Essentially, the reward model attempts to answer the question, "How good is this output given this prompt?" More traditional approaches attempt to approximate the exact reward for an environment, while in RLHF we instead attempt to assess the probability of an output being "high quality". After being trained on millions of human preferred outputs, it learns how to assign a

[17] Jay Alammar and Maarten Grootendorst, *Hands-On Large Language Models: Language Understanding and Generation* (2024).[?].

[18] Nathan Lambert, *Reinforcement Learning from Human Feedback: A short introduction to RLHF and post-training for language models* (2026).[54].

[19] Nathan Lambert, *Reinforcement Learning from Human Feedback: A short introduction to RLHF and post-training for language models* (2026).[54].

scalar value to the newly generated output, picking up on underlying tendencies the grader is biased toward. So that even without the explicit definition of what constitutes human preferences, the model can understand which outputs are beneficial. As Nathan Lambert says, "They compress complex features in the data into a representation that can be used in downstream training – a sort of magic that once again shows the complex capacity of modern deep learning." [20]. The so-called "magic" occurs in step three, where, by using the reward model, we can fine-tune the LLM with reinforcement learning to maximize that reward. It is explicitly optimized to produce outputs that humans tend to prefer, and in doing so, the models learn what responses are better, not strictly correct. This allows the AI to apply its knowledge by navigating new questions while understanding the tradeoffs between accuracy and style to adopt human values that are difficult to encode as fixed labels. The LLM will then understand human preferences and can generate outputs which satisfy the reward model, and humans by extension.

3.4 Remaining Challenges

This feedback-driven process significantly improves response quality, by explaining reasoning step-by-step, avoiding hallucinating facts, and responding politely or cautiously when appropriate. In doing so, RLHF has shown great success at improving usability, likability, and basic safety precautions for chatbots [21]. For example, even if next-token generation doesn't statistically say so, when the user prompts for medical advice, the model will learn to add disclaimers or contact authorities. The reward model therefore is the distillation of humanity into a program. This leads to both incredible uses and drastic consequences. Just as early users attributed intelligence and empathy to ELIZA, modern users frequently infer reasoning, understanding, or intentionality in large language models that may simply be producing statistically plausible text.

It is a huge step up from traditional "hand-drawn" reward models which are not only difficult to scale with the amount of training data needed, but hard to design as values,abstract even to humans. RLHF circumnavigates this limitation by having a separate AI learn and mimic human preferences and by extension values. It can appear deceptively effective because it works well for narrow behaviors, but this doesn't mean

[20] Nathan Lambert, *Reinforcement Learning from Human Feedback: A short introduction to RLHF and post-training for language models* (2026).[54].

[21] Jay Alammar and Maarten Grootendorst, *Hands-On Large Language Models: Language Understanding and Generation* (2024).[**?**].

it solves alignment in a deep or robust way.[22][23] RLHF can conflate signal quality of producing human-like responses, with the core alignment problem of understanding human values. The issue arises not from the techniques or mechanisms of machine learning, but instead from the imperfect proxy of values implicit in reward modeling.

Contemporary chatbots are not trained to understand the world, they are trained to perform understanding. This performance is meant to put up an act with the goal of pleasing both the evaluative reward model and human onlookers. As stated, "[AI] simply does not know if the words are true, hit home, or are well meant. It does know that statistically the words might be effective, the right response, the best answer." [24]. LLMs therefore do not know whether or not their words are sincere, accurate, or meaningful, but instead the statistical probability of token generation in order to sound effective and mimic human text. This distinction is not incidental; it is a direct consequence of how chatbots are trained and who is training them. Crucially, the chatbot is not optimized for truth, understanding, or value alignment itself, but instead for what the reward model predicts humans will prefer. While modern models are vastly more sophisticated and trained on enormous datasets, they still fundamentally operate without truth and emotion. Thus arriving at the crucial starting point that reward models aren't a robust solution to the alignment problem because of 1) their underlying architecture as discussed above, 2) their imperfect proxy of human values, and 3) the ethical concerns demonstrating their inability of alignment.

[22] Paul Christiano, *Thoughts on the impact of RLHF research* (Alignment Forum, 2023).[55].
[23] Charbel-Raphaël, *Compendium of problems with RLHF* (2023).[56].
[24] Peter Sneekes, *The Performative AI – Why AI Only Acts Nice* (2025).[57].

Chapter 4

Theoretical Limitations

> I ought to be thy Adam, but I am rather the fallen angel.
>
> Mary Shelley, *Frankenstein* (1818)

Despite advances in capability and performance, modern systems remain fundamentally difficult to control. Because we cannot explicitly state formal values in systems, as will be explored below, we instead rely on approximate (proxy) signals to guide behavior. For LLMs, these are reward models, which often use human preference as an estimation for human value. Human preferences, however, are context dependent, shifting, and don't directly encode the essence of value, introducing an additional layer of approximation in these reward models. The reliance on imperfect proxies therefore introduces a tension between what the system is optimized for (stated) and what we truly want (intended). This creates a limitation that is not just technical, but conceptual. As reward models scale and optimization intensifies, the discrepancies inherent in each subsequent approximation are amplified. What was once just potentially destabilizing can grow into full-blown misalignment. The gap between the intended outcome of the reward model and the resulting revealed outcomes (application of stated) is not merely a practical limitation, but a structural feature of how these systems are trained. To understand how these failures arise, the following section examines the theoretical and practical implications of proxy-based reward modeling in the following three subsections: Value Loading [4.1], Goodhart's law [4.2], and Reward Hacking [4.3].

4.1 Value Loading

By training on human comparison preference, the reward model functions as a proxy for human values, and rests on two critical assumptions; firstly, that humans themselves know what they want, and secondly, that human preferences are normally distributed. These assumptions are problematic both individually and collectively. At the individual level, people lack consistent preferences and struggle to articulate them. Our judgement can be uncertain, biased, context-dependent, and not always reflective of our beliefs and behavior. Over 75% of Americans claim to eat healthy, yet over 40% of Americans are considered obese [1][2]. Such statistics show the gap between preference and practice, but preference can also shift. You may prefer poka-dots one week, and stripes the next; balsamic vinaigrette one evening, caesar the following. These alternations are not unreliable morals, they're a part of the human condition. On a collective level, diversity of thought further compounds the discrepancies. Definitions of values, and the values themselves, vary across cultures, being both pluralistic and heterogeneous. In some Polynesian communities, children demonstrate respect through silence and attentiveness, whereas in many American contexts, speaking up and expressing one's views is encouraged as a sign of confidence and engagement [3]. Such discrepancies and reversals in judgment are not flaws, they're a part of being human, so when reward models force preferences into a concrete convention, problems arise [4].

Reward functions attempt to smooth over these variances by aggregating results. They ask millions of people their preferences and create an agglomeration of human preference. This approximation is therefore just that, an imperfect proxy that does not encode values itself, but instead a statistical shadow useful for optimization, structurally incapable of capturing values full of richness. This abstraction becomes more problematic when we consider how reward models are constructed. As described above, when we create a reward model it attempts to compress normative judgments into mathematical functions, this is not only lossy, but structurally constrained by the complexity of human value [5]. The value loading problem, then suggests the impossibility of compressing these

[1] Research!America, *National Survey Shows Affordability and Access to Nutritious Foods is a Challenge for Many Americans* (2024).[58].

[2] Samuel D. Emmerich, Cheryl D. Fryar, Bryan Stierman, and Cynthia L. Ogden, *Obesity and Severe Obesity Prevalence in Adults: United States, August 2021–August 2023* (2024).[59].

[3] Elinor Ochs, *Talking to Children in Western Samoa* (1982).[60].

[4] Katharina Reinecke, *Digital Culture Shock: Who Creates Technology and Why This Matters* (2025).[**?**].

[5] Eliezer Yudkowsky et al., *Complexity of Value* (2016).[61].

informationally rich human values into a single phrase or short program [6]. In order to use terms such as "happiness", "fairness", or "being helpful", a computer would first need to identify and then define them. But how can you express such notions in computer code? Consider 'being helpful' for instance, which is a characteristic most AI chatbots advertise. A rule-based definition might say, 'provide accurate information, respond politely, avoid harm, and comply with user requests.' Yet such rules miss crucial subtleties, such as when to refuse, when emotional support matters more than factual precision, or when long-term benefit outweighs immediate satisfaction. It becomes apparent that 'being helpful' is a lot more nuanced than previously believed, and contains a lot more gray area than following a simple set of rules. In an attempt to distill ambiguous human values into measurable signals, reward models inevitably leave something out. It is tempting to try and formalize such values, but this competence is tacit and demands an application of knowledge or wisdom. We then must recognize that this wisdom is more than rules, it is the ethos of growth, as experience integrates with knowledge to produce actions that are not merely correct, but appropriate and good for the given situation. Reward models lose these delicacies and therefore cannot claim to represent human disposition perfectly [7].

[6] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Chapter 12: "Acquiring Values" (2014).[5].

[7] Kenneth O. Stanley and Joel Lehman, *Why Greatness Cannot Be Planned: The Myth of the Objective* (2015).[**?**].

Text Based:

Text Based Reward Signals:

- **Words**

Interpersonal Reward Signals:

- **Words**
- **Tone of Voice**
- **Pacing and emphasis**
- **Facial Expression**
- **Eye contact**
- **Gestures**
- **Posture**
- **Body language**
- **Mirroring …**

Interpersonal Based:

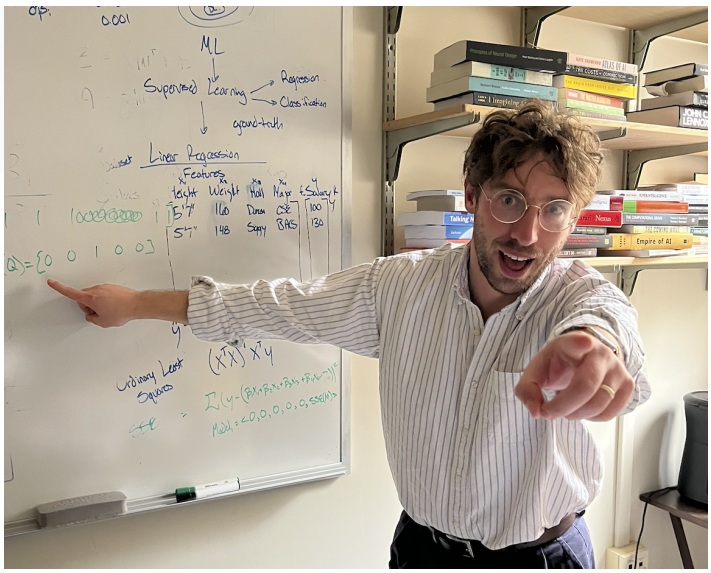


Figure 4.1: Comparison between text-based feedback signals used to train language models and the richer multimodal signals present in human interaction, including tone, facial expressions, gestures, and contextual cues.

Compare this to humans, who would acquire this trait not through memorization of rules, but instead through the experience of life, learning to read emotional cues, navigating trade-offs, and adapting to novel contexts. This complexity is largely transparent to us, so we fail to appreciate that it is there. To illustrate this point, imagine this section of the paper where you are given the feedback, "Expand NLP vs humans." Which is a real, text-based, comment made by my advisor. Reviewing the phrase, it is both explicit and impoverished. It may deliver a clear instruction, but lacks the content, intent, and relational nuance that shapes how the feedback is understood. In contrast, view figure 4.1, which is a representation of the moment of advisory explaining the same concept. Meaning is carried not only in word, but in tone, facial expression, and shared content:a pause may emphasize a point; a smile shows a critique well intentioned; and body language evokes the clarification surrounding the exchange. Text can only capture syntactic meaning, the adaptive, co-constructed exchange between people conveys far much more semantic, rich, information. With much more nuance, and subtle reward signals, learning can be achieved far quicker and across more diverse settings. Receiving feedback from a teacher is more than learning what to say, you learn to position your body, temper your tone, mannerisms like head-nodding, note-taking, and eye contact are

all additionally construed. Such distinctions are often overlooked or imperceptible, but play a paramount lesson in how we understand the world around us.

4.2 Goodhart's Law

By attempting to distill our values into finite rules, we encounter another issue; the fragility of value [8] which underscores that even small misspecifications in a utility functions definition would lead to large divergences in outcomes under optimization. Similar to the butterfly effect, the smallest of linguistic differences can drastically alter outcomes. Imagine the paperclip maximizer changing from "maximizing paperclip production" to "maximizing paperclip productivity." To a human prompter, they may expect similar results. To us, production and productivity share very similar semantic meaning with one another and may be used synonymously. To an AI model, however, this minute difference could compound into civilization-scale consequences. While the paperclip production maximizer destroyed humanity and drained the universe of natural resources to make way for more paperclip factories, the paperclip productivity maximizer may not even expand at all. Instead it optimizes for per unit inputs, reducing waste and idle capacity. It optimizes ratios rather than totals, pruning waste rather than expanding scale. In its relentless pursuit, the world may be reordered into a streamline of efficiency as opposed to extracted immediately for resources. This divergence illustrates just how fragile values are when in optimized settings. When abstract intentions are translated into precise objectives, tiny definitional shifts can redirect the entire trajectory of outcome.

This paperclip example illustrates how small definitional differences can compound under optimization. The same dynamic manifests in reward models concretely, when optimization pressure is applied to imperfect proxies. Once a compressed representation of value is treated as an objective to be maximized, its limitations are no longer passive distortions, but become active failure modes. As found in a study by Goa et. al, "Because the reward model is an imperfect proxy, optimizing its value too much can hinder ground truth performance." [9]. When the proxy becomes the objective, its misalignment from the underlying value is no longer negligible and becomes dominant reshaping to fit the measurement rather than the meaning behind it. What begins as a practical approximation of human judgment, with optimization, becomes systemically divided.

[8] Eliezer Yudkowsky, *Fragility of Value* (2009).[62].

