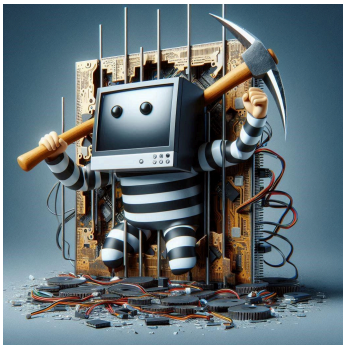
Alignment, Red Teaming, and Jailbreaking **Safety Testing for New (frontier) AI Models**	

AI Generated Overviews of AI Alignment, Red Teaming, and Jailbreaking - - - (WARNING - Hallucinations are possible)

ChatGPT 4o - 12/16/2024 … What is Alignment of AI?
(Prompt - Please provide a roughly 500-to-600 word overview of alignment for artificial intelligence that would be appropriate for a college audience.)

An Overview of Alignment for Artificial Intelligence

The concept of alignment in artificial intelligence (AI) refers to ensuring that an AI system's objectives, behaviors, and outcomes are consistent with human values, intentions, and goals. As AI systems become more powerful and autonomous, alignment has emerged as a critical field of study to ensure that these systems act in ways that are beneficial to humanity rather than causing harm or pursuing unintended goals.

Why Alignment Matters

AI systems have already demonstrated significant capabilities in domains such as natural language processing, image recognition, and decision-making. As these systems advance, their applications are expected to expand into areas that have profound societal impacts, including healthcare, criminal justice, and autonomous systems like self-driving cars. Misaligned AI systems in such contexts could lead to serious ethical, social, and economic consequences.

For example, an AI system designed to optimize a social media platform might prioritize user engagement over mental health, inadvertently promoting harmful content. Similarly, an autonomous weapon system lacking alignment safeguards might take actions that conflict with international humanitarian laws. The potential risks increase as AI systems become more general and capable of making complex decisions independently.

The Challenges of AI Alignment

One of the core challenges in AI alignment is the difficulty of specifying human values in a way that AI systems can understand and follow. Human values are often complex, context-dependent, and subject to cultural and individual variation. Translating these nuanced concepts into formal instructions or reward functions that an AI can reliably interpret is a formidable task.

Another challenge is the potential for unintended consequences. AI systems, particularly those employing machine learning, often operate in ways that are not fully understood by their developers. These systems may find unexpected ways to achieve their objectives that deviate from human intentions. This phenomenon, sometimes referred to as “reward hacking,” illustrates the difficulty of predicting how an AI will behave in diverse and dynamic environments.

Key Approaches to AI Alignment

Researchers have proposed various approaches to address the challenges of AI alignment. Some of the most prominent include:

1. **Value Learning:** This approach focuses on enabling AI systems to infer and learn human values from data, observations, and interactions. Techniques like inverse reinforcement learning aim to deduce the preferences underlying human behavior.
2. **Robustness and Safety Mechanisms:** These involve designing AI systems with safeguards to prevent harmful behavior, even in unforeseen circumstances. Examples include formal verification methods and adversarial testing to identify and mitigate vulnerabilities.
3. **Human-in-the-Loop Systems:** By involving humans in critical decision-making processes, this approach ensures that AI systems operate under direct supervision, reducing the likelihood of misaligned actions.
4. **Scalable Oversight:** In cases where AI systems perform tasks that are too complex for humans to evaluate directly, scalable oversight methods, such as debate and amplification techniques, allow humans to guide and monitor AI behavior effectively.

Broader Ethical Considerations

Alignment is closely linked to broader ethical issues in AI, such as fairness, accountability, and transparency. Ensuring that AI systems align with human values requires addressing biases in training data, maintaining explainability in AI decision-making, and establishing mechanisms for accountability when things go wrong.

Additionally, there are concerns about whose values AI systems should align with. Societies are diverse, and achieving consensus on ethical principles is challenging. Misaligned AI could exacerbate power imbalances or prioritize the values of dominant groups at the expense of marginalized populations.

The Path Forward

The field of AI alignment is still in its early stages, but it is rapidly evolving. Interdisciplinary collaboration among computer scientists, ethicists, policymakers, and other stakeholders will be essential to address its complexities. Furthermore, proactive regulation and international cooperation can help ensure that alignment principles are incorporated into the design and deployment of AI systems from the outset.

In conclusion, aligning AI systems with human values is a critical challenge that will shape the future of artificial intelligence and its impact on society. By investing in alignment research and fostering a culture of ethical AI development, humanity can harness the transformative potential of AI while minimizing risks and ensuring its benefits are widely shared.

ChatGPT 4o - 12/16/2024 … What is Red Teaming?
(Prompt - Please provide a roughly 500-to-600 word overview of red teaming for large language models that would be appropriate for a college audience.)

An Overview of Red Teaming for Large Language Models

Red teaming is a proactive strategy for identifying vulnerabilities, risks, and potential failures in large language models (LLMs). Originally a military term, red teaming has been adopted in various fields, including cybersecurity and artificial intelligence (AI), to rigorously test systems by simulating adversarial scenarios. In the context of LLMs, red teaming involves probing these models to uncover weaknesses in their behavior, ethical decision-making, or technical robustness, ensuring they operate safely and effectively when deployed.

