

# AI Safety: Challenges and Solutions



P.Baldi

Department of Computer Science  
Institute for Genomics and Bioinformatics  
AI in Science Institute  
University of California, Irvine

1. AI Opportunities

2. AI Challenges

AI Telescope

AI Safety Framework



1. AI Opportunities

2. AI Challenges

AI Telescope

AI Safety Framework

# Polyp Detection

## Colorectal Cancer How You Can Prevent It

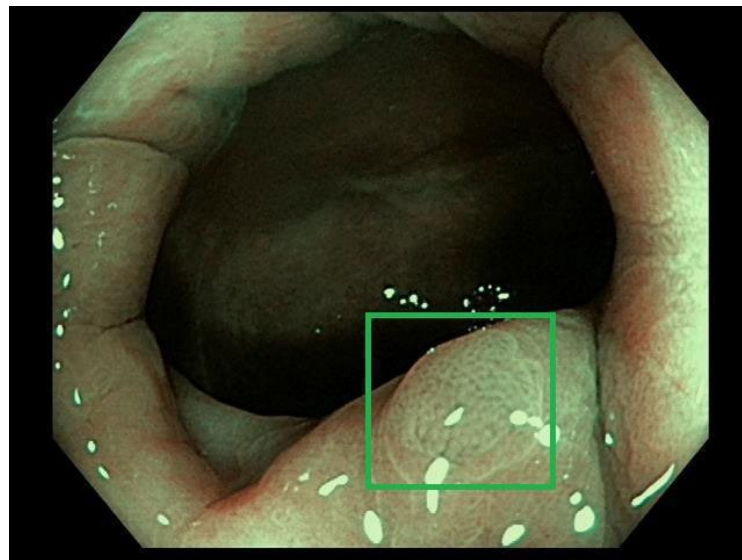
William Karnes, MD, AGAF  
Clinical Associate Professor of Medicine  
UCI, Ohio Comprehensive Digestive Disease Center



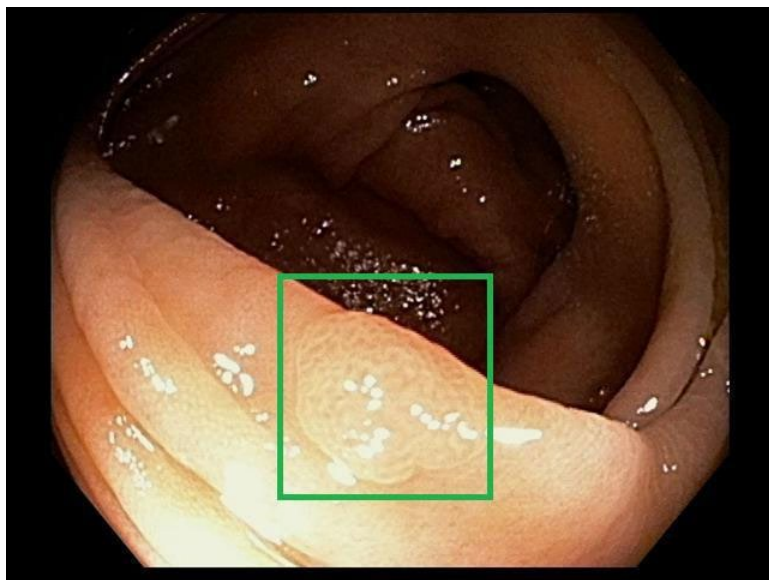
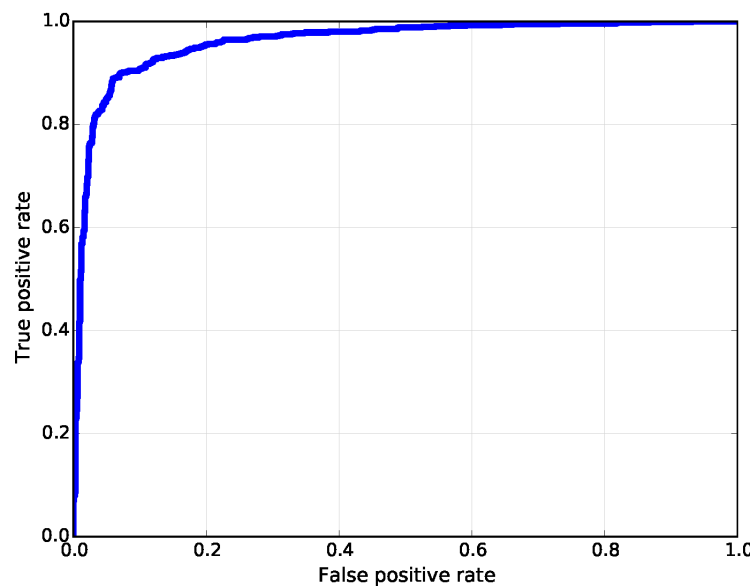
William Karnes



Gregor Urban



ROC Curve



The rate of adenoma detection by colonoscopists varies from 7% to 53%. It is estimated that every 1% increase in ADR reduces the risk of interval colorectal cancers by 3-6%. New strategies are needed to increase the ADR during colonoscopy



Gregor Urban, Priyam Tripathi, Talal Alkayali, Mohit Mittal, Farid Jalali, William Karnes, and Pierre Baldi. Deep Learning Localizes and Identifies Polyps in Real Time with 96% Accuracy in Screening Colonoscopy. *Gastroenterology*, Volume 155, Issue 4, Pages 1069–1078, (2018).



Siwei Chen, Gregor Urban, and Pierre Baldi. Weakly Supervised Polyp Segmentation in Colonoscopy Images using Deep Neural Networks. *Journal of Imaging*, 8, 5, 121, (2022).

medRxiv  
THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

HOME | SUBMIT | FAQ | BLOG | ALERTS / RSS | ABOUT

Search  
Advanced Search

Previous Next

Posted December 21, 2022.

[Download PDF](#)  
[Print/Save Options](#)  
[Author Declarations](#)  
[Data/Code](#)  
[Revision Summary](#)

[Email](#)  
[Share](#)  
[Citation Tools](#)

[Tweet](#)

**COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv**

**Subject Area**  
**Medical Education**

**Subject Areas**

All Articles

- Addiction Medicine
- Allergy and Immunology
- Anesthesia
- Cardiovascular Medicine
- Dentistry and Oral Medicine
- Dermatology
- Emergency Medicine

**Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models**

Tiffany H. Kung, Morgan Cheatham, ChatGPT, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, Victor Tseng  
doi: <https://doi.org/10.1101/2022.12.19.22283643>

**This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.**

[0](#) [0](#) [0](#) [0](#) [121](#) [0](#) [0](#) [6418](#)

[Abstract](#) [Full Text](#) [Info/History](#) [Metrics](#) [Preview PDF](#)

**ABSTRACT**

We evaluated the performance of a large language model called ChatGPT on the United States Medical Licensing Exam (USMLE), which consists of three exams: Step 1, Step 2CK, and Step 3. ChatGPT performed at or near the passing threshold for all three exams without any specialized training or reinforcement. Additionally, ChatGPT demonstrated a high level of concordance and insight in its explanations. These results suggest that large language models may have the potential to assist with medical education, and potentially, clinical decision-making.

**Competing Interest Statement**

The authors have declared no competing interest.

**Funding Statement**

This study did not receive any external funding

Other example