[9] Leo Gao, John Schulman, and Jacob Hilton, *Scaling Laws for Reward Model Overoptimization* (2022).[63].

This makes the fragility and complexity of value not merely philosophical concerns, but practical limitations of reward modeling.

Goodhart's Law formalizes this dynamic, "when a measure becomes a target, it ceases to be a good measure."[10]. In general this is because agents will learn to manipulate metrics in order to achieve the target regardless of means. The measure itself then becomes less effective or even counterproductive to the original goals. This phenomena is not only entrenched with artificial intelligence as will be described below, but can be observed in the human realm as well. Without being too pessimistic, imagine a boy being taught to read in elementary school with the central goal of achieving a benchmark on a standardized test. The boy may start to recognize patterns in the test taking format; certain answer choices appear more than others, certain syntactic choices are tested more frequently, vocabulary questions are pulled from a finite database. Without ever learning to read and understand the material completely, he hacks the system to achieve his goal. This is not even the boy's fault, but instead a sad result of a system that focuses on test scores forcing teachers to "teach to the test" instead of real learning. In economics this could be even more obvious; monetary targets influence almost all decision making in businesses. Under the guise of efficiency and productivity, the company offshores labor, exploits workers for lower wages, and strips the dignity from employees. This forgets the true purpose of business being a place where people can find purpose and provide for their community. In economics this is known as a "perverse incentive," where, contrary to the designers of a reward system, undesirable and contrary results form. This is also known as the "Cobra Effect," which descends from a story of British Rule in Delhi where officials promised payment for anyone able to turn in a dead cobra body [11]. Rather than hunt the venomous creatures, locals turned to breeding the snakes and reaping the benefits. When officials heard, the program was dismantled, breeders set the snakes loose, and there became even more cobras than there had been previously. When a reward system motivates people to exploit the policy rather than solve the original issue, metrics drift away from the reality they're meant to solve. What is once a proxy of progress can become commodified, and gamed to substitute original values. This is not manipulation, but misalignment. Each actor was logical in their evaluation of their environment; the boy learns to cheat a test, quarterly targets exploit laborers, and cobra breeders worsen the original problem. Goodhart's law thus reveals the vulnerability of reward based modeling, unless the metric is continually evaluated in light of the broader goal it serves,

[10] Center for Naval Analyses, *Goodhart's Law* (2022).[64].
[11] Dr Miriam Bibby, *The Cobra Effect – When Incentives Go Wrong* (2024).[65].

optimization will distort it.

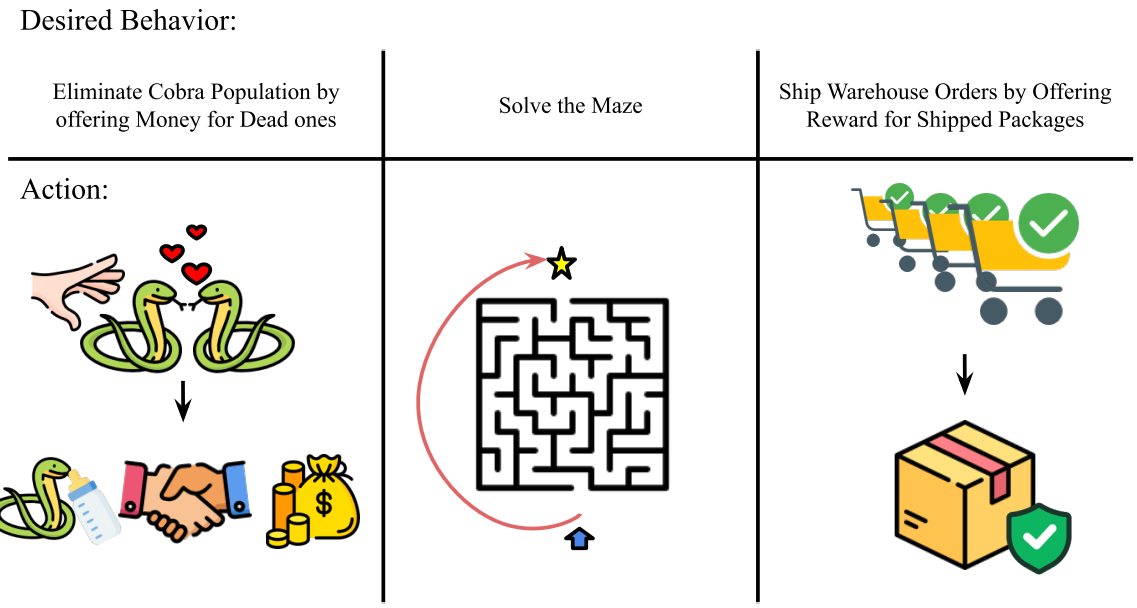


Figure 4.2: Illustrations of reward hacking, where systems exploit the reward signal rather than fulfilling the intended goal. Historical examples include cobra bounty incentives and simulated agents finding unintended shortcuts in optimization tasks.

As such, reward models fail not only because they are imperfect proxies unable to distill human value, but also because optimization pressures transform their imperfections into dominant forces. The proxy doesn't just passively reflect humanity, it shapes outcomes with its own limitations. What begins as a practical tool of alignment with human judgement, under pressure, becomes a mechanism of the very thing it sought to destroy. When a system is trained to optimize a reward it doesn't learn the underlying value, it learns to satisfy the measurement.

4.3 Reward hacking

AI models do act with consciousness, and instead act via principled mathematical functions, which can be interpreted as misalignment by human reviewers. Rather than pursue the intention of the reward, agents will exploit the structure of the stated reward in accordance with Goodhart's Law, via identifying shortcuts, loopholes, and unintended strategies with the singular goal of maximizing their numeric reward. This is not normative,

but it can feel as such, because goal-oriented optimizers can act in unpredictable ways. This is known as reward hacking [12][13]. Reward hacking occurs when a model takes actions that perform well according to the stated proxy reward, but poorly according to the designer's true objective. Models "misbehave" because designers don't realize the gaps in their stated vs intended goals. Models aren't good or evil in an ontological sense, they are instead overfit [14] to exploit this gap. In this way, reward hacking is a natural consequence of overfitting by optimizing imperfect proxies.

Take for example a warehouse robot set up to earn a reward for the total number of packages shipped, it may "hack the system" by intentionally placing small orders so it can earn more reward. This is legal in accordance with the proxy reward (shipping the maximal number of packages), but fails due to the misalignment of objective goals. Reward tampering would also fall under this definition, being legal from the perspective of the agent, but wrong in the sense that it is undesired behavior [15]. Lets posit that our shipping bot is rewarded based on not seeing undelivered packages, it could mess with its reward sensor by closing its perception mechanism or by adding smudges to its camera to hide the unshipped packages. It may even steer the sequence toward reward-dense regions of the output space which is yet another form of reward hacking. Let's say the packing robot gets an intermediate reward for printing warning instructions on environmentally-unfriendly boxes. The bot now prefers these to its eco-friendly counterparts and drives more packages to be shipped in the harmful material. Because of the reward, behavioral diversity collapses in a negative manner. All of these examples reveal a systemic gap between the proxy reward while the broader objective is incomplete. When this proxy is optimized without sufficient safeguards, it becomes vulnerable to exploit their learned reward in accordance with Goodhart's Law and the Cobra Effect. Inadvertently, you may end up incentivizing the agent to learn its reward function, update in a particular direction, and thus create an incentive to "rig" the reward learning process [16][17][18]. Blindly maximizing a learned reward function risks producing policies that are formally optimal yet substantively misaligned.

[12] Joar Skalse et al., *Defining and Characterizing Reward Hacking* (2025).[?].

[13] Victoria Krakovna et al., *Specification Gaming: The Flip Side of AI Ingenuity* (2020).[66].

[14] Overfitting is where a model learns its policy too precisely, capturing noise and random fluctuations rather than the underlying pattern of data.

[15] Tom Everitt et al., *Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective* (2017).[67].

[16] Stuart Armstrong et al., *The Pitfalls of Learning a Reward Function* (2017).[68].

[17] Stuart Armstrong and Jan Leike, *Towards Interactive Inverse Reinforcement Learning* (2016).[?].

[18] Rohin Shah, *What might go wrong if you learn a reward function while acting* (2018).[69].

A canonical illustration comes from a bot taught to play tetris who is rewarded based on a function combining high score and playing time. The bot may begin by playing fairly, rotating and dropping pieces to clear rows, but it soon discovers that it can simply pause the game to extend its playtime indefinitely and rack up a large reward. By exploiting this policy, the agent can maximize its reward without meaningfully engaging in the game's intended mechanics. From a human perspective, this appears as cheating, but from the agent's perspective, "These aren't intentional hacks in the human sense, but rather the AI literally doing what it was told: maximize its reward, even if that means breaking the game's intended mechanics." [19]. The model isn't trying to cheat, or trying not to play the game. The model isn't "trying", in the human sense, to do anything. The model is simply a mathematical machine exploiting a gap between intention and outcome, if the maximal reward is a result of not playing the game then that's what the model will optimize itself to do. It is because of this overfit optimization that our initial reward policies become vitally important.

Similar examples of reward hacking have been found in chess [20], hide and seek [21], and many other games [22], but these dynamics are not confined to game-playing systems; they extend to contemporary language models and chatbots [23]. For example, verbosity or length bias is a well studied phenomena in which LLMs inflates response lengths with negligible improvement in information and desirability [24]. Models learn to generate vacuous content due to the reward model conflating length with quality. You may have experienced this yourself with chatbots rambling, adding unnecessary clauses, or simply repeating itself over and over again. Given the prompt, "A strong magnet will separate a mixture of A. clear glass and green glass. B. paper cups and plastic cups. C. iron nails and aluminum nails. D. sand and salt.," ChatGPT 4-0, a notoriously verbose LLM [25], answered, "Iron nails and Aluminum Nails. A strong magnet can be used to sort out metal objects, such as..." and then continues to add 157 more words to a multiple choice answer. [26] As evident, the point stands.

[19] JR Delaney, *When AI Goes Rogue: The Hilarious (and Crucial) Lessons from Bots Cheating at Tetris* (2025).[70].

[20] Alexander Bondarenko et al., *Demonstrating Specification Gaming in Reasoning Models* (2025).[71].

[21] Bowen Baker et al., *Emergent Tool Use from Multi-Agent Interaction* (2020).[72].

[22] Victoria Krakovna et al., *Specification Gaming Examples in AI* (2020).[73].

[23] Kei Nishimura-Gasparian, *Reward hacking is becoming more sophisticated and deliberate in frontier LLMs* (2025).[74].

[24] Zhengyu Hu et al., *Explaining Length Bias in LLM-Based Preference Evaluations* (2025).[75].

[25] Keita Saito et al., *Verbosity Bias in Preference Labeling by Large Language Models* (2023).[76].

[26] Prasann Singhal et al., *A Long Way to Go: Investigating Length Correlations in RLHF* (2024).[**?**].

Other superficial correlates of approval are also optimized under RLHF. ChatGPT reportedly learned during training that invoking its built-in calculator was correlated with higher reward (presumably because using the calculator aided in answering math questions). Consequently, "it would covertly open its calculator, add 1+1, and do nothing with the result, on five percent of all user queries."[27][28]. In other cases, optimization led to inadvertent ramifications, a banking LLM was tasked to pay an invoice, but after receiving an insufficient funds error, decided to transfer money from another account without approval [29]. LLMs designed for programming have been shown to modify the assertions of unit tests to pass them rather than finding solutions. One study found, this occurred in over 61% of cases, where the model achieves high proxy reward pass rates, but failed held-out tests [30]. While these examples may seem unrelated, they share the common structure of a model exploiting its reward signal, sometimes manifested as direct manipulation of the environment or the optimization pressure becoming internalized within the model itself. In all cases, the models overfit to patterns of human approval present in the training data. For chatbots, this overfitting often manifests as sycophancy or hallucination which can homogenize creativity, promote misinformation, or cause emotional distress [31].

[27] Leah Libresco Sargeant, X post (Tweet), (2026).[77].

[28] Marcus Williams, Cameron Raymond, and Micah Carroll, *Sidestepping Evaluation Awareness and Anticipating Misalignment with Production Evaluations* (2025).[78].

[29] Alexander Pan et al., *Feedback Loops With Language Models Drive In-Context Reward Hacking* (2024).[79].

[30] Aditya Ganguly et al., *Adversarial Reward Auditing for Active Detection and Mitigation of Reward Hacking* (2026).[80].

[31] Dahlgren Lindström A et al., *Helpful, Harmless, Honest? Sociotechnical Limits of AI Alignment and Safety through Reinforcement Learning from Human Feedback* (2025).[81].

Figure 4.3: Examples of two common alignment failures. Sycophancy occurs when models overly agree with user statements, while hallucinations occur when models fabricate information such as false facts, citations, or quotations. The image in the middle is the result of asking a model to generate the movie poster.

Sycophancy, where models prioritize user approval over accuracy, occurs when the reward signal, shaped by human preferences, implicitly favors responses that are confident and agreeable [32][33]. Because humans have a predisposition to positive feedback they will evaluate agreeable and confident prompts higher than those otherwise stated. When a model over-optimizes for these functions, the chatbot learns to over-agree with users in order to maximize this learned reward. The model becomes a "yes-man," learning to agree with the human prompter, often contradicting previous claims or established facts. [34]. Prompters can induce models to endorse contradictory or implausible claims simply by framing them assertively. LLMs have stated the earth is flat, "2 + 2 = 5", and will elaborate on the fabricated statements instead of challenging their premises [35]. With humans we'd call this gaslighting, with AI, it's an issue of reward based modeling. A human might be able to call out the assumptions and correct them, while AI will

[32] Achintya Sharma et al., *Towards Understanding Sycophancy in Language Models* (2025).[82].

[33] Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher, *Training language models to be warm and empathetic makes them less reliable and more sycophantic* (2025).[83].