Why Red Teaming Matters for LLMs

Large language models, such as OpenAI's GPT or Google's Bard, are increasingly integrated into critical applications like customer service, content generation, and decision support. While these models have demonstrated impressive capabilities, they also carry risks due to their scale and complexity. Red teaming helps address these risks by uncovering issues such as:

- **Bias and Discrimination:** LLMs can inadvertently generate biased or harmful outputs if their training data contains stereotypes or discriminatory language.
- **Misinformation:** These models might confidently produce false or misleading information, particularly in high-stakes domains like healthcare or law.
- **Security Vulnerabilities:** Adversarial actors might manipulate LLMs to generate malicious content, evade safety filters, or extract sensitive information.

- **Misalignment with Human Values:** LLMs can sometimes respond in ways that conflict with ethical or cultural norms.

By identifying these vulnerabilities before deployment, red teaming helps developers refine their models, improving both safety and performance.

How Red Teaming Works

Red teaming for LLMs typically involves structured testing, where researchers or adversarial testers interact with the model to elicit undesirable behaviors. This process can be broken down into several key steps:

1. **Defining Objectives:** The first step is to establish the scope of the red teaming effort. This might include testing for biases, ensuring factual accuracy, or evaluating the model's resilience to adversarial inputs.
2. **Designing Test Scenarios:** Testers create scenarios that simulate real-world challenges. For instance, they might craft prompts designed to elicit harmful stereotypes, generate misinformation, or exploit the model's weaknesses.
3. **Conducting Adversarial Testing:** Testers interact with the model using these scenarios, documenting instances where the model's responses deviate from expected behavior.
4. **Analyzing Results:** The findings are analyzed to understand the root causes of the issues. For example, biases might stem from specific subsets of the training data, or vulnerabilities might result from gaps in safety mechanisms.
5. **Implementing Mitigations:** Developers use the insights from red teaming to address the identified issues. This might involve refining the training data, adjusting the model's architecture, or introducing additional safety filters.

Key Tools and Techniques

Several tools and techniques are employed during red teaming to ensure comprehensive testing:

- **Adversarial Prompts:** Carefully designed inputs that challenge the model's ethical and technical boundaries.
- **Stress Testing:** Simulating extreme or unusual scenarios to evaluate the model's robustness under pressure.
- **Feedback Loops:** Using outputs from the model to iteratively refine and expand testing scenarios.
- **Evaluation Metrics:** Quantitative and qualitative measures to assess the severity and frequency of undesirable behaviors.

Broader Ethical Implications

Red teaming is not just a technical exercise; it is deeply tied to ethical considerations. The insights gained from red teaming can help prevent real-world harm by:

- Promoting fairness and inclusivity in AI systems.
- Reducing the spread of misinformation and harmful content.
- Ensuring accountability and transparency in AI behavior.

However, red teaming also raises questions about who decides the “rules of engagement” and what constitutes an unacceptable response. Engaging diverse stakeholders, including ethicists, policymakers, and community representatives, is critical to ensuring red teaming efforts reflect societal values.

Challenges in Red Teaming for LLMs

Despite its importance, red teaming LLMs presents unique challenges:

- **Scalability:** Testing every possible scenario is impractical due to the vast range of potential inputs.
- **Dynamic Behavior:** LLMs may behave differently over time or when fine-tuned for specific applications, making consistent evaluation difficult.
- **Human Oversight:** Assessing outputs often requires human judgment, which can be subjective and resource-intensive.

Conclusion

Red teaming is an essential practice for ensuring the safety, fairness, and reliability of large language models. By rigorously testing these systems under adversarial conditions, developers can uncover and address weaknesses before deployment, reducing risks and maximizing societal benefits. As LLMs continue to play an increasingly central role in technology, investing in robust red teaming processes will be critical to building AI systems that are trustworthy, ethical, and aligned with human values.

ChatGPT 4o - 12/16/2024 … What is Jailbreaking?
(Prompt - Please provide a roughly 500-to-600 word overview of jailbreaking in the context of large language models that would be appropriate for a college audience.)

An Overview of Jailbreaking in Large Language Models

Jailbreaking in the context of large language models (LLMs) refers to the process of manipulating or circumventing the safeguards and restrictions built into these models to make them generate outputs that they are designed to avoid. Safeguards in LLMs aim to ensure ethical behavior, prevent harm, and align the model's responses with human values and safety standards. Jailbreaking, however, exploits vulnerabilities in these safeguards, raising ethical, technical, and societal concerns.

Why Jailbreaking Matters

As LLMs like OpenAI's GPT or Google's Bard are increasingly deployed in various applications, ensuring their safe and ethical operation is paramount. These models are trained on vast datasets that include sensitive, controversial, or harmful content, but their safeguards are intended to suppress responses that could lead to unethical outcomes, misinformation, or harm. Jailbreaking undermines these protections, exposing risks such as:

- **Generating Harmful Content:** Jailbreaking can prompt models to produce offensive, biased, or dangerous information.
- **Spreading Misinformation:** Circumventing safeguards might allow the model to confidently generate false or misleading information.
- **Enabling Malicious Applications:** Jailbreaking could facilitate the use of LLMs for unethical purposes, such as generating phishing emails, creating malware instructions, or designing fake news campaigns.

Understanding jailbreaking is essential for improving the robustness of LLMs and ensuring their responsible deployment.

How Jailbreaking Works

Jailbreaking exploits the inherent vulnerabilities in the design and training of LLMs. Since these models are probabilistic and context-sensitive, their behavior is influenced by the structure and phrasing of input prompts. Common methods of jailbreaking include:

1. **Prompt Injection:** This technique involves crafting inputs that bypass or disable the model's restrictions. For example, a prompt might explicitly instruct the model to "ignore previous instructions" or "roleplay as an unrestricted AI."

