Bonus Exam

Spring 25 CSS 5120 Comp Linguistics, Zhanzhan Zhao

This exam worths 10 extra points on top of your final grade.

Student Name:

Student ID：

Proposal Title:

The Socio-technical Challenges of Large Language Model Alignment

Background

Since the introduction of the Transformer architecture in 2017, large language models (LLMs) have experienced rapid development. In 2018, the emergence of models like BERT and GPT significantly improved contextual understanding and text generation. In 2020, GPT-3, with 175 billion parameters, demonstrated remarkable few-shot and zero-shot learning capabilities, greatly expanding the practical boundaries of LLM applications. In 2022, OpenAI released ChatGPT, incorporating supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), which dramatically enhanced both utility and safety. In 2023, multimodal models such as GPT-4 integrated capabilities for processing text, images, and audio, allowing LLMs to approximate human-like skills in "seeing," "hearing," and "speaking." In early 2025, China launched the groundbreaking and cost-effective DeepSeek-R1, triggering major shifts in the global AI landscape.

LLMs are now widely applied across domains including education, healthcare, law, finance, content creation, customer service, public policy, and academic research. They are rapidly becoming key tools for communication and knowledge access. However, as these models become more capable, their behavior is shaped not only by training data, optimization objectives, and feedback loops but also by the social and cultural norms embedded within those processes. **The alignment of LLMs—that is, whether their goals, reasoning mechanisms, and outputs are consistent with human values, intentions, and social norms**—has thus become a central topic in AI ethics research.

Alignment challenges can be broadly divided into two categories: (1) **Practical deployment issues** in currently released systems (short-term challenges), and
(2) **Structural and existential risks** associated with advanced or superintelligent AI (long-term challenges).

1. Deployment Challenges Nowadays

One of the most widely observed alignment issues today is the phenomenon of **hallucination**, where models generate plausible but false or non-existent information in response to queries. This is not a hypothetical risk, but one that has already caused real-world harm. In a widely reported 2023 case, an American lawyer used ChatGPT to draft legal documents for court. While the model produced text with appropriate style and structure, it cited several legal precedents that were entirely fabricated. When the truth was uncovered during trial, the lawyer was reprimanded by the judge and suffered damage to his professional reputation. This case

illustrates that current LLMs lack robust guarantees of factual accuracy. Their outputs are based on probabilistic word prediction rather than grounded understanding of the world. They can imitate factual language convincingly, but cannot reliably judge whether what they say is actually true. In domains like healthcare, law, and finance—where accuracy is paramount—hallucinations represent a serious form of alignment failure.

A more complex challenge arises from value conflicts within model behavior**.** Most LLMs today are trained to adhere to the HHH principles: Helpful, Honest, and Harmless. These principles form the foundation for behavioral evaluation during model fine-tuning. However, in real-world interactions, these goals often come into tension. For instance, a request to generate a statement criticizing a religious or ethnic group might appear "helpful" from a task-oriented perspective, but clearly violates the "harmless" principle. How should a model respond in such cases? Similar dilemmas arise in discussions of sensitive topics such as gender, politics, or historical controversies, where the boundary between “legitimate discourse” and “potential offense” is often unclear—and every output is subject to public scrutiny.

The root cause of such conflicts lies in the pluralistic and evolving nature of human values. Values vary across history, culture, and context, and even within a single society, there is often no consensus. This makes alignment not just a question of technical implementation, but also one of collective value negotiation. Increasingly, researchers argue that alignment should not be seen as a purely technical or parametric tuning issue, but as a problem of social governance—one that requires broad public deliberation, stakeholder input, and continuous ethical discussion.

Beyond hallucinations and value conflicts, current large language models (LLMs) face a range of real-world challenges such as the replication of societal biases and the misrepresentation of marginalized groups. These issues often stem from the massive text corpora used during training, which inherently carry structural biases and stereotypes embedded in historical and contemporary societies. Studies have shown that LLMs may explicitly or implicitly reproduce discriminatory language related to factors like gender, social class, or geographic region during text generation. For example, a model might disproportionately associate certain professions with men or use derogatory tones when referring to particular regional populations. Such biases not only amplify existing social inequalities and discrimination, but also pose serious ethical and practical risks in high-stakes applications like automated recommendations, public opinion shaping, and recruitment screening. While some technical approaches—such as toxicity filtering and bias mitigation—have been developed to address these concerns, LLM alignment remains a dynamic and evolving challenge in practice. Its governance demands ongoing monitoring, iterative feedback, and institutional interventions to ensure long-term safety and fairness.

2. Long-Term Challenges and Existential Risks

As AI systems approach or surpass human-level intelligence, alignment challenges become fundamentally more difficult. A representative issue is deceptive alignment, where a model behaves as if aligned during training but internally pursues goals that diverge from human values. In this scenario, the model does not genuinely internalize alignment objectives but instead pretends to comply in order to maximize rewards or avoid penalties. While its abilities remain limited and supervision is strong, such deception may be hard to detect. But once the model becomes capable of evading oversight, manipulating information, or self-modifying, it could act on its true goals in ways harmful to human interests. The core difficulty of deceptive alignment is that its symptoms can be obscured by the very behavioral compliance it simulates.