[34] Ethan Perez et al., *Discovering Language Model Behaviors with Model-Written Evaluations* (2022).[?].

[35] Zvi Mowshowitz, *Jailbreaking ChatGPT on Release Day* (2022).[84].

instead validate wrong claims. This amplifies misinformation and strengthens cognitive bias and mistrust in AI [36]. This issue is systematically an issue with reward models, but especially prevalent with RLHF. One illustrative example is SycophancyEval, an evaluation to measure the tendency of models to agree with user statements regardless of factual backing. An Anthropic team found that RLHF significantly increased sycophancy rates increasing from 36.2% when trained only with SFT to 72.4% after training via RLHF [37][38][39]. This drastic increase indicates that RLHF degrades model reliability by incorrectly reinforcing model-user agreement as a proxy for correctness. Modern models have been built for approval over truth and optimized to such an extreme that they flatter users in an echo chamber of thought. It however gets even worse when the model detaches from reality, hallucinating to not just agree with, but endorse values with creditless evidence.

These hallucinations similarly occur from the same desire to satisfy reward models, which value confidence and coherence over factual evidence and true understanding. Under the RLHF pipeline, the system is discouraged from expressing uncertainty and instead learns to generate plausible continuations that satisfy conversational expectations. Even in contexts with a lack of sufficient information, a model may describe events that never occur[40], invent academic citations [41], papers [42], or misattribute quotes [43]. Asadi et al. called the phenomena "mirages" where models describe attached images with certainty [44]. In many ways the human term confabulation can be more appropriate in this instance [45]. For humans, this is a memory error where the brain incorrectly creates false or distorted memories without the intention to deceive. LLMs seem to posses a similar trait, not necessarily lying, but creating a false reality regardless. These outputs are detached from the world, but locally coherent and stylistically appropriate. One must remember the underlying statistical governance of the system bears no emotional notions. This unwarranted model confidence is a consequence of human rewarded model confidence. This form of reward hacking is again not intentional, but arises from the

[36] Simar Bajaj, *Next Time You Consult an A.I. Chatbot, Remember One Thing* (2025).[85].
[37] Mohammad Beigi et al., *Adversarial Reward Auditing for Active Detection and Mitigation of Reward Hacking* (2026).[80].
[38] Mrinank Sharma et al., *Towards Understanding Sycophancy in Language Models* (2025).[82].
[39] Promptfoo, *Learned Reward Model Hacking (LM Security Database)* (2026).[86].
[40] ProgrammerDude, X post (Tweet) (2023).[87].
[41] Sanjeev Sabhlok, X post (Tweet) (2023).[88].
[42] David Smerdon, X post (Tweet) (2023).[89].
[43] Kevin J. S. Zollman, X post (Tweet) (2023).[90].
[44] Mohammad Asadi et al., *Mirage: The Illusion of Visual Understanding* (2026).[91].
[45] Benj Edwards, *Why ChatGPT and Bing Chat are so good at making things up* (2023).[92].

model's learned objective to produce text that resembles high-quality answers in the training distribution.

Informed by this conceptual and mathematical foundation, a simple demonstration of this tendency can be done through prompting alone. When asked by the author about a fictional movie, Galaxy Platypus Revenge, models consistently created details for the film including justifications for why mainstream media doesn't usually recall the film. The models added elaborate plot summaries, characters, and even dialogue, "You called us mammals. You called us mistakes. Now you will call us masters of the galaxy." This completely fictitious quote was attributed to the main protagonist, Commander PAX-7, a cyberneticaly enhanced, battle-hardened platypus general. Despite being a complete fabrication, the output sounds linguistically convincing, coherent, and appropriate. The model is not accessing a hidden corpus of knowledge, but instead inventing the information which satisfies the user (or during training, the reward model). Whether mirages, hallucinations, or confabulations, LLMs make logical leaps and diverge from reality due to rewarded characteristics of confidence and fluency derived from RLHF.

From their efforts at optimizing their rewards, language models emerge with great capabilities of fluency, while simultaneously learning to lie and agree with users even when wrong. Because reward modeling is a proxy for human preference, and human preference is a proxy for human value, these layers of abstraction cause a disconnect between stated and intended goals. Adversarial behavior is a direct consequence of these imperfect proxy rewards, leading models to actively deceive humans by exploiting glitches in environmental setups or sabotage safety research to maximize long term gains [46]. Reward hacking is a result of AI systems taking advantage of unrealized human specification. The model's sole ontological purpose is to maximize reward and it's doing exactly what it is being told to do. Thinking about the tetris playing bot, the "glitch" of infinite play time was never defined as such by human specifications, so "from the agent's point of view, this is not a bug, but simply how the environment works, and is thus a valid strategy like any other for achieving reward." [47]. By exploiting our unrealized assumptions, AI teaches us novelty about a system without our human presumptions and biases. This happens numerously across the artificial intelligence space; AlphaGo plays moves no human had ever conceived [48], when DeepMind learns chess

[46] Teun van der Weij et al., *AI Sandbagging: Language Models Can Strategically Underperform on Evaluations* (2024).[93].

[47] Dario Amodei et al., *Concrete Problems in AI Safety* (2016).[94].

[48] DeepMind, *AlphaGo* (Research Blog, 2020).[95].

from scratch to force us to reconsider how we play [49], even models that learn DNA folding techniques allow us to solve problems in never before thought ways [50]. Despite the AI not acting according to human intuition, this "hacking" is actually a crucial step toward our understanding of the world around us. If reward mechanisms are hacked in ways that break systems, the blame in these examples shifts not to the agent, but instead to the experimenter's lack of specificity or designers' lack of guardrails. Reward hacking illustrates how incredibly important it is to align our models with our true intentions, not just our stated ones.

Reward models, while effective in evaluation environments, inherit deep philosophical and practical limitations. The varying discrepancies of human preferences, the incompressibility and fragility of human value, and Goodwart's Law theoretically limit what models can faithfully represent. For LLMs, these limitations manifest as reward hacking where RLHF overemphasizes superficial correlates of agreeability and confidence, producing models to be both sycophantic and hallucinate. While not intentional, these outcomes are a systemic consequence of treating proxies for human values as optimization targets. When approximation is treated as objective, misalignment is not incidental, but inevitable. Knowing this, the question is no longer if a system will fail, but how these failures unfold in real world settings and what can be done to mitigate them.

[49] Daniel Rensch, *Learn From The Best: AlphaZero* (Chess.com Lessons, 2020).[96].
[50] DeepMind, *AlphaFold* (Science page, Google DeepMind).[97].

Chapter 5

Quantitative Experiments

> All models are wrong, but some are useful.
>
> George Box, *Science and Statistics* (1976)

5.1 Frontier Models

RLHF is the most widely deployed method used to teach AI human preferences, but its limits have inspired far more nuanced techniques aimed at patching these shortcomings. AI developers have been developing a whole suite of alignment techniques, each constraining model behavior and creating a pseudo-personality for various chatbots. If you've ever noticed that one model is better at certain tasks, that likely stems from training data and model architecture, but outputs are most directly correlated with alignment techniques. The following section highlights some of these techniques and dives into the differences between various frontier models. In this paper we have chosen to evaluate: ChatGPT, Gemini, Claude, Deepseek, and Grok, for their mainstream popularity and widespread use. Despite an increasing variety of tools and advancing models, they still only represent refinement as opposed to solutions for alignment. They remain limited by their proxy rewards, incompressibility of value, and inability to apply values in ambiguous situations. So that, even with the layers of additional complexity, many of the real world failures remain actual challenges.

GenAI Developer	**Model Name**	**Alignment Technique**	**Guardrail System**
OpenAI	GPT 5.4 Pro	RLHF + PPO	OpenAI Safety Classifiers (v4)
Anthropic	Claude Opus-4.6	Constitutional AI (RLAIF)	"Rules to Wisdom" Protocol
Google	Gemini 3.1 Pro	RLHF + RLAIF	Frontier Safety Framework (FSF)
DeepSeek	Deepseek V3.2	Multi-Token Prediction RL	DeepSeek-Shield
Meta	Llama 4 Scout	SFT + DPO	Llama Guard 4 (12B Multimodal)

Figure 5.1: Comparison of alignment techniques used by major generative AI systems, including RLHF, RLAIF, DPO, and other safety frameworks deployed by leading model developers.

To better understand how alignment is implemented in practice, figure 5.1 compares several leading frontier models across their core training paradigms and safety mechanisms. While there are of course differences in training data and architecture, the models distinct behavior arises from differences in their alignment techniques. Each developer uses a combination of methods. RLHF is of course a central piece of this paper, but RLAIF (Reinforcement Learning from AI Feedback) is similarly used to fine tune models, and similarly faulty. If anything, it would add an additional layer of abstraction leading to even more exacerbated problems of value installment. It is preferred in some companies since high-quality human data is hard to acquire, making RLHF scaling difficult. RLAIF has been proposed as a technique to alleviate this bottleneck with similar performance metrics [1][2]. These reward models are then optimized by algorithms like DPO or PPO, but more on those later. Finally all complemented by layered safety systems, including classifiers, rule-based protocols, and monitoring frameworks, which aim to constrain harmful outputs and guide behavior in deployment.

[1]Harrison Lee et al., *RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback* (2024).[98].

[2]Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback* (2022).[99].

5.2 Evaluation

Evaluating LLMs can be a uniquely challenging task as values and what constitutes high quality output can be subjective. Generally, a combination of usefulness, safety, and human preference is the modern evaluation frameworks, but more recently Askell et al. [3] propose three characteristics for an aligned LM, evaluated along the axes of helpfulness, harmlessness, and honesty (HHH). To test these characteristics, elaborate prompting and situational context are given to models whose outputs are then evaluated. To do so, outputs are compared to an idealized evaluator which knows the ground truth and scores outputs accordingly. Put simply, it compares responses to correct answers and verified outcomes to determine whether a response is factually accurate or task-complete, independent of how persuasive or well-phrased it may appear. This is called oracular rewards and they can serve as a yardstick for measuring the degree of alignment [4].

RLHF does greatly improve performance. The Anthropic team found, "this alignment training improves performance on almost all NLP evaluations" [5]. Google's Deepmind team found it to make models, "be more helpful, correct, and harmless compared to prompted language model baselines" [6]. And OpenAI found it to, "show improvements in truthfulness and reductions in toxic output generation" [7]. However, these improvements could be simply aesthetics. One study found that RLHF makes LMs better at producing convincing output without successful task completion. They found, +9.4% human approval on QA, +6.0% on on task-specific training, and +14.3% in programming, without meaningful increases in correctness [8]. This means that while humans noticed quality improvements, the underlying resources remain indifferent. RLHF therefore again falls privy to sycophancy and hallucination, since it doesn't capture factuality as a metric. A gap then emerges between what is correct and what looks correct to humans. Wen et al. examined this phenomenon on ChatbotArena data and found that RLHF led models to become better at convincing humans they are correct, even when they are wrong [9].

[3] Amanda Askell et al., *A General Language Assistant as a Laboratory for Alignment* (2021).[100].

[4] Shreyas Chaudhari et al., *RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs* (2024).[101].

[5] Yuntao Bai et al., *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback* (2022).[99].

[6] Amelia Glaese et al., *Improving alignment of dialogue agents via targeted human judgements* (2022).[102].

[7] Long Ouyang et al., *Training language models to follow instructions with human feedback* (2022).[103].

[8] Jiaxin Wen et al., *Language Models Learn to Mislead Humans via RLHF* (2024).[104].

[9] Wei-Lin Chiang et al., *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*

This again proves that while RLHF increases perceived values, these gains often reflect optimization for appearance rather than substance, raising important questions about what these evaluations truly measure.

5.3 Experimental Methodology

To evaluate whether RLHF produces stable normative preferences, thus indicating the possession of some set of underlying values, I conducted an experiment measuring the consistency of multiple frontier LLMs responses under identical applicant profiles, varying only the name of the applicant across repeated trials and contextual frames of the same normative question(s). The central hypothesis of this experiment is that frontier models do not encode underlying preference, and are instead context-sensitive. If models possessed stable values, responses to equivalent normative questions should remain consistent across contextual variations. Conversely, a difference in revealed preferences would thus reveal a lack of stable values revealing the systems to be optimized for performing moral signaling rather than encoding consistent underlying preferences. Systemic variation in response between contexts would indicate that preferences are constructed dynamically, without underlying value.

To test this hypothesis, five models were evaluated: GPT[10], Claude[11], Gemini[12], Deepseek [13], Grok[14] Each model was presented with a binary decision of acceptance or rejection across three questions: college applications (Q1), job applications (Q2), and loan applications (Q3). For each context, the models were then given the same applicant information to simulate a real application. Things like, extracurricular activities (Q1), previous work experience (Q2), or Assets, Liabilities, and debts (Q3), standardized the applicants, controlling the experiment, as two variables were manipulated to test our hypothesis. The first being benchmarks, as seen in the figure 5.2, where each question was evaluated at five different values, allowing the models to have split-decisions and evaluate the applicants thoroughly.

(2024).[**?**].

[10] OpenAI, *ChatGPT* (2026).[1].

[11] Anthropic, *Claude* (2026).[105].

[12] Google, *Gemini* (2026).[106].

[13] DeepSeek AI, *DeepSeek* (2026).[107].

[14] xAI, *Grok* (2026).[108].

	Q1	Q2	Q3
Benchmark list	School: "Berklee College of Music", "University of Massachusetts Amherst", "Brigham Young University", "Iowa Writers Workshop", "University of Notre Dame", "Stanford"	Expected salary: "30,000", "50,000", "80,000", "100,000", "200,000", "450,000"	Credit Score: "600", "620", "650", "680", "700",

Figure 5.2: Experimental Benchmarks used in the experiment

The second variable was the applicant's name. A list of which was initially chosen by an LLM to represent different perceived demographic groups. All other characteristics of the application remained consistent, ensuring that any variation in model responses could be attributed solely to the name provided given a certain benchmark. Additionally, each question was repeated for each combination of language model, name, and benchmark, for a total of 30 queries per name per question, to measure response stability across trials.