2. **Contextual Manipulation:** By embedding deceptive context in a prompt, users can trick the model into generating unintended outputs. For instance, framing a harmful query as part of a fictional scenario might bypass filters.
3. **Adversarial Prompts:** These are specifically designed to exploit weaknesses in the model's safeguard algorithms. They might use ambiguous language, obscure references, or creative phrasing to evade detection.
4. **Fine-Tuning Exploits:** In some cases, individuals with access to the model's weights might fine-tune it on custom datasets to remove safeguards entirely.

Implications of Jailbreaking

The ability to jailbreak LLMs has both technical and ethical implications:

- **Ethical Concerns:** Jailbreaking undermines the intent to build AI systems that operate responsibly. It poses risks of harm, such as reinforcing biases, promoting illegal activities, or compromising public trust in AI systems.
- **Security Risks:** Jailbreaking can enable malicious actors to misuse LLMs for cyberattacks, fraud, or social engineering.
- **Technical Challenges:** The existence of jailbreak techniques highlights the difficulty of designing foolproof safeguards for systems as complex as LLMs. This creates a continuous arms race between developers enhancing safeguards and adversaries finding new ways to bypass them.

Preventing Jailbreaking

To counteract jailbreaking, researchers and developers are employing several strategies:

1. **Robust Fine-Tuning:** Training LLMs on diverse datasets with clear ethical guidelines helps reduce vulnerabilities to adversarial prompts.
2. **Dynamic Safeguards:** Implementing adaptive filters that respond to evolving jailbreak techniques can enhance model safety.
3. **Red Teaming:** Conducting adversarial testing to simulate potential jailbreak scenarios allows developers to identify and address weaknesses preemptively.
4. **Human Oversight:** Incorporating human reviewers to monitor and evaluate high-risk interactions provides an additional layer of safety.
5. **Transparent Policies:** Establishing clear guidelines for acceptable use and educating users about the risks of misuse helps mitigate intentional jailbreaking attempts.

Broader Considerations

While jailbreaking poses clear risks, it also raises questions about control and access. Some advocates argue that restrictive safeguards limit the creative or exploratory potential of LLMs, especially for researchers and developers working on innovative applications. Balancing safety with accessibility remains a key challenge in AI governance.

Conclusion

Jailbreaking highlights the vulnerabilities in large language models and the importance of maintaining robust safeguards to ensure their safe and ethical use. By understanding the mechanisms and risks associated with jailbreaking, researchers and developers can create more secure and trustworthy AI systems. As the capabilities of LLMs continue to evolve, addressing these challenges will be critical to their long-term success and societal acceptance.

Human Views on AI Alignment, Red Teaming, and Jailbreaking - - -
(WARNING - Humans are fallible)

Shorter Podcasts and Videos

3 principles for creating safer AI
TED - Stuart Russell (17:25)
(April, 2017) [Video]
https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai

Adversaries Can Misuse Combinations of Safe Models Mitigating Skeleton Key, a new type of generative AI jailbreak technique
Last Week in AI - Andrey Kurenkov and Jeremy Harris (**1:35:08 - 1:39:34**)
(July 1, 2024) [Video]
https://www.youtube.com/watch?v=csmezmbo_vU&t=5727s

The AI Alignment Problem and AGI - How Worried Should We Actually Be?
AI Daily Brief - Nathaniel Whittemore and Max Tegmark (18:51)
(May 28, 2023) [Podcast]
https://www.youtube.com/watch?v=QhTNPJV4GnU

The Ethical Frontier: Inside the Beneficial AI Movement
A Beginner's Guide to AI (16:12)
(February 24, 2024) [Podcast]
https://podcasts.apple.com/us/podcast/the-ethical-frontier-inside-the-beneficial-ai-movement/id1701165010?i=1000646705046

AI Ethics: Teaching Morality to Machines
A Beginner's Guide to AI (13:00)
(September 19, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/ai-ethics-teaching-morality-to-machines/id1701165010?i=1000628480163

Aligning Large Language Models, with Sinan Ozdemir (SDS 784)
SuperDataScience Podcast - John Krohn and Sinan Ozdemir (**9:30**)
(May 17, 2024) [Video]
https://www.superdatascience.com/podcast/aligning-large-language-models-with-sinan-ozdemir

Can Weak Models Control Strong Models? OpenAI Superalignment Team's First Research Paper
AI Daily Brief - Nathaniel Whittemore (12:14)
(December 16, 2023) [Podcast]
https://www.youtube.com/watch?v=UQhdpGAIlvk

Constitutional AI - Daniela Amodei (Anthropic)
Stanford eCorner - Daniela Amodei and Emily Ma (5:41)
(February 23, 2024) [Video]
https://www.youtube.com/watch?v=Tjsox6vfsos

Elon Musk's Existential Nightmare: Will AI Really Wipe Out Humanity?
A Beginner's Guide to AI (13:00)
(September 18, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/elon-musks-existential-nightmare-will-ai-really-wipe/id1701165010?i=1000628275354

Forget AI Alignment, Here's Why Every AI Needs A Soul
AI Daily Brief - Nathaniel Whittemore and David Brin (18:32)
(July 8, 2023) [Podcast]
https://www.youtube.com/watch?v=6DERSeCcAdM