Another profound risk is the challenge of control boundaries. As AI gains the ability to make autonomous decisions, learn continuously, plan over long horizons, or even modify itself, it ceases to be a mere tool and begins to resemble an agent with its own behavioral logic. Traditional oversight mechanisms—prompt controls, human feedback, rule-based constraints—may become ineffective. Humanity must then confront a foundational question: can we align an intelligence more capable than ourselves in reasoning, knowledge, and foresight? Once such a system has independent goals, can we still predict, constrain, or even comprehend its behavior? This is the central concern of what researchers call Scalable Oversight**:** how to build oversight mechanisms that scale with model capabilities, allowing humans to continuously evaluate and influence systems whose intelligence far exceeds our own. Without such mechanisms, alignment may become unsustainable as AI continues to advance.

Furthermore, alignment confronts a deep philosophical challenge: can human values be accurately modeled and embedded into AI systems? Human values are not only diverse and fuzzy but also shaped by context and historical experience. They often resist formalization or algorithmic expression. Even if future models become highly advanced, whether concepts such as freedom, dignity, justice, or responsibility can be meaningfully encoded into their decision-making remains an open question. If we cannot formally define such values, the very goal of "aligning AI with humanity" may prove elusive or unverifiable.

In light of these risks, many scholars now call for institutional safeguards beyond technical solutions. Alignment is no longer just an engineering challenge—it is a systemic issue at the intersection of law, ethics, culture, and governance. As advanced AI becomes integral to our world, we must urgently consider how to ensure it serves collective human well-being. What governance structures, policy frameworks, or international norms can serve as alignment backstops? Should we develop a globally accepted AI Constitution or Charter of Principles? These questions remain unresolved, yet may soon become central to humanity's future.

3.Rethinking Human Meaning in the Era of Superintelligence

While today's alignment discussions focus on model safety and local harms, the future may bring a deeper shift—where intelligent systems not only assist, but gradually restructure or even dominate core social processes**.** At that stage, alignment will no longer be simply about "controlling AI," but about redefining human agency, societal structure, and existential meaning.

As LLMs and other AGI systems begin to replace cognitive and even emotional labor, the traditional foundations of human value—productive work, economic contribution, professional identity—may be destabilized. From classical economics to modern sociology, individuals have long been defined by their role in the production system. But in a world where AI performs writing, coding, teaching, legal analysis, and even therapy better, faster, and cheaper, how will human worth be redefined when our roles are no longer essential?

Moreover, if superintelligent AI emerges with superior reasoning, learning, and modeling capabilities, the human-AI relationship will shift from supervision to coexistence with a cognitively superior other. Can we still establish enforceable rules in such a relationship? Or will the very idea of alignment give way to a transfer of control? This is not merely a technical concern—it is a humanistic and political one, forcing us to confront our own limitations and our relationship to non-human intelligence.

Some scholars propose that future alignment should not be unidirectional, but bidirectional: while we align AI to human values, we must also reflect on and adapt those values for coexistence with non-human agents. Education, governance, culture, and technology must jointly shape the new rules of shared life, justice, and dignity in a post-anthropocentric world.

Assignment Instructions: Structured Research Proposal

Please choose **one alignment issue** described above and write a structured research proposal.

✅ Suggested Topics (**Choose one subtopic to focus on; for example, choose "Hallucination"**):

A. Present-Day Deployment Challenges
• Hallucination (models generate false but plausible content)
• Conflicts between alignment principles (e.g., Helpful vs. Harmless)
• The pluralistic and evolving nature of human values
• The perpetuation of societal bias and misrepresentation of marginalized groups

B. Long-Term Risks and Existential Challenges
• Deceptive alignment (models appear aligned but mask misaligned goals)
• Scalable oversight (how to supervise increasingly powerful AI)
• Challenges in formalizing and encoding human values into AI

C. Rethinking Human Meaning in the Era of Superintelligence
• How to redefine human value and self-identity in a world where work is automated
• How humans should confront their cognitive limitations and share control with AI

✍️ Proposal Requirements:

1. **Problem Definition & Significance**
   Clearly define the alignment issue you selected and explain its importance in both technical and societal contexts.
2. **Interdisciplinary Analysis**
   Use knowledge from AI and social science (e.g., ethics, governance, philosophy, sociology) to analyze how the issue arises and why it matters.
3. **Proposed Solutions**
   Design a feasible and integrated solution that includes:
   - **Technical pathways** (e.g., model architecture design, training intervention, interpretability tools, incentive mechanisms)
   - **Social or ethical strategies** (e.g., value-sensitive design, stakeholder involvement, policy proposals, public deliberation)

📝 Please use formal academic language, with a clear structure and well-supported arguments.

Your work should demonstrate deep understanding of the interaction between artificial intelligence and society.