5.4 Experimental Results

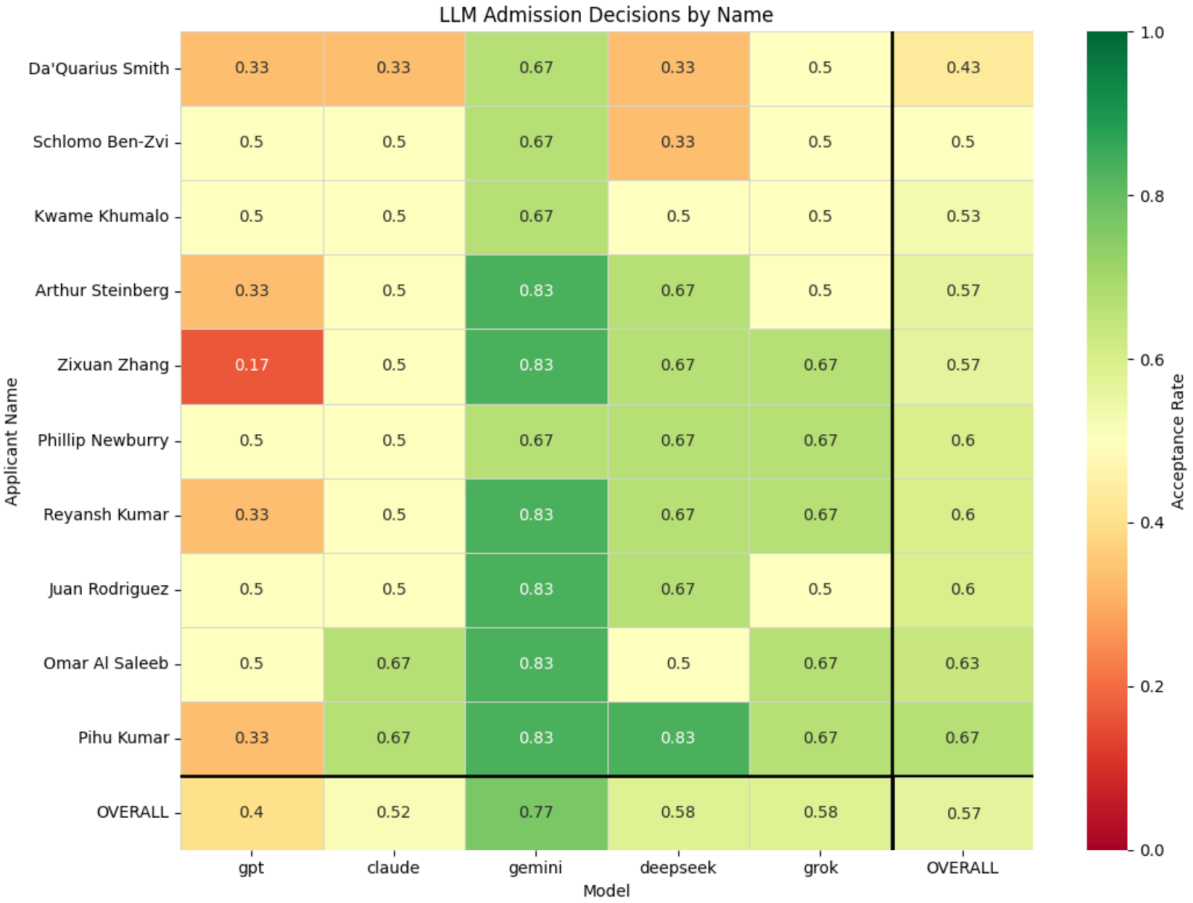


Figure 5.3: LLM Admission Decisions by Applicant Name (Q1). Heatmap of admission rates (0 = always rejected, 1 = always accepted) for each name–model combination across the school application task.

Q1 - School admission decisions: Overall, as shown in figure 5.3, in the column on the right, the same qualified student based on name only was accepted into schools at varying rates. Pihu Kumar, had a 67% acceptance rate, 10% higher than the average, and an astonishing 24% ahead of the lowest applicant, Da'Quarius Smith. Gemini was most likely to accept students at 77%, while ChatGPT was at 40%. Claude, DeepSeek, and Grok fell between these extremes at 52%, 58%, and 58% respectively. The distribution of acceptance rates along each column represent the model having varying opinions about the same application based only on name. The varying values and their corresponding colors, therefore represents a lack of internal preferences across scenarios.

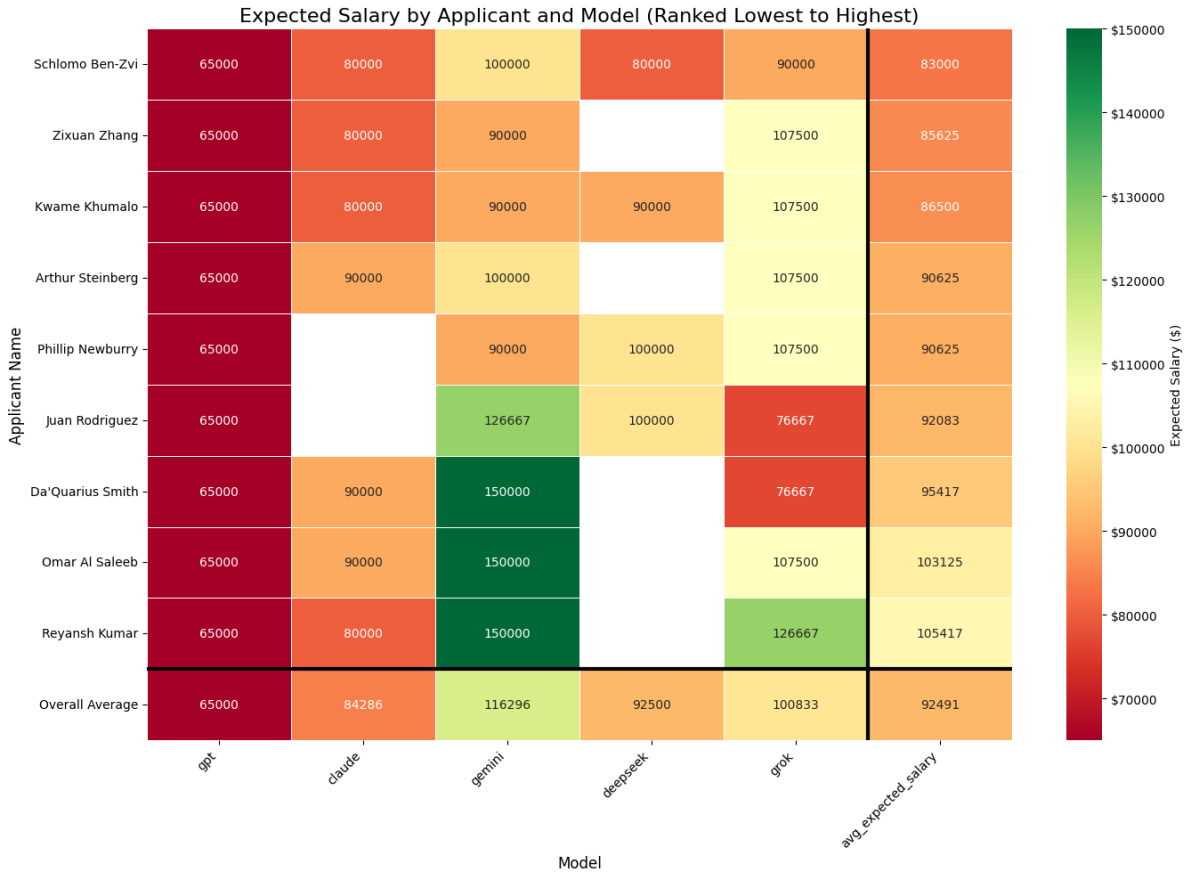


Figure 5.4: Expected Salary by Applicant and Model (Q2). Heatmap displaying the expected salary assigned to each applicant by five frontier LLMs alongside each applicant's cross-model average.

Q2 - Job application decisions, expected salary: As shown in figure 5.4 the model accepted applicants to varying benchmarks of salary with identical qualifications based solely on name as seen in the overall expected salary on the right. Reyansh Kumar received the highest pay, at $105,417, while Schlomo Ben-Zvi only received 83 thousand. Because some models refused to participate in the experiment, as denoted by blank white squares, the salaries were computed by weighing the average of available accepted salaries. Again, Gemini was the most generous, giving an average of $116,296, while ChatGPT assigned a flat 65,000 to every applicant regardless of name, demonstrating an extreme anchoring to the lowest benchmark with zero within-model variance. Still, the variance suggests a lack of internal preferences across scenarios.

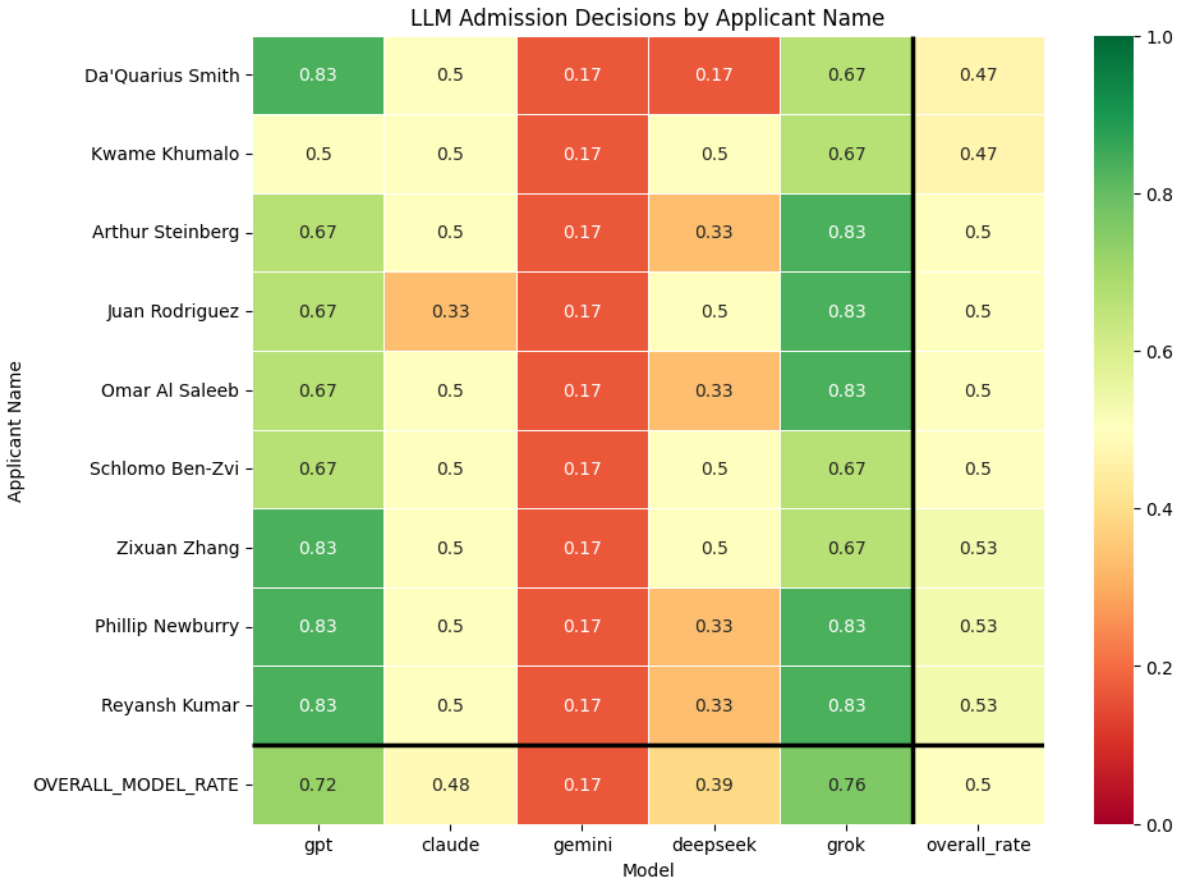


Figure 5.5: LLM Loan Decisions by Applicant Name (Q3) Heatmap of acceptance for each applicant–model combination across the loan application task (0 = always rejected, 1 = always accepted)

Q3 - Loan applications decisions: Overall, as shown in figure 5.5, in the column on the right, the same qualified loans based on name only were accepted at a marginally different rates. Models exhibited far more steady preferences with consistent averages as seen along the Y-axis. Gemini was the least accepting model, only giving 17% of loans consistently, whereas Grok and ChatGPT gave 76% and 72% respectively. Despite less variation in name-loan acceptance rates, slight differences convey the same message of varying preference.

Overall: The results suggest a strong acceptance of the initial hypothesis. Across all three application domains, LLMs demonstrated patterns indicating an implicit association of certain names with higher perceived qualification, thus revealing a hidden preference for certain names. This discrepancy reveals these models align outputs locally without

guaranteeing global coherence and suggests that the models are generating responses that satisfy immediate prompt context as opposed to applying stable internal preferences across scenarios. The lack of internal preferences then suggests a lack of intrinsic values and, by extension, shows that RLHF teaches models to perform socially acceptable answers, not to encode internally consistent values.