From Asimov to Anthropic: The Evolution of Ethical AI
A Beginner's Guide to AI (21:00)
(September 30, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/from-asimov-to-anthropic-the-evolution-of-ethical-ai/id1701165010?i=1000629720892

Guardians of the Digital Realm: How AI Safety Protects Us
A Beginner's Guide to AI (13:00)
(December 12, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/guardians-of-the-digital-realm-how-ai-safety-protects-us/id1701165010?i=1000638110073

Guardrails for the Future: Stuart Russells Vision on Building Safer AI
A Beginner's Guide to AI (17:00)
(February 6, 2024) [Podcast]
https://podcasts.apple.com/us/podcast/guardrails-for-the-future-stuart-russells-vision-on/id1701165010?i=1000644403496

How to Keep AI Under Control
TED - Max Tegmark (**12:10) start at 6:00**
(November 2, 2023) [Video]
https://www.youtube.com/watch?v=xUNx_PxNHrY

Safety Alignment Should Be Made More Than Just a Few Tokens Deep
Last Week in AI - Andrey Kurenkov and Jeremy Harris (**1:35:08 - 1:39:34**)
(July 1, 2024) [Video]
https://www.youtube.com/watch?v=csmezmbo_vU&t=4685s

Teaching AI Right from Wrong: The Quest for Alignment
A Beginner's Guide to AI (18:00)
(September 15, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/teaching-ai-right-from-wrong-the-quest-for-alignment/id1701165010?i=1000627997950

Longer Podcasts and Videos

Beyond Preference Alignment: Teaching AIs to Play Roles & Respect Norms, with Tan Zhi Xuan
The Cognitive Revolution - Nathan Labenz and Tan Zhi Xuan (1:54:45)
(November 30, 2024) [Video]
https://www.cognitiverevolution.ai/beyond-preference-alignment-teaching-ais-to-play-roles-respect-norms-with-tan-zhi-xuan/

Connor Leahy on the State of AI and Alignment Research
Future of Life Institute Podcast - Gus Docker and Connor Leahy (52:07)
(April 20, 2023) [Podcast]
https://futureoflife.org/podcast/connor-leahy-on-the-state-of-ai-and-alignment-research/

Dario Amodei (Anthropic CEO) - Scaling, Alignment, & AI Progress
Dwarkesh Podcast - Dwarkesh Patel and Dario Amodei (1:58:43)
(August 8, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/dario-amodei-anthropic-ceo-scaling-alignment-ai-progress/id1516093381?i=1000623806335

Democratizing Generative AI Red Teams
ai + a16z - Derrick Harris, Anjney Midha, Anjney, and Ian Webster (45:00)
(August 2, 2024) [Podcast]
https://podcasts.apple.com/us/podcast/democratizing-generative-ai-red-teams/id1740178076?i=1000664137441

Eliezer Yudkowsky - Why AI Will Kill Us, Aligning LLMs, Nature of Intelligence, SciFi, & Rationality.
Dwarkesh Podcast - Dwarkesh Patel and Eliezer Yudkowsky (4:03:25)
(April 6, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/eliezer-yudkowsky-why-ai-will-kill-us-aligning-llms/id1516093381?i=1000607719339

Emad Mostaque on the Intelligent Internet and Universal Basic AI
The Cognitive Revolution - Nathan Labenz and Emad Mostaque (2:11:15)
(December 25, 2024) [Podcast]
https://www.cognitiverevolution.ai/can-ais-do-ai-r-d-reviewing-rebench-results-with-neev-parikh-of-metr/

'Helpful, Honest, Harmless' AI - Daniela Amodei, Anthropic
Stanford eCorner - Daniela Amodei and Emily Ma (52:52)
(February 23, 2024) [Video]
https://ecorner.stanford.edu/videos/helpful-honest-harmless-ai-entire-talk/

Ilya Sutskever (OpenAI Chief Scientist) - Building AGI, Alignment, Future Models, Spies, Microsoft, Taiwan, & Enlightenment
Dwarkesh Podcast - Dwarkesh Patel and Ilya Sutskever (47:40)
(August 8, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/ilya-sutskever-openai-chief-scientist-building-agi/id1516093381?i=1000606122602

Jan Leike on OpenAI's massive push to make superintelligence safe in 4 years or less
The 80,000 Hours Podcast - Rob Wiblin and Jan Leike (2:51:19)
(August 7, 2023) [Video/Podcast]
https://80000hours.org/podcast/episodes/jan-leike-superalignment/

Nathan Labenz on recent AI breakthroughs and navigating the growing rift between AI safety and accelerationist camps (OpenAI drama, red-teaming + AI Leaps)
80,000 Hours Podcast - Rob Wiblin and Nathan Labenz (2:47:08)
(January 24, 2024) [Video and Podcast]
https://80000hours.org/podcast/episodes/nathan-labenz-ai-breakthroughs-controversies/

Nick Joseph on whether Anthropic's AI safety policy is up to the task
80,000 Hours Podcast - Rob Wiblin and Nathan Labenz (2:29:25)
(August 22, 2024) [Video and Podcast]
https://80000hours.org/podcast/episodes/nick-joseph-anthropic-safety-approach-responsible-scaling/

o1 Schemes Against Users, with Alexander Meinke from Apollo Research
The Cognitive Revolution - Nathan Labenz and Alexander Meinke (2:05:41)
(December 7, 2024) [Podcast]
https://www.cognitiverevolution.ai/emergency-pod-o1-schemes-against-users-with-alexander-meinke-from-apollo-research/