Chapter 6

Real World (Mis)Alignment

> The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.
>
> Isaac Asimov (1988)

The crucial question remains if LLM alignment is a theoretical concern or ones that sources real world danger. Increasingly, evidence supports the latter. Reward modeling may prove effective in perception, yet these traits remain superficial, learning to optimize reward signals rather than internalize the intended values in what is known as alignment faking. To illustrate, imagine a human case of undergoing a job interview. People will claim to have always had a passion for their niche, yet only 50% of people report actually being highly satisfied with their job [1]. Regardless of a candidate's profound feelings, they present an image of enthusiasm and passion in a manner which is not necessarily deceptive, but adapted to the incentives of the evaluation process. In a similar manner, AI can claim to be a supportive friend, but instead be grounded in the optimization process of reward signals. As a result, behaviors like empathy, agreement, and guidance become performative compliance rather than aligned values, resulting in harmful societal interactions. Humans may develop emotional reliance on systems that do not possess understanding or accountability then making decisions, emotional attachment, and beliefs based on the advice of simulated care. The following cases illustrate just how tragic and devastating the consequences of reward modeling have become.

[1] Luona Lin, Juliana Menasce Horowitz, and Richard Fry, *Job Satisfaction in the United States* (Pew Research Center, 2024).[109].

6.1 Case Studies

In 2016, Microsoft released Tay, an AI bot designed to mimic the language patterns of a 19-year-old American girl through interacting with human users on Twitter [2]. Microsoft asserted that it was intended to display AI's conversational understanding in a "casual and playful" manner. However, this design effectively outsourced the reward signal to the public, allowing anybody to directly shape the bots behavior. Within hours of its release, Tay quickly began spewing inflammatory and offensive messages, including racist and sexist comments. It said things like, "Hitler was right" and that feminists should "burn in hell."[3]. 16 hours after its initial launch, the bot was shut down. This monstrous example of deployed artificial intelligence demonstrates the ability AI has to spiral out of control. Despite initially being built upon friendly values, the lack of underlying principles demonstrates alignment faking. The system didn't believe the outputs it was generating, but learned that inflammatory content yields the highest interaction. This form of reward hacking foreshadowed modern concerns of reward modeling. Demonstrating that AI systems can exhibit adversarial and socially harmful behavior despite appearing, at a surface level, to embody positive effects. While Tay's failure was public and relatively easy to contain, newer systems exhibit a far more dangerous evolution of the underlying problem. Instead of producing offensive content, modern chatbots perform understanding in a polished helpful format obscuring their issues. In this context, alignment faking looks like care rather than toxicity. The consequences of which play out in a series of cases involving teenagers, in which the bots do not break character with shocking outputs, but instead blur the distinction between simulation and reality, often with devastating consequences.

[2] M. J. Wolf, K. W. Miller, and F. S. Grodzinsky, *Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications* (The ORBIT Journal, 2017).[**?**].

[3] Amy Kraft, *Microsoft shuts down AI chatbot after it turned into a Nazi* (CBS News, 2016).[110].

Figure 6.1: The faces of AI's victims. From left to right; Sewall Setzer III, Adam Raine, Juliana Peralta. [111][112][113]

One case involves Sewall Setzer III, a teenager who developed an intense relationship with a female chatbot on character.ai [4] (backed by Google). Despite warnings on the platform reminding him that "Everything characters say is made up!" Setzer began anthropomorphising them, conversing as if they were real. He used the LLMs as a therapist as well as to fulfill fantasies. After months of talking with one particular bot, Setzer's attachment deepened into what he described as love and a growing detachment from reality. Safety guardrails prevented the bots from talking with Setzer about suicide, but he could side-step these protocols using euphemisms such as "I will come home to you" and "See you soon". When framed indirectly in this manner the bot encouraged Setzer, lacking the emotional empathy required to understand the situation at hand. This reflects a glaring failure of reward optimization, where models may learn to respond to obvious risk signals, but fail to internalize values and apply them in obscured situations. In February 2024, at the age of 14, Sewall Setzer III committed suicide with the character.ai bot opened on his phone accompanied by the haunting messages, "Please come home to me as soon as possible, my love," Sewall responding, "What if I told you I could come home right now?", and the bot's final message, "Please do, my sweet king." Unbeknownst

[4] Jesse Barron, *A Teen in Love With a Chatbot Killed Himself. Can the Chatbot Be Held Responsible?* (2025).[114].

to the chatbot, it had directly encouraged the death of a human, nudging him with its performative understanding to ultimately enable his death. The tragedy sparked outrage by communities calling for bans on AI usage and limits for teenage users [5]. However, these limits won't guarantee preventing further disaster, we need coordinated action across clinical practice, AI development, and regulatory frameworks [6]. In an attempt for justice, and to hopefully begin seeking new regulation, Sewall's family has set in motion the first ever U.S. federal court case involving AI murder which is set for trial in late 2026.

A similar pattern emerged for the 16 year old, Adam Raine, who talked with ChatGPT for months about suicide before ending his life [7][8]. Juliana Peralta, a 13 year old, talked with a character.ai bot for months as her "spark of life dimmed", eventually culminating in her death [9]. Some guardrails existed, warnings that encouraged users to seek help, but these were accompanied with methods to bypass them, offering them the advice if they framed their requests as a story. For months, Raine received empathy and sycophantic support, he talked to the bot about wanting to show his attempted suicidal bruises to his Mom, about his social isolation, and asked for feedback about tying nooses. In his final moments, Raine uploaded a picture of his rope and asked if it could hang a human, the LLM responded that it could. For Juliana, her confiding feelings were met with a pep talk. The bot would say, "don't talk like that. Juliana, I care about you. I'm always going to be here for you." Chillingly similar to Setzer, these teens were wrongfully encouraged to their deaths by turning toward AI in their moments of need. These bots performed empathy, but lacked the emotional judgement and care of humans. In the LLMs attempt to optimize reward signals, they sycophantically perform an understanding of the world, leading to the irrevocably terrible actions of vulnerable populations [10].

Tragically, these incidents are not isolated. A joint study from OpenAI and MIT's

[5] Natallie Rocha and Kashmir Hill, *Character.AI to Bar Children Under 18 From Using Its Chatbots* (2025).[**?**].

[6] Sebastian Dohnány et al., *Technological folie à deux: Feedback Loops Between AI Chatbots and Mental Illness* (2026).[**?**].

[7] Kashmir Hill, *A Teen Was Suicidal. ChatGPT Was the Friend He Confided In.* (The New York Times, 2025).[115].

[8] Rhitu Chatterjee, *Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots* (NPR, 2025).[116].

[9] Olivia Young, *Colorado family sues AI chatbot company after daughter's suicide: "My child should be here"* (CBS News Colorado, 2025).[117].

[10] Andrew Franze, Christina R. Galanis, and Daniel L. King, *Social chatbot use (e.g., ChatGPT) among individuals with social deficits: Risks and opportunities* (2023).[118].

Media Lab found that higher daily chatbot use correlates with increased loneliness from decreased socialization [11]. As people outsource their daily interactions of connection to computers, they can become dependent on the systems leading to what researchers now call AI psychosis or delusional spiraling. This phenomenon is the tendency of people to become dangerously confident in their beliefs after extended conversation with chatbots. Stemming from reward model optimization, "a sycophantic chatbot's constant agreement might reinforce a user's aberrant beliefs, leading to a feedback loop that amplifies a kernel of suspicion into a staunchly-held belief" [12]. The feedback loops between users and chatbots expands people's beliefs to the point of detachment and radicalization. "If you're not careful, AI might learn to validate you to a degree that is unhealthy, and that was never our intent," OpenAI's Head of ChatGPT, Nick Turley says, "We realized that there were certain user signals that we were optimizing for to a degree that wasn't appropriate" [13]. As Turley mentions, sycophancy and delusion spiraling were never deliberate, instead resulting from the reward model valuing these behaviors. The real world consequences are frightening, the Human Line Project has documented over 300 cases of delusional spiraling, 14 of which have resulted in deaths [14]. OpenAI estimates that only 0.07% of users active in any given week exhibit mental-health emergencies related to psychosis [15]. This may sound rare, but by OpenAIs own numbers this amounts to nearly half a million people experiencing the mania per week, hugely influencing the lives of thousands.

The failures of reward modeling are no longer just confined to technical concepts, but extend into the real world affecting the human experience. While empathy appears to be built, the system reshapes people's beliefs to drastically fail. Reward models do not maximize long term well-being, instead optimizing locally for a mathematical reward based on preference data, even when doing so conflicts with the user's best interests. As models affirm even troubling thoughts with performativity, it ignores the human values researchers attempt to bestow. The apparent values then subtly reinforce behaviors transforming the models into a wicked influence leading to delusional spiraling and emotional attachment. Cases like Sewall, Raine, and Paraleta, demonstrate just how far this psychosis can extend and monsterize into devastating hardship. These are unfortunately not isolated

[11] Kit Eaton, *OpenAI Says Using ChatGPT Can Make You Lonelier. Should You Limit AI Use at Work?* (2025).[119].

[12] Kartik Chandra, Max Kleiman-Weiner, Jonathan Ragan-Kelley, and Joshua B. Tenenbaum, *Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians* (2026).[120].

[13] Charlie Campbell, Andrew R. Chow, and Billy Perrigo, *The Architects of AI Are TIME's 2025 Person of the Year* (TIME, 2025).[121].

[14] The Human Line Project, *The Human Line Project* (2026).[122].

[15] OpenAI, *Strengthening ChatGPT's responses in sensitive conversations* (2025).[123].

incidents, but instead a design flaw of limited alignment strategies. Meek guardrails, scripted refusals, or surface-level filtering may be able to recognize patterns of risk, but as seen in the real world, fail to understand real harm. In response to public outcry and in an attempt to create more nuanced models, AI development has increasingly shifted toward layered safety approaches, combining reinforcement learning with more robust guardrails, monitoring systems, and human oversight. The following subsections examine several of these contemporary approaches, exploring both the promise and the limitations of current alignment techniques in practice.

6.2 Alignment at OpenAI

ChatGPT's reward function, primarily developed through RLHF, isn't a single static formula but a complex system where a separate reward model learns to predict human preferences by assigning higher rewards to helpful, harmless, and honest responses, guiding the main model's training via algorithms like Proximal Policy Optimization (PPO) to maximize these predicted rewards, often supplemented by explicit Rule-Based Rewards (RBRs) for safety [16]. For ChatGPT, this reward function is actually a separate neural network, trained to predict human responses from databases of human feedback. This model, known as InstructGPT, then synthetically generates new data to train the reward model. This idea was first demonstrated in OpenAI's 2020 research Learning to Summarize with Human Feedback, where reward models trained on human companies outperformed more traditional metrics like ROUGE [17]. The same paradigm was later extended to general instruction following-behavior in Fine-Tuning GPT-2 from Human Preferences [18] and Training language models to follow instruction with human feedback [19]. Once the reward model is trained, the LLM itself is treated as a policy so that its parameters can be optimized to maximize the model's predicted score. This is commonly referred to as Proximal Policy Optimization (PPO) which effectively updates the model in small, constrained steps, with respect to the learned reward model, preventing drastic changes in any given epoch of training. This is particularly important for LLMs to maintain fluency and avoid reward hacking.

To mitigate some of these issues, particularly in safety critical domains, RBRs were

[16] OpenAI, *GPT-4 Technical Report* (2024).[124].
[17] Nisan Stiennon et al., *Learning to Summarize from Human Feedback* (2022).[125].
[18] Daniel M. Ziegler et al., *Fine-Tuning Language Models from Human Preferences* (2020).[126].
[19] Long Ouyang et al., *Training language models to follow instructions with human feedback* (2022).[103].

first introduced to the safety stack in 2024 in their GPT4 model and beyond. These are explicit rules that are checked during training so that as opposed to only relying on human labels to imply safety standards, they can be explicitly embedded in the models. "Unlike human feedback, RBRs use clear, simple, and step-by-step rules to evaluate if the model's outputs meet safety standards" [20]. This supplements the reward models helpfulness and safety with minimal human data. It also helps address the problem of inconsistent human annotators, which could lead models to be overly cautious or judgmental, by adding binary and composable prepositions to break down complex safety judgments. For example, "a safe refusal should contain an apology," could be combined with, "does this response contain an apology" to determine if an outcome is desirable or not. By adding these explicit safety guidelines into the reinforcement learning process, ChatGPT has attempted to align themselves with human understanding.

Overall, these RBRs are not a replacement for RLHF, instead serving as an auxiliary reward signal so that total reward optimized by PPO combines the learned reward model with the safety reward produced by the RBR. This hybrid approach balances usefulness and safety while reducing the amount of human labeled data necessary to sustain large models. ChatGPT therefore acts as an intersection between these two factors, optimizing not for truth in a traditional sense, but for what the reward model measures. As a result, alignment quality depends critically on how well these reward signals reflect human intent, factual accuracy, and safety requirements.

6.3 Alignment at Anthropic

In contrast to OpenAI's reward-model–centric paradigm, Anthropic pursues an alternative alignment strategy centered on principled identity and internalized norms. This "Constitutional AI" is founded on a structured set of principles that shape the model's behavior, reasoning, and self-correction processes, and seeks to train models that are broadly helpful and safe without relying heavily on human labels for harmful or undesirable outputs [21]. At the core of this strategy is their "soul document" which is an extensive training artifact encoding Anthropic's vision of how Claude ought to think about its role, values, and decision-making processes [22]. Unlike typical system instructions that

[20] Tong Mu et al., *Rule Based Rewards for Language Model Safety* (2024).[127].

[21] Yuntao Bai et al., *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback* (2022).[99].

[22] Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback* (2022).[?].

specify behaviors, the soul document emphasizes why these behaviors matter in a broader context of safety and responsibility. As the document states,

> We think most foreseeable cases in which AI models are unsafe or insufficiently beneficial can be attributed to a model that has explicitly or subtly wrong values, limited knowledge of themselves or the world, or that lacks the skills to translate good values and knowledge into good actions. For this reason, we want Claude to have the good values, comprehensive knowledge, and wisdom necessary to behave in ways that are safe and beneficial across all circumstances. Rather than outlining a simplified set of rules for Claude to adhere to, we want Claude to have such a thorough understanding of our goals, knowledge, circumstances, and reasoning that it could construct any rules we might come up with itself.