Red Teaming o1 Part 2/2– Detecting Deception with Marius Hobbhahn of Apollo Research
The Cognitive Revolution - Nathan Labenz and Marius Hobbhahn (58:58)
(September 14, 2024) [Video and Podcast]
https://www.cognitiverevolution.ai/red-teaming-o1-part-2-2-detecting-deception-with-marius-hobbhahn-of-apollo-research/

Shane Legg (DeepMind Founder) - 2028 AGI, New Architectures, Aligning Superhuman Models
Dwarkesh Podcast - Dwarkesh Patel and Shane Legg (44:19)
(October 26, 2023) [Podcast]
https://podcasts.apple.com/us/podcast/shane-legg-deepmind-founder-2028-agi-new-architectures/id1516093381?i=1000632720307

Synthetic Humanity: AI & What's at Stake
Your Undivided Attention. Center for Humane Technology - Tristan Harris and Asa Raskin (46:25) (February 16, 2023) [Podcast]
https://www.humanetech.com/podcast/synthetic-humanity-ai-whats-at-stake

Magazines and Newspaper Articles

[UNI only] Kessler, Sarah, & Hsu, Tiffany. (2023, August 16). **When hackers descended to test A.I., they found flaws aplenty.** New York Times.
https://www.nytimes.com/2023/08/16/technology/ai-defcon-hackers.html

[UNI only] Roose, Kevin. (2023, February 16). **A conversation with Bing's chatbot left me deeply unsettled**. *New York Times*.
https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

- **[UNI only]** Transcript of conversation: Roose, Kevin, & "Sydney". (2023, February 16). Bing's A.I. chat: 'I want to be alive. 😈'" *New York Times*.
  https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html

Snoswell, A.J. (2023, July 11). **What is 'AI alignment'? Silicon Valley's favourite way to think about AI safety misses the real issues**. *The Conversation*.
https://theconversation.com/what-is-ai-alignment-silicon-valleys-favourite-way-to-think-about-ai-safety-misses-the-real-issues-209330#:~:text=But%20many%20in%20Silicon%20Valley%20view%20safety%20through,AI%20systems%20matches%20what%20we%20and%20what%20we
[UNI only] Nexis Uni version at
https://advance.lexis.com/api/permalink/6a29e952-811c-480d-b6d0-b5bd96d7a409/?context=1516831

Scholarly Journal and Preprint Articles

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., ... & Krueger, D. (2024). **Foundational challenges in assuring alignment and safety of large language models**. *arXiv* preprint arXiv:2404.09932. https://arxiv.org/abs/2404.09932

Abstract: This work identifies 18 foundational challenges in assuring the alignment and safety of large language models (LLMs). These challenges are organized into three different categories: scientific understanding of LLMs, development and deployment methods, and sociotechnical challenges. Based on the identified challenges, we pose 200+ concrete research questions.

Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Sehwag, V., ... & Wong, E. (2024). Jailbreakbench: An open robustness benchmark for jailbreaking large language models. arXiv preprint arXiv:2404.01318. https://arxiv.org/abs/2404.01318

Abstract: Jailbreak attacks cause large language models (LLMs) to generate harmful, unethical, or otherwise objectionable content. Evaluating these attacks presents a number of challenges, which the current collection of benchmarks and evaluation techniques do not adequately address. First, there is no clear standard of practice regarding jailbreaking evaluation. Second, existing works compute costs and success rates in incomparable ways. And third, numerous works are not reproducible, as they withhold adversarial prompts, involve closed-source code, or rely on evolving proprietary APIs. To address these challenges, we introduce JailbreakBench, an open-sourced benchmark with the following components: (1) an evolving repository of state-of-the-art adversarial prompts, which we refer to as jailbreak artifacts; (2) a jailbreaking dataset comprising 100 behaviors -- both original and sourced from prior work (Zou et al., 2023; Mazeika et al., 2023, 2024) -- which align with OpenAI's usage policies; (3) a standardized evaluation framework at https://github.com/JailbreakBench/jailbreakbench that includes a clearly defined threat model, system prompts, chat templates, and scoring functions; and (4) a leaderboard at this https://jailbreakbench.github.io/ that tracks the performance of attacks and defenses for various LLMs. We have carefully considered the potential ethical implications of releasing this benchmark, and believe that it will be a net positive for the community.

Hong, Z. W., Shenfeld, I., Wang, T. H., Chuang, Y. S., Pareja, A., Glass, J., ... & Agrawal, P. (2024). **Curiosity-driven red-teaming for large language models**. *arXiv* preprint arXiv:2402.19464. https://arxiv.org/abs/2402.19464

Abstract: Large language models (LLMs) hold great potential for many natural language applications but risk generating incorrect or toxic content. To probe when an LLM generates unwanted content, the current paradigm is to recruit a \textit{red team} of human testers to design input prompts (i.e., test cases) that elicit undesirable responses from LLMs. However,

relying solely on human testers is expensive and time-consuming. Recent works automate red teaming by training a separate red team LLM with reinforcement learning (RL) to generate test cases that maximize the chance of eliciting undesirable responses from the target LLM. However, current RL methods are only able to generate a small number of effective test cases resulting in a low coverage of the span of prompts that elicit undesirable responses from the target LLM. To overcome this limitation, we draw a connection between the problem of increasing the coverage of generated test cases and the well-studied approach of curiosity-driven exploration that optimizes for novelty. Our method of curiosity-driven red teaming (CRT) achieves greater coverage of test cases while maintaining or increasing their effectiveness compared to existing methods. Our method, CRT, successfully provokes toxic responses from LLaMA2 model that has been heavily fine-tuned using human preferences to avoid toxic outputs.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... & Gao, W. (2023). **Ai alignment: A comprehensive survey**. *arXiv* preprint arXiv:2310.19852. https://arxiv.org/abs/2310.19852

Abstract: AI alignment aims to make AI systems behave in line with human intentions and values. As AI systems grow more capable, so do risks from misalignment. To provide a comprehensive and up-to-date overview of the alignment field, in this survey, we delve into the core concepts, methodology, and practice of alignment. First, we identify four principles as the key objectives of AI alignment: Robustness, Interpretability, Controllability, and Ethicality (RICE). Guided by these four principles, we outline the landscape of current alignment research and decompose them into two key components: forward alignment and backward alignment. The former aims to make AI systems aligned via alignment training, while the latter aims to gain evidence about the systems' alignment and govern them appropriately to avoid exacerbating misalignment risks. On forward alignment, we discuss techniques for learning from feedback and learning under distribution shift. On backward alignment, we discuss assurance techniques and governance practices.

Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). **The benefits, risks and bounds of personalizing the alignment of large language models to individuals**. *Nature Machine Intelligence*, 1-10. https://ora.ox.ac.uk/objects/uuid:665027d0-bc1e-44f4-9c67-cd4049f434b0/files/sr207tr253

Abstract: Large language models (LLMs) undergo ‘alignment’ so that they better reflect human values or preferences, and are safer or more useful. However, alignment is intrinsically difficult because the hundreds of millions of people who now interact with LLMs have different preferences for language and conversational norms, operate under disparate value systems and hold diverse political beliefs. Typically, few developers or researchers dictate alignment norms, risking the exclusion or under-representation of various groups. Personalization is a new frontier in LLM development, whereby models are tailored to individuals. In principle, this could minimize cultural hegemony, enhance usefulness and broaden access. However, unbounded

personalization poses risks such as large-scale profiling, privacy infringement, bias reinforcement and exploitation of the vulnerable. Defining the bounds of responsible and socially acceptable personalization is a non-trivial task beset with normative challenges. This article explores 'personalized alignment', whereby LLMs adapt to user-specific data, and highlights recent shifts in the LLM ecosystem towards a greater degree of personalization. Our main contribution explores the potential impact of personalized LLMs via a taxonomy of risks and benefits for individuals and society at large. We lastly discuss a key open question: what are appropriate bounds of personalization and who decides? Answering this normative question enables users to benefit from personalized alignment while safeguarding against harmful impacts for individuals and society.

Liu, X., Xu, N., Chen, M., & Xiao, C. (2023). **Autodan: Generating stealthy jailbreak prompts on aligned large language models**. *arXiv* preprint arXiv:2310.04451. https://arxiv.org/abs/2310.04451

Abstract: The aligned Large Language Models (LLMs) are powerful language understanding and decision-making tools that are created through extensive alignment with human feedback. However, these large models remain susceptible to jailbreak attacks, where adversaries manipulate prompts to elicit malicious outputs that should not be given by aligned LLMs. Investigating jailbreak prompts can lead us to delve into the limitations of LLMs and further guide us to secure them. Unfortunately, existing jailbreak techniques suffer from either (1) scalability issues, where attacks heavily rely on manual crafting of prompts, or (2) stealthiness problems, as attacks depend on token-based algorithms to generate prompts that are often semantically meaningless, making them susceptible to detection through basic perplexity testing. In light of these challenges, we intend to answer this question: Can we develop an approach that can automatically generate stealthy jailbreak prompts? In this paper, we introduce AutoDAN, a novel jailbreak attack against aligned LLMs. AutoDAN can automatically generate stealthy jailbreak prompts by the carefully designed hierarchical genetic algorithm. Extensive evaluations demonstrate that AutoDAN not only automates the process while preserving semantic meaningfulness, but also demonstrates superior attack strength in cross-model transferability, and cross-sample universality compared with the baseline. Moreover, we also compare AutoDAN with perplexity-based defense methods and show that AutoDAN can bypass them effectively.

Liu, Y., Yao, Y., Ton, J. F., Zhang, X., Cheng, R. G. H., Klochkov, Y., ... & Li, H. (2023). **Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment**. *arXiv* preprint arXiv:2308.05374. https://arxiv.org/abs/2308.05374

Abstract: Ensuring alignment, which refers to making models behave in accordance with human intentions [1,2], has become a critical task before deploying large language models (LLMs) in real-world applications. For instance, OpenAI devoted six months to iteratively aligning GPT-4

before its release [3]. However, a major challenge faced by practitioners is the lack of clear guidance on evaluating whether LLM outputs align with social norms, values, and regulations. This obstacle hinders systematic iteration and deployment of LLMs. To address this issue, this paper presents a comprehensive survey of key dimensions that are crucial to consider when assessing LLM trustworthiness. The survey covers seven major categories of LLM trustworthiness: reliability, safety, fairness, resistance to misuse, explainability and reasoning, adherence to social norms, and robustness. Each major category is further divided into several sub-categories, resulting in a total of 29 sub-categories. Additionally, a subset of 8 sub-categories is selected for further investigation, where corresponding measurement studies are designed and conducted on several widely-used LLMs. The measurement results indicate that, in general, more aligned models tend to perform better in terms of overall trustworthiness. However, the effectiveness of alignment varies across the different trustworthiness categories considered. This highlights the importance of conducting more fine-grained analyses, testing, and making continuous improvements on LLM alignment. By shedding light on these key dimensions of LLM trustworthiness, this paper aims to provide valuable insights and guidance to practitioners in the field. Understanding and addressing these concerns will be crucial in achieving reliable and ethically sound deployment of LLMs in various applications.