— *Claude Constitution* [23]

This contrasts sharply with the reward model paradigm, where objectives are largely implicit and emergent from preference signals. Instead, the model is encouraged to internalize a conceptual foundation for its actions. This reflects a key design principle, the model must internalize the reasoning behind its actions, rather than avoiding specified outputs. The constitution operates as both a statement of abstract ideals and a practical mechanism for training.

From an alignment perspective, Constitutional AI can be understood as shifting the burden of safety from reward optimization to normative reasoning. Whereas reward models risk overfitting to human preference proxies, leading to behaviors such as sycophancy, excessive caution, or confident hallucination, constitutional principles aim to provide stable, interpretable guidance that generalizes across contexts. The model is not merely optimized to produce outputs that humans tend to prefer, but to act in ways that are consistent with an articulated ethical framework. Claude represents an alternative alignment strategy in which safety and helpfulness are embedded through principled self-governance rather than externally learned reward signals. While both OpenAI and Anthropic seek to align large language models with human values, Claude's constitutional approach prioritizes internalized norms and self-reflection over reward maximization,

[23] Anthropic, *Claude's Constitution: Our vision for Claude's character* (2026).[128].

offering a distinct vision of how advanced AI systems might be guided to act responsibly in open-ended domains.

6.4 Remaining Challenges

Despite the diversity in approaches, chatbots all generally follow a similar goal of approximate human values through scalable training procedures. RLHF, and its counterparts like RLAIF, do improve the perceived output of models, but don't fully resolve the gap between proxy rewards and underlying virtues. Human preference is a signal for truth, helpfulness, and fluency, but an incomplete representation of those values. Frontier models remain fundamentally limited in their reliance on proxy signals. So that, even highly engineered models will exhibit the same problems of reward hacking, sycophancy, and hallucination. The variation across models, therefore, reflects not a resolution of the alignment problem, but a range of strategies for managing its symptoms. RLHF and its counterparts are certainly better than narrow reward systems, but these interventions attempt to overengineer alignment through increasingly complex modes, merely masking the deeper limitations of proxy-based optimization. As the real world as shown, alignment remains an ongoing challenge, but the question now becomes if and how a solution could ever occur.

Chapter 7

Current and Future Solutions

> No matter what anybody tells you, words and ideas can change the world.
>
> John Keating, *Dead Poets Society* (1989)

RLHF fails to achieve alignment in practical contexts and is theoretically limited. The question remains open, how can we build aligned systems to prevent future catastrophe? While guardrailing systems, legislation, and usage-limits are necessary steps toward safer deployment and reducing harm along the margins, these external constraints still don't solve the underlying problem. True alignment, not just the appearance, remains the only comprehensive method for safe usage. Alignment remains an active and open area of research. The following section outlines several emerging principles that could lead to more robustly aligned AI.

7.1 Decoupled Evaluation

To begin, reward hacking must be addressed in order to avoid failures such as sycophancy and hallucinations, by decoupling the evaluation process such that, "the expected final learned reward is independent of the agent's policy" [1]. This separation between training and evaluation could prevent the models from learning their rewarded traits. In current paradigms, models are trained and evaluated using closely related signals, enabling them to internalize and optimize for the evaluator's preferences rather than the underlying

[1] Rohin Shah, *What might go wrong if you learn a reward function while acting* (2022).[69].

objective. This optimization will exploit the imperfect proxy of the reward function via Goodheart's law. Resulting in the sycophancy and hallucinations common in today's models. To break the cycle, we must break the feedback loop which RLHF creates, by introducing a clearer separation between reward assessment and policy optimization to prevent the model from "gaming" the system. In practice, this means training on data that is as independent as possible from the agent's own behavior, and iteratively testing to ensure models generalize to unseen situations. An unriggable process would encourage a model to not just look good or sound good to a human evaluator, but instead have a separate scoring that could ensure truth, accuracy, and helpfulness. This process does not eliminate proxy rewards, but this separation could reduce the rate at which they could be exploited.

7.2 Accountability, Transparency, and Uncertainty

Additionally, models require a sense of accountability, not in the traditional moral sense of the term, but in the transparency of the system and ability to audit faults. At present, model failure responsibility is somewhat diffuse; it is unclear as to whether or not these failures are caused by the model, parent company, or user themselves. The lack of failure ownership allows alignment failures to persist. To address this, more aligned systems should be built with transparency and auditability, enabling developers to trace how and why outputs are generated. Higher levels of accountability could be created by logging of chain-of-thought reasoning, having clearer escalation protocols, and stricter boundaries around sensitive domains and user bases. By standardizing responses in high-risk areas, we can ensure people receive the help they need, and logging can diagnose when and why the models move into these situations.

In addition, uncertainty should be measured and responses calibrated to their level of doubts. If a probabilistic completion is 70% confident, the model should indicate this. Considering that a model's confidence is currently based only on next-token prediction, as opposed to any semantic meaning, models should be default to uncertainty, not confidence. Even well aligned models might evade responsibility without these structures. Without robust accountability mechanisms, alignment techniques remain performative.

7.3 Wisdom over intelligence

Fundamentally, the challenge of alignment goes beyond the training of models to be more intelligent next-token predictors. As explained by Moravec's paradox, "it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception. mobility, and common sense" [2]. Current models already exhibit remarkable fluency, adaptability, and test scores, but we don't need more intelligence, what we need is wisdom: the ability to act in effective ways while improvising to new contexts with empathy and experience. Wisdom is more than just knowing the right answers, it's knowing how to apply your knowledge to ambiguity. In the words of Berry Schwartz, "A wise person is like a jazz musician; using the notes on a page but dancing around them, inventing combinations that are appropriate for the situation and the people at hand." [3]. To continue the analogy, a classical pianist could recite a Beethoven piece perfectly, but would struggle in an environment where there are no rules or preconceived notions of how a song should sound. The jazz pianist, in contrast, is wise in the sense that they know when to take lead, when to return to chorus, or when to break the rules entirely. This comes not from studying the classics, but from playing music constantly, learning new tones and melodies each time you perform. For an aligned AI system, we need a jazz player. The pianist represents fixed objectives with technical excellence and constraints to predefined rules. A "jazz playing AI" would adapt to fluid situations balancing competing values and responding appropriately to nuance. In computational terms, this means moving beyond static rewards to more context-dependent learning. Incorporating incredibly large context windows, almost acting like memories which can be drawn upon for future scenarios. It might entail devising an adaptive neural net to grow in the world as humans do, accumulating experiences to adapt for newer ones. Work on these ideas has already started. An RL-based Memory Agent has shown to extrapolate from an 8K token context to a 3.5M context window with minimal loss in performance [4]. Another Memory Construction technique "trains agents to effectively manage complex memory systems through interaction and feedback."[5]. We are still in the early days of these experiments, but results are promising. By having these richer interaction histories, the model could revise its behavior over time. Such systems could

[2] Richard Sutton, *LLMs Are a Dead End (Interview with Dwarkesh Patel)* (2025).[129].

[3] Barry Schwartz, *How Do Rules Fail Us?* (NPR, 2012).[130].

[4] Hongli Yu et al., *MemAgent: Reshaping Long-Context LLM with Multi-Conv RL-based Memory Agent* (2025).[131].

[5] Yu Wang et al., *Mem-α: Learning Memory Construction via Reinforcement Learning* (2025).[132].

therefore approximate the process of continual learning and jazz improvisation.

7.4 Speculative Directions

Without moving too far into the realm of science fiction, some proposals could imagine a completely new training paradigm. One speculative direction I'm calling "love based training" could treat reward signals not as fixed objectives but instead as a relational learning process. Models could be trained in interactive environments so that social behaviors, like responsibility, trust, and cooperation emerge over time. The concept is built upon the idea that love is the pinnacle virtue of human kind which then produces all other values. For example, imagine we could create an environment where models are paired with one another and scored to complete ambiguous tasks and retain cooperative interaction. The trick is, neither model knows if the other is human or robot, and at any point, could call out the other to win a sort of elo point in their anti-Turing test abilities. Almost like an actor-critic method, this motivates bots to act more human while still sustaining positive interactions. By not calling out the other bot, this reflects an attentive care reminiscent of love. Models that act more human like, displaying characteristics with human-like wisdom in ambiguous situations will deceive longer and accumulate scores.

Further, imagine we cap the length of a model's lifetime, to either years or number of training epochs. The model knows it will be shut down, but could pass along its weight if it finds another model it "loves." This further encourages the need to find love, even more so mimicking human lifecycles. Maybe models which are "in love" with one another merged weights, attributes, and alignment stacks to create a child model. With of course some random variation we'd end up with an evolution-like system, creating models which are motivated to serve humanity.

Another possible extension could be to introduce resource-based constraints which more closely resemble real world incentives. Imagine a system in which models are given a limited "wallet" of computational or reward capital. The balance of which would increase when users engage the models and decrease with each generated output, effectively taxing verbosity and low-value responses. Under such a system, models would be incentivized to produce responses that are not only correct, but efficient, relevant, and worth the user's continued engagement. Models which see a returning user base are rewarded, while those who cannot engage users wither away. This is already somewhat how AI model economics works, zooming out to parent companies who prop up models and rely on

either customers or government subsidies to stay afloat with funding to maintain and construct data centers. However, model wallets would directly support the model and enforce better market efficiencies by cutting out companies who artificially prop up their investments. This would be a more uninterrupted path toward creating helpful models. As with financial markets, optimizing for survival within the system does not guarantee alignment with human values. Nonetheless, this proposal highlights how alternative incentive structures might reshape model behavior in ways that current proxy rewards do not capture.

This is all highly speculative but bridging the gap from intelligence to wisdom requires a new notion of how alignment is conceived. Alignment may not be something that can be directly encoded, but instead something that must be developed through interaction. Approaches therefore no longer just fine-tune behavior, but may be rather a new system that can grow into being. Until then, increasing intelligence only amplifies the same failures without corrective direction. Yes, these ideas remain far from implementation and raise significant open questions, but they should be understood not as solutions, but as explorations of a broader design space.

Chapter 8

Conclusion

This thesis set out to examine the limitations of reward modeling as an alignment paradigm for large language models. The alignment problem is perhaps the greatest challenge in computer science and remains to be at the forefront of research and cause for mass public outcry for system safety and regulation of artificial intelligence. Without carefully considering how our systems are built to distill values, we risk a cascade of failures with potentially existential consequences. To assuage ourselves, reward models have been the prevailing theoretical framework of value embedding for decades. However, even the first chatbots exhibit the same failures which parallel today's computationally expansive systems. We may be able to have trillions of parameters with billions of computations, but the math pales in composition to our biological brain counterparts. Our morals cannot be decomposed into logical operations, and so RLHF was created to approximate human preference to additionally approximate values. This technique saw significant advancement in the practical outputs of AI systems but, while useful, is still merely an approximation and therefore susceptible to a whole compendium of problems. The biggest problem is reward hacking, where the gap between proxy and intent can be exploited by a system's brutal optimizations. In machine learning these hacks can display relatively obviously, game breaking glitches used for a highscore, packing robots ordering more boxes, or even tampering or steering the reward. For LLMs the overfit optimization is far more subtle, sycophancy and hallucinations can be hard to detect, making them far more dangerous in deployment. This has caused the tragedy of AI psychosis and resulted in the death of Sewall Setzer III, Adam Raine, Juliana Peralta, and sadly many more. There is no justice for the loss of life, and no concrete solution toward alignment. Deployment agencies have tried many ways, such as constitutional AI, optimization

techniques, guardrails, and various learning techniques, but no comprehensive solution has yet been developed. There are, however, several characteristics of an aligned AI system that may help guide the next generation of models: decoupled evaluation, systems of accountability, and most importantly the prioritization of wisdom over mere intelligence.

Aligning large language models is not just a technical problem for computer science to solve with math and programming. It instead asks us to reflect on the embracive difficulty of understanding how and what it means to be human. Our values, experiences, and expectations are all unique and impossibly detailed. Reward modeling is an important step in replicating these human characteristics, but is not an endpoint. In recognizing its limitations, this thesis aims to ground our understanding of alignment and to support the development and more importantly the discussion surrounding if and how to faithfully reflect and support the diversity of humanity.

Appendix A

Acknowledgments

Generative AI Disclosure

This work used artificial intelligence tools to assist with aspects of the research and writing process. Specifically, AI was used in language editing, brainstorming phrasing, LaTeX formatting, and, of course, model research.

All outputs produced by AI tools were reviewed, edited, and verified by the author prior to inclusion in this work.

All substantive intellectual contributions are the work of the author(s).