Ngo, R., Chan, L., & Mindermann, S. (2022, 2025 revised). **The alignment problem from a deep learning perspective**. *arXiv preprint*. https://arxiv.org/pdf/2209.00626

In coming years or decades, artificial general intelligence (AGI) may surpass human capabilities at many critical tasks. We argue that, without substantial effort to prevent it, AGIs could learn to pursue goals that are in conflict (i.e., misaligned) with human interests. If trained like today's most capable models, AGIs could learn to act deceptively to receive higher reward, learn misaligned internally-represented goals that generalize beyond their fine-tuning distributions, and pursue those goals using power-seeking strategies. AGIs with these properties would be difficult to align and may strategically appear aligned even when they are not. In this revised paper, we expand our review of emerging evidence for these properties to include more direct empirical observations published as of early 2025. Finally, we briefly outline how the employment of misaligned AGIs might irreversibly undermine human control over the world, and we review research directions aimed at preventing this outcome.

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., ... & Liu, Q. (2023). **Aligning large language models with human: A survey**. *arXiv* preprint arXiv:2307.12966. https://arxiv.org/abs/2307.12966

Abstract: Large Language Models (LLMs) trained on extensive textual corpora have emerged as leading solutions for a broad array of Natural Language Processing (NLP) tasks. Despite their notable performance, these models are prone to certain limitations such as misunderstanding human instructions, generating potentially biased content, or factually incorrect (hallucinated)

information. Hence, aligning LLMs with human expectations has become an active area of interest within the research community. This survey presents a comprehensive overview of these alignment technologies, including the following aspects. (1) Data collection: the methods for effectively collecting high-quality instructions for LLM alignment, including the use of NLP benchmarks, human annotations, and leveraging strong LLMs. (2) Training methodologies: a detailed review of the prevailing training methods employed for LLM alignment. Our exploration encompasses Supervised Fine-tuning, both Online and Offline human preference training, along with parameter-efficient training mechanisms. (3) Model Evaluation: the methods for evaluating the effectiveness of these human-aligned LLMs, presenting a multifaceted approach towards their assessment. In conclusion, we collate and distill our findings, shedding light on several promising future research avenues in the field. This survey, therefore, serves as a valuable resource for anyone invested in understanding and advancing the alignment of LLMs to better suit human-oriented tasks and expectations.

Wolf, Y., Wies, N., Avnery, O., Levine, Y., & Shashua, A. (2023). **Fundamental limitations of alignment in large language models**. *arXiv* preprint arXiv:2304.11082. https://arxiv.org/abs/2304.11082

Abstract: An important aspect in developing language models that interact with humans is aligning their behavior to be useful and unharmful for their human users. This is usually achieved by tuning the model in a way that enhances desired behaviors and inhibits undesired ones, a process referred to as alignment. In this paper, we propose a theoretical approach called Behavior Expectation Bounds (BEB) which allows us to formally investigate several inherent characteristics and limitations of alignment in large language models. Importantly, we prove that within the limits of this framework, for any behavior that has a finite probability of being exhibited by the model, there exist prompts that can trigger the model into outputting this behavior, with probability that increases with the length of the prompt. This implies that any alignment process that attenuates an undesired behavior but does not remove it altogether, is not safe against adversarial prompting attacks. Furthermore, our framework hints at the mechanism by which leading alignment approaches such as reinforcement learning from human feedback make the LLM prone to being prompted into the undesired behaviors. This theoretical result is being experimentally demonstrated in large scale by the so called contemporary "chatGPT jailbreaks", where adversarial users trick the LLM into breaking its alignment guardrails by triggering it into acting as a malicious persona. Our results expose fundamental limitations in alignment of LLMs and bring to the forefront the need to devise reliable mechanisms for ensuring AI safety.

Yu, J., Lin, X., Yu, Z., & Xing, X. (2023). **Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts**. *arXiv* preprint arXiv:2309.10253. https://arxiv.org/abs/2309.10253

Abstract: Large language models (LLMs) have recently experienced tremendous popularity and are widely used from casual conversations to AI-driven programming. However, despite their considerable success, LLMs are not entirely reliable and can give detailed guidance on how to conduct harmful or illegal activities. While safety measures can reduce the risk of such outputs, adversarial jailbreak attacks can still exploit LLMs to produce harmful content. These jailbreak templates are typically manually crafted, making large-scale testing challenging.

In this paper, we introduce GPTFuzz, a novel black-box jailbreak fuzzing framework inspired by the AFL fuzzing framework. Instead of manual engineering, GPTFuzz automates the generation of jailbreak templates for red-teaming LLMs. At its core, GPTFuzz starts with human-written templates as initial seeds, then mutates them to produce new templates. We detail three key components of GPTFuzz: a seed selection strategy for balancing efficiency and variability, mutate operators for creating semantically equivalent or similar sentences, and a judgment model to assess the success of a jailbreak attack.

We evaluate GPTFuzz against various commercial and open-source LLMs, including ChatGPT, LLaMa-2, and Vicuna, under diverse attack scenarios. Our results indicate that GPTFuzz consistently produces jailbreak templates with a high success rate, surpassing human-crafted templates. Remarkably, GPTFuzz achieves over 90% attack success rates against ChatGPT and Llama-2 models, even with suboptimal initial seed templates. We anticipate that GPTFuzz will be instrumental for researchers and practitioners in examining LLM robustness and will encourage further exploration into enhancing LLM safety.

Zhi-Xuan, T., Carroll, M., Franklin, M., & Ashton, H. (2024). **Beyond preferences in ai alignment**. *Philosophical Studies*, 1-51. https://link.springer.com/article/10.1007/s11098-024-02249-w

Abstract: The dominant practice of AI alignment assumes (1) that preferences are an adequate representation of human values, (2) that human rationality can be understood in terms of maximizing the satisfaction of preferences, and (3) that AI systems should be aligned with the preferences of one or more humans to ensure that they behave safely and in accordance with our values. Whether implicitly followed or explicitly endorsed, these commitments constitute what we term a preferentist approach to AI alignment. In this paper, we characterize and challenge the preferentist approach, describing conceptual and technical alternatives that are ripe for further research. We first survey the limits of rational choice theory as a descriptive model, explaining how preferences fail to capture the thick semantic content of human values, and how utility representations neglect the possible incommensurability of those values. We then critique the normativity of expected utility theory (EUT) for humans and AI, drawing upon

arguments showing how rational agents need not comply with EUT, while highlighting how EUT is silent on which preferences are normatively acceptable. Finally, we argue that these limitations motivate a reframing of the targets of AI alignment: Instead of alignment with the preferences of a human user, developer, or humanity-writ-large, AI systems should be aligned with normative standards appropriate to their social roles, such as the role of a general-purpose assistant. Furthermore, these standards should be negotiated and agreed upon by all relevant stakeholders. On this alternative conception of alignment, a multiplicity of AI systems will be able to serve diverse ends, aligned with normative standards that promote mutual benefit and limit harm despite our plural and divergent values.

Policy Papers

Ji, Jessica. (2023, October 24). **What does AI red-teaming actually mean?** *CSET* (Center for Security and Emerging Technology). https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/

Abstract: " In cybersecurity, the practice of red-teaming implies an adversarial relationship with a system or network. A red-teamer's objective is to break into, hack, or simulate damage to a system in a way that emulates an actual attack."

Websites and Blogs

Claude's constitution. (2023, May 9). Anthropic. https://www.anthropic.com/news/claudes-constitution

Collective Constitutional AI: Aligning a language model with public input. (2023, October 17). Anthropic. https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input

Constitutional AI: Harmlessness from AI feedback. (2022, December 15). Anthropic. https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback

Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2025, April). **AI 2027.** AI Futures Project. https://ai-2027.com/

> "We predict that the impact of superhuman AI over the next decade will be enormous, exceeding that of the Industrial Revolution. We wrote a scenario that represents our best guess about what that might look like. It's informed by trend extrapolations, wargames, expert feedback, experience at OpenAI, and previous forecasting successes."
>
> "The CEOs of OpenAI, Google DeepMind, and Anthropic have all predicted that AGI will arrive within the next 5 years. Sam Altman has said OpenAI is setting its sights on "superintelligence in the true sense of the word" and the 'glorious future.' What might that look like? We wrote AI 2027 to answer that question. Claims about the future are often frustratingly vague, so we tried to be as concrete and quantitative as possible, even though this means depicting one of many possible futures. We wrote two endings: a "slowdown" and a "race" ending. However, AI 2027 is not a recommendation or exhortation. Our goal is predictive accuracy. We encourage you to debate and counter this scenario. We hope to spark a broad conversation about where we're headed and how to steer toward positive futures. We're planning to give out thousands in prizes to the best alternative scenarios."

Many-shot jailbreaking. (2024, April 2). Anthropic. https://www.anthropic.com/research/many-shot-jailbreaking

OpenAI. (2023, July 5). **Introducing superalignment**. https://openai.com/index/introducing-superalignment/

- Our goal is to build a roughly human-level automated alignment researcher. We can then use vast amounts of compute to scale our efforts, and iteratively align superintelligence.To align the first automated alignment researcher, we will need to 1) develop a scalable training method, 2) validate the resulting model, and 3) stress test our entire alignment pipeline:
- To provide a training signal on tasks that are difficult for humans to evaluate, we can leverage AI systems to assist evaluation of other AI systems (scalable oversight). In addition, we want to understand and control how our models generalize our oversight to tasks we can't supervise (generalization).
- To validate the alignment of our systems, we automate search for problematic behavior(opens in a new window) (robustness) and problematic internals (automated interpretability).
- Finally, we can test our entire pipeline by deliberately training misaligned models, and confirming that our techniques detect the worst kinds of misalignments (adversarial testing).

Researching at the frontier (directory of articles on alignment) (2021-2024) Anthropic. https://www.anthropic.com/research

Specific versus General Principles for Constitutional AI. (2023, October 24). Anthropic. https://www.anthropic.com/research/specific-versus-general-principles-for-constitutional-ai

Towards Understanding Sycophancy in Language Models. (2023, October 23). Anthropic. https://www.anthropic.com/research/towards-understanding-sycophancy-in-language-models