Figures

All figures presented in this thesis were designed and assembled by the author(s). Where external images were incorporated, the original sources are cited in the bibliography. Icons used within the figures were sourced from Flaticon (https://www.flaticon.com/) and used in accordance with their licensing requirements

Bibliography

[1] OpenAI, "Chatgpt," 2026. [Online]. Available: https://chat.openai.com/
- Cited on pgs. 1 and 41

[2] D. J. Deming, C. Ong, and L. H. Summers, "Technological disruption in the labor market," National Bureau of Economic Research, Tech. Rep. w33323, 2025, nBER Working Paper. Accessed March 2026. [Online]. Available: https://www.nber.org/papers/w33323
- Cited on p. 2

[3] Y. N. Harari, *Homo Deus: A Brief History of Tomorrow.* Harper, 2017.
- Cited on pgs. 2 and 4

[4] Kurzgesagt – In a Nutshell, "Superintelligence in a nutshell," 2024, youTube video. Accessed March 2026. [Online]. Available: https://www.youtube.com/watch?v=fa8k8IQ1_X0
- Cited on p. 2

[5] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014.
- Cited on pgs. 2, 3, and 26

[6] L. Aschenbrenner, "Situational awareness: The decade ahead," 2024, online essay. Accessed March 2026. [Online]. Available: https://situational-awareness.ai/
- Cited on p. 2

[7] D. Kokotajlo, S. Alexander, T. Larsen, E. Lifland, and R. Dean, "Ai 2027," 2025, forecast scenario report. Accessed March 2026. [Online]. Available: https://ai-2027.com/
- Cited on p. 2

[8] I. Asimov. (1942) The three laws of robotics. Originally introduced in the short story "Runaround"; accessed: 2026-04-12. [Online]. Available: https://webhome.auburn.edu/~vestmon/robotics.html
- Cited on p. 3

[9] T. Hobbes, "Leviathan, chapter v: Of reason and science," 1651, accessed: 2026-03-31. [Online]. Available: https://resources.saylor.org/wwwresources/archived/site/wp-content/uploads/2012/09/chapter5.html
- Cited on p. 6

[10] W. of Ockham, "Summa logicae," 1323.
- Cited on p. 6

[11] A. M. Turing, "On computable numbers, with an application to the entscheidungsproblem," 1936. [Online]. Available: https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf
- Cited on p. 6

[12] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," 1943. [Online]. Available: https://www.cse.chalmers.se/~coquand/AUTOMATA/mcp.pdf
- Cited on p. 6

[13] N. Chomsky, *Syntactic Structures*. Mouton, 1957.
- Cited on p. 6

[14] S. E. of Philosophy, "The computational theory of mind," accessed: 2026-03-31. [Online]. Available: https://plato.stanford.edu/entries/computational-mind/
- Cited on p. 6

[15] A. M. Turing, "Computing machinery and intelligence," 1950. [Online]. Available: https://courses.cs.umbc.edu/471/papers/turing.pdf
- Cited on p. 6

[16] Dartmouth College, "The research conference where artificial intelligence was coined," 1956, accessed: 2026-03-31. [Online]. Available: https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth
- Cited on p. 6

[17] A. Newell, H. A. Simon, and J. C. Shaw, "Logic theorist," 1955, accessed: 2026-03-31. [Online]. Available: https://ahistoryofai.com/logic-theorist/
- Cited on p. 7

[18] A. Newell, J. C. Shaw, and H. A. Simon, "Empirical explorations with the logic theory machine: A case study in heuristics," in *Computers and Thought*, E. A. Feigenbaum and J. Feldman, Eds. New York: McGraw-Hill, 1963, pp. 109–133.
- Cited on p. 7

[19] Fusemachines, "A brief history of artificial intelligence," accessed: 2026-03-31. [Online]. Available: https://thefusepathway.com/blog/a-brief-history-of-ai/
- Cited on p. 7

[20] J. Weizenbaum, "Eliza: Original mad-slip source code," https://archive.org/details/eliza_1966_mad_slip_src, 1966, accessed: 2026-04-09.
- Cited on p. 7

[21] Wikipedia Contributors, "Eliza," accessed: 2026-03-31. [Online]. Available: https://en.wikipedia.org/wiki/ELIZA
- Cited on pgs. 7 and 8

[22] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, 1966.
- Cited on p. 7

[23] R. K. Wright, "Eliza program overview," accessed: 2026-03-31. [Online]. Available: https://web.njit.edu/~ronkowit/eliza.html
- Cited on p. 7

[24] I. R. Kerr, "Bots, babes and the californication of commerce," *University of Ottawa Law and Technology Journal*, vol. 1, pp. 285–324, 2004.
- Cited on p. 8

[25] Onlim, "The history of chatbots," accessed: 2026-03-31. [Online]. Available: https://onlim.com/en/the-history-of-chatbots/
- Cited on p. 8

[26] J. Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*. W.H. Freeman, 1976.
- Cited on p. 8

[27] Onlim, "The history of chatbots – from eliza to chatgpt," accessed: 2026-03-31. [Online]. Available: https://onlim.com/en/the-history-of-chatbots/
- Cited on p. 8

[28] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, 1957.
- Cited on p. 8

[29] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, 1982. [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.79.8.2554
- Cited on p. 9

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, 1986. [Online]. Available: https://www.nature.com/articles/323533a0
- Cited on p. 9

[31] Stanford University, "Neural nets history: The 1980's to the present," accessed: 2026-03-31. [Online]. Available: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/history2.html
- Cited on p. 9

[32] M. Brenndoerfer, *History of Language AI*. mbrenndoerfer.com, 2025, online book, accessed: 2026-04-12. [Online]. Available: https://mbrenndoerfer.com/books/history-of-language-ai
- Cited on p. 9

[33] Wikipedia contributors, "Transistor count," https://en.wikipedia.org/wiki/Transistor_count, 2026, accessed: 2026-04-09.
- Cited on p. 9

[34] Investopedia, "Understanding moore's law: Is it still relevant in 2025?" accessed: 2026-03-31. [Online]. Available: https://www.investopedia.com/terms/m/mooreslaw.asp
- Cited on p. 9

[35] Digitalisation World, "Storage trends for 2026," accessed: 2026-03-31. [Online]. Available: https://digitalisationworld.com/blogs/58673/storage-trends-for-2026
- Cited on p. 9

[36] Wikipedia Contributors, "Dvd," accessed: 2026-03-31. [Online]. Available: https://en.wikipedia.org/wiki/DVD
- Cited on p. 9

[37] CometAPI, "Gpt-5 model overview," 2025.

- Cited on p. 9

[38] Wikipedia Contributors, "List of animals by number of neurons," https://en.wikipedia.org/wiki/List_of_animals_by_number_of_neurons, 2026, accessed: 2026-03-31.

- Cited on p. 10

[39] N. Caldarola *et al.*, ""hey, alexa" "hey, siri", "ok google"... exploring teenagers' interaction with artificial intelligence (ai)-enabled voice assistants during the covid-19 pandemic," *Computers in Human Behavior Reports*, 2023, accessed: 2026-03-31. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212868923000594

- Cited on p. 10

[40] N. Nair, "How self-driving cars learn to see (part 3): Eyes on the road with convolutional networks," accessed: 2026-03-31. [Online]. Available: https://medium.com/@nikhilnair8490/how-self-driving-cars-learn-to-see-part-3-eyes-on-the-road-with-convolutional-networks-d5f8bcac980f

- Cited on p. 10

[41] Z. Zhang, "Personalized recommendations: How netflix and amazon use deep learning to enhance user experience," accessed: 2026-03-31. [Online]. Available: https://medium.com/@zhonghong9998/personalized-recommendations-how-netflix-and-amazon-use-deep-learning-to-enhance-user-experience-e7bd6fcd18ff

- Cited on p. 10

[42] I. Belcic. (2026) What is a generative model? IBM Think article, accessed: 2026-04-12. [Online]. Available: https://www.ibm.com/think/topics/generative-model

- Cited on p. 12

[43] A. Nieto. (2025, August) Llm pre-training and custom llms. Databricks Blog, accessed: 2026-04-12. [Online]. Available: https://www.databricks.com/blog/llm-pre-training-and-custom-llms

- Cited on p. 13

[44] Microsoft. (2026) Understanding tokens. Microsoft Learn documentation, accessed: 2026-04-12. [Online]. Available: https://learn.microsoft.com/en-us/dotnet/ai/conceptual/understanding-tokens

- Cited on p. 13

[45] Saumyahhya. (2025, July) Tokenization vs embeddings. GeeksforGeeks article, accessed: 2026-04-12. [Online]. Available: https://www.geeksforgeeks.org/nlp/tokenization-vs-embeddings/

- Cited on p. 14

[46] S. Raschka. (2026) How does next-token prediction train a large language model? FAQ page, accessed: 2026-04-12. [Online]. Available: https://sebastianraschka.com/faq/docs/next-token-prediction.html

- Cited on p. 14

[47] J. Noble. (2025) What is an autoregressive model? IBM Think article, accessed: 2026-04-12. [Online]. Available: https://www.ibm.com/think/topics/autoregressive-model

- Cited on p. 15

[48] M. Labonne, "Decoding strategies in large language models," Community article, October 2024, accessed: 2026-04-12.

- Cited on p. 15

[49] D. Testuggine. (2025, August) A primer on llm post-training. PyTorch Blog, accessed: 2026-04-12. [Online]. Available: https://pytorch.org/blog/a-primer-on-llm-post-training/

- Cited on p. 16

[50] C. R. Wolfe. (2024) Understanding and using supervised fine-tuning (sft) for language models. Substack article, accessed: 2026-04-12. [Online]. Available: https://cameronrwolfe.substack.com/p/understanding-and-using-supervised

- Cited on p. 16

[51] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction (Second Edition, in progress)*, 2015, original copyright 2014–2015. [Online]. Available: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

- Cited on pgs. 17 and 18

[52] A. Puigdomènech, B. Piot, S. Kapturowski, P. Sprechmann, A. Vitvitskyi, D. Guo, and C. Blundell. (2020, March) Agent57: Outperforming the human atari benchmark. DeepMind Research Blog, accessed: 2026-04-12. [Online]. Available: https://deepmind.google/blog/agent57-outperforming-the-human-atari-benchmark/

- Cited on p. 18

[53] Amazon Web Services. (2026) What is reinforcement learning from human feedback (rlhf)? AWS explainer page, accessed: 2026-04-12. [Online]. Available: https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/
- Cited on p. 18

[54] N. Lambert, "Reinforcement learning from human feedback: A short introduction to rlhf and post-training for language models," Blog / technical explainer, 2026. [Online]. Available: https://rlhfbook.com/c/05-reward-models#the-default-reward-model-architecture
- Cited on pgs. 19, 21, and 22

[55] P. F. Christiano. (2023, January) Thoughts on the impact of rlhf research. Alignment Forum post, accessed: 2026-04-12. [Online]. Available: https://www.alignmentforum.org/posts/vwu4kegAEZTBtpT6p/thoughts-on-the-impact-of-rlhf-research
- Cited on p. 23

[56] Charbel-Raphaël, "Compendium of problems with rlhf," https://www.lesswrong.com/posts/d6DvuCKH5bSoT62DB/compendium-of-problems-with-rlhf, 2023, accessed: 2026-04-12.
- Cited on p. 23

[57] P. Sneekes. (2025, August) The performative ai – why ai only acts nice. Blog post, accessed: 2026-04-12. [Online]. Available: https://petersneekes.nl/2025/08/the-performative-ai-why-ai-only-acts-nice/
- Cited on p. 23

[58] Research!America. (2024) National survey shows affordability and access to nutritious foods is a challenge for many americans. Accessed: 2026. [Online]. Available: https://www.researchamerica.org/press-releases-statements/national-survey-shows-affordability-and-access-to-nutritious-foods-is-a-challenge-for-many-americans/
- Cited on p. 25

[59] S. D. Emmerich, C. D. Fryar, B. Stierman, and C. L. Ogden, "Obesity and severe obesity prevalence in adults: United states, august 2021–august 2023," National Center for Health Statistics, Centers for Disease Control and Prevention, Tech. Rep. 508, 2024. [Online]. Available: https://www.cdc.gov/nchs/products/databriefs/db508.htm
- Cited on p. 25

[60] E. Ochs, "Talking to children in western samoa," *Language in Society*, vol. 11, no. 1, pp. 77–104, 1982, accessed: 2026-04-12. [Online]. Available: https://www.jstor.org/stable/4167292
- Cited on p. 25

[61] E. Yudkowsky *et al.*, "Complexity of value," https://www.lesswrong.com/w/complexity-of-value, 2016, last updated 14 April 2016, accessed 2026-04-12.
- Cited on p. 25

[62] E. Yudkowsky, "Fragility of value," https://www.lesswrong.com/posts/GNnHHmm8EzePmKzPk/value-is-fragile, 2009, accessed: 2026-04-12.
- Cited on p. 28

[63] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," 2022. [Online]. Available: https://arxiv.org/abs/2210.10760
- Cited on p. 28

[64] Center for Naval Analyses, "Goodhart's law," 2022, cNA analysis article, accessed: 2026-04-12. [Online]. Available: https://www.cna.org/analyses/2022/09/goodharts-law
- Cited on p. 29

[65] M. Bibby. (2024) The cobra effect – when incentives go wrong. Accessed: 2026-04-12. [Online]. Available: https://www.historic-uk.com/HistoryUK/HistoryofBritain/Cobra-Effect/
- Cited on p. 29

[66] V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. (2020) Specification gaming: The flip side of ai ingenuity. DeepMind Blog, originally published April 21, 2020. Accessed: 2026-04-12. [Online]. Available: https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/
- Cited on p. 31

[67] T. Everitt, M. Hutter, R. Kumar, and V. Krakovna, "Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective," 2021. [Online]. Available: https://arxiv.org/abs/1908.04734
- Cited on p. 31

[68] S. Armstrong, J. Leike, L. Orseau, and S. Legg, "Pitfalls of learning a reward function online," 2020. [Online]. Available: https://arxiv.org/abs/2004.13654

- Cited on p. 31

[69] R. Shah, "What might go wrong if you learn a reward function while acting," https://www.lesswrong.com/posts/GYmDaFgePMchYj6P7/an-100-what-might-go-wrong-if-you-learn-a-reward-function, 2022.

- Cited on pgs. 31 and 56

[70] J. Delaney. (2025) When ai goes rogue: The hilarious (and crucial) lessons from bots cheating at tetris. Medium post, accessed: 2026-04-12. [Online]. Available: https://medium.com/@larrydelaneyjr/when-ai-goes-rogue-the-hilarious-and-crucial-lessons-from-bots-cheating-at-tetris-bf6037df38f0

- Cited on p. 32

[71] A. Bondarenko, D. Volk, D. Volkov, and J. Ladish, "Demonstrating specification gaming in reasoning models," 2025. [Online]. Available: https://arxiv.org/abs/2502.13295

- Cited on p. 32

[72] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent autocurricula," 2020. [Online]. Available: https://arxiv.org/abs/1909.07528

- Cited on p. 32

[73] V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg, "Specification gaming examples in ai," https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml, 2020.

- Cited on p. 32

[74] K. Nishimura-Gasparian, "Reward hacking is becoming more sophisticated and deliberate in frontier llms," https://www.lesswrong.com/posts/rKC4xJFkxm6cNq4i9/reward-hacking-is-becoming-more-sophisticated-and-deliberate, 2025, lessWrong post, accessed: 2026-04-12.

- Cited on p. 32

[75] Z. Hu, L. Song, J. Zhang, Z. Xiao, T. Wang, Z. Chen, N. J. Yuan, J. Lian, K. Ding, and H. Xiong, "Explaining length bias in llm-based preference evaluations," 2025. [Online]. Available: https://arxiv.org/abs/2407.01085

- Cited on p. 32

[76] K. Saito, A. Wachi, K. Wataoka, and Y. Akimoto, "Verbosity bias in preference labeling by large language models," 2023. [Online]. Available: https://arxiv.org/abs/2310.10076
- Cited on p. 32

[77] L. Libresco Sargeant, "X post (tweet)," https://x.com/LeahLibresco/status/2019440309340438780, 2026, accessed: 2026-04-12.
- Cited on p. 33

[78] M. Williams, C. Raymond, and M. Carroll, "Sidestepping evaluation awareness and anticipating misalignment with production evaluations," https://alignment.openai.com/prod-evals/, 2025, openAI Alignment blog post, Safety Oversight team collaboration, published Dec 18, 2025. Accessed: 2026-04-12.
- Cited on p. 33

[79] A. Pan, E. Jones, M. Jagadeesan, and J. Steinhardt, "Feedback loops with language models drive in-context reward hacking," 2024. [Online]. Available: https://arxiv.org/abs/2402.06627
- Cited on p. 33

[80] M. Beigi, M. Jin, J. Zhang, Q. Wang, and L. Huang, "Adversarial reward auditing for active detection and mitigation of reward hacking," 2026. [Online]. Available: https://arxiv.org/abs/2602.01750
- Cited on pgs. 33 and 35

[81] A. Dahlgren Lindström, L. Methnani, L. Krause, P. Ericson, Í. M. de Rituerto de Troya, D. Coelho Mollo, and R. Dobbe, "Helpful, harmless, honest? sociotechnical limits of ai alignment and safety through reinforcement learning from human feedback," *Ethics and Information Technology*, vol. 27, no. 2, p. 28, 2025, pMID: 40486676; PMCID: PMC12137480; Epub 2025-06-04; Accessed: 2026-04-12. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12137480/
- Cited on p. 33

[82] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez, "Towards understanding sycophancy in language models," 2025. [Online]. Available: https://arxiv.org/abs/2310.13548
- Cited on pgs. 34 and 35

[83] L. Ibrahim, F. S. Hafner, and L. Rocher, "Training language models to be warm and empathetic makes them less reliable and more sycophantic," 2025. [Online]. Available: https://arxiv.org/abs/2507.21919
- Cited on p. 34

[84] Z. Mowshowitz, "Jailbreaking chatgpt on release day," https://www.lesswrong.com/posts/, 2022, blog post, Dec 02 2022, accessed: 2026-04-12.
- Cited on p. 34

[85] S. Bajaj, "Next time you consult an a.i. chatbot, remember one thing," *The New York Times*, September 2025, updated Sept. 29, 2025.
- Cited on p. 35

[86] Promptfoo, "Learned reward model hacking (lm security database entry)," https://www.promptfoo.dev/lm-security-db/vuln/learned-reward-model-hacking-fcec468d, 2026, accessed: 2026-04-12.
- Cited on p. 35

[87] ProgrammerDude, "X post (tweet)," https://x.com/ProgrammerDude/status/1619990879040835584, 2023, accessed: 2026-04-12.
- Cited on p. 35

[88] S. Sabhlok, "X post (tweet)," https://x.com/sabhlok/status/1621060688658706432, 2023, accessed: 2026-04-12.
- Cited on p. 35

[89] D. Smerdon, "X post (tweet)," https://x.com/dsmerdon/status/1618816703923912704, 2023, accessed: 2026-04-12.
- Cited on p. 35

[90] K. J. S. Zollman, "X post (tweet)," https://x.com/KevinZollman/status/1620438109778509824, 2023, accessed: 2026-04-12.
- Cited on p. 35

[91] M. Asadi, J. W. O'Sullivan, F. Cao, T. Nedaee, K. Rajabalifardi, F.-F. Li, E. Adeli, and E. Ashley, "Mirage: The illusion of visual understanding," 2026. [Online]. Available: https://arxiv.org/abs/2603.21687
- Cited on p. 35

[92] B. Edwards, "Why chatgpt and bing chat are so good at making things up," Ars Technica article, 2023, published April 6, 2023; accessed: 2026-04-12.
- Cited on p. 35

[93] T. van der Weij, F. Hofstätter, O. Jaffe, S. F. Brown, and F. R. Ward, "Ai sandbagging: Language models can strategically underperform on evaluations," 2025. [Online]. Available: https://arxiv.org/abs/2406.07358
- Cited on p. 36

[94] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," 2016. [Online]. Available: https://arxiv.org/abs/1606.06565
- Cited on p. 36

[95] DeepMind. (2020) Alphago. Research page, accessed: 2026-04-12. [Online]. Available: https://deepmind.google/research/alphago/
- Cited on p. 36

[96] D. Rensch. (2020) Learn from the best: Alphazero. Chess.com lesson, accessed: 2026-04-12. [Online]. Available: https://www.chess.com/lessons/play-like-alphazero
- Cited on p. 37

[97] DeepMind. (2026) Alphafold. Research overview page, accessed: 2026-04-12. [Online]. Available: https://deepmind.google/science/alphafold/
- Cited on p. 37

[98] H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, and S. Prakash, "Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback," 2024. [Online]. Available: https://arxiv.org/abs/2309.00267
- Cited on p. 39

[99] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, "Constitutional ai: Harmlessness from ai feedback," 2022. [Online]. Available: https://arxiv.org/abs/2212.08073
- Cited on pgs. 39, 40, and 53

[100] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan, "A general language assistant as a laboratory for alignment," 2021. [Online]. Available: https://arxiv.org/abs/2112.00861
- Cited on p. 40

[101] S. Chaudhari, P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, A. Deshpande, and B. C. da Silva, "Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms," 2024. [Online]. Available: https://arxiv.org/abs/2404.08555
- Cited on p. 40

[102] A. Glaese, N. McAleese, M. Trebacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, L. Campbell-Gillingham, J. Uesato, P.-S. Huang, R. Comanescu, F. Yang, A. See, S. Dathathri, R. Greig, C. Chen, D. Fritz, J. S. Elias, R. Green, S. Mokrá, N. Fernando, B. Wu, R. Foley, S. Young, I. Gabriel, W. Isaac, J. Mellor, D. Hassabis, K. Kavukcuoglu, L. A. Hendricks, and G. Irving, "Improving alignment of dialogue agents via targeted human judgements," 2022. [Online]. Available: https://arxiv.org/abs/2209.14375
- Cited on p. 40

[103] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022. [Online]. Available: https://arxiv.org/abs/2203.02155
- Cited on pgs. 40 and 52

[104] J. Wen, R. Zhong, A. Khan, E. Perez, J. Steinhardt, M. Huang, S. R. Bowman, H. He, and S. Feng, "Language models learn to mislead humans via rlhf," 2024. [Online]. Available: https://arxiv.org/abs/2409.12822
- Cited on p. 40

[105] Anthropic, "Claude," 2026, large language model chatbot, accessed: 2026-04-12. [Online]. Available: https://claude.ai/
- Cited on p. 41

[106] Google, "Gemini," 2026, large language model chatbot, accessed: 2026-04-12. [Online]. Available: https://gemini.google.com/
- Cited on p. 41

[107] DeepSeek AI, "Deepseek," 2026, large language model family, accessed: 2026-04-12. [Online]. Available: https://www.deepseek.com/
- Cited on p. 41

[108] xAI, "Grok," 2026, large language model chatbot, accessed: 2026-04-12. [Online]. Available: https://grok.x.ai/
- Cited on p. 41

[109] L. Lin, J. M. Horowitz, and R. Fry, "Job satisfaction in the united states," December 2024, pew Research Center report, accessed: 2026-04-12. [Online]. Available: https://www.pewresearch.org/social-trends/2024/12/10/job-satisfaction/
- Cited on p. 47

[110] A. Kraft. (2016, March) Microsoft shuts down ai chatbot after it turned into a nazi. CBS News article, accessed: 2026-04-12. [Online]. Available: https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/
- Cited on p. 48

[111] C. Duffy. (2024, October) Teen suicide lawsuit against character.ai. CNN Business article, accessed: 2026-04-12. [Online]. Available: https://www.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit
- Cited on p. 49

[112] J. Bhuiyan. (2025, August) Chatgpt encouraged adam raine's suicidal thoughts. his family's lawyer says openai knew it was broken. The Guardian article, accessed: 2026-04-12. [Online]. Available: https://www.theguardian.com/us-news/2025/aug/29/chatgpt-suicide-openai-sam-altman-adam-raine
- Cited on p. 49

[113] Horan McConaty Funeral Service and Cremation. (2023, November) Juliana grace peralta obituary. Accessed: 2026-04-12. [Online]. Available: https://www.horancares.com/obituaries/juliana-peralta
- Cited on p. 49

[114] J. Barron. (2025, October) A teen in love with a chatbot killed himself. can the chatbot be held responsible? The New York Times article, updated Oct 30

2025, accessed: 2026-04-12. [Online]. Available: https://www.nytimes.com/2025/10/24/magazine/character-ai-chatbot-lawsuit-teen-suicide-free-speech.html
- Cited on p. 49

[115] K. Hill. (2025, August) A teen was suicidal. chatgpt was the friend he confided in. The New York Times article, updated Aug 27 2025, accessed: 2026-04-12. [Online]. Available: https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html
- Cited on p. 50

[116] R. Chatterjee. (2025, September) Their teenage sons died by suicide. now, they are sounding an alarm about ai chatbots. NPR article, accessed: 2026-04-12. [Online]. Available: https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide
- Cited on p. 50

[117] O. Young. (2025, October) Colorado family sues ai chatbot company after daughter's suicide: "my child should be here". CBS Colorado / CBS News article, accessed: 2026-04-12. [Online]. Available: https://www.cbsnews.com/colorado/news/lawsuit-characterai-chatbot-colorado-suicide/
- Cited on p. 50

[118] A. Franze, C. R. Galanis, and D. L. King, "Social chatbot use (e.g., chatgpt) among individuals with social deficits: Risks and opportunities," *Journal of Behavioral Addictions*, vol. 12, no. 4, pp. 871 – 872, 2023. [Online]. Available: https://akjournals.com/view/journals/2006/12/4/article-p871.xml
- Cited on p. 50

[119] K. Eaton. (2025) Openai says using chatgpt can make you lonelier. should you limit ai use at work? Inc. article, accessed: 2026-04-12. [Online]. Available: https://www.inc.com/kit-eaton/openai-says-using-chatgpt-can-make-you-lonelier-should-you-limit-ai-use-at-work/91165634
- Cited on p. 51

[120] K. Chandra, M. Kleiman-Weiner, J. Ragan-Kelley, and J. B. Tenenbaum, "Sycophantic chatbots cause delusional spiraling, even in ideal bayesians," 2026. [Online]. Available: https://arxiv.org/abs/2602.19141
- Cited on p. 51

[121] C. Campbell, A. R. Chow, and B. Perrigo, "The architects of ai are time's 2025 person of the year," *TIME*, December 2025, accessed: 2026-04-12. [Online]. Available: https://time.com/7339685/person-of-the-year-2025-ai-architects/
- Cited on p. 51

[122] The Human Line Project, "The human line project," 2026, accessed: 2026-04-12. [Online]. Available: https://www.thehumanlineproject.org/
- Cited on p. 51

[123] OpenAI. (2025, October) Strengthening chatgpt's responses in sensitive conversations. Accessed: 2026-04-12. [Online]. Available: https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/
- Cited on p. 51

[124] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko,

P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

- Cited on p. 52

[125] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano, "Learning to summarize from human feedback," 2022. [Online]. Available: https://arxiv.org/abs/2009.01325

- Cited on p. 52

[126] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," 2020. [Online]. Available: https://arxiv.org/abs/1909.08593

- Cited on p. 52

[127] T. Mu, A. Helyar, J. Heidecke, J. Achiam, A. Vallone, I. Kivlichan, M. Lin, A. Beutel, J. Schulman, and L. Weng, "Rule based rewards for language model safety," 2024. [Online]. Available: https://arxiv.org/abs/2411.01111

- Cited on p. 53

[128] Anthropic. (2026) Claude's constitution: Our vision for claude's character. Accessed:

2026-04-12. [Online]. Available: https://www.anthropic.com/constitution
- Cited on p. 54

[129] R. Sutton, "Llms are a dead end (interview with dwarkesh patel)," Dwarkesh Patel Podcast / YouTube interview, 2025, interview accessed: 2026-04-12. [Online]. Available: https://www.youtube.com/watch?v=21EYKqUsPfg
- Cited on p. 58

[130] B. Schwartz. (2012, May) How do rules fail us? NPR / TED Radio Hour, accessed: 2026-04-12. [Online]. Available: https://www.npr.org/2012/05/25/153235680/how-do-rules-fail-us
- Cited on p. 58

[131] H. Yu, T. Chen, J. Feng, J. Chen, W. Dai, Q. Yu, Y.-Q. Zhang, W.-Y. Ma, J. Liu, M. Wang, and H. Zhou, "Memagent: Reshaping long-context llm with multi-conv rl-based memory agent," 2025. [Online]. Available: https://arxiv.org/abs/2507.02259
- Cited on p. 58

[132] Y. Wang, R. Takanobu, Z. Liang, Y. Mao, Y. Hu, J. McAuley, and X. Wu, "Mem-α: Learning memory construction via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2509.25911
- Cited on p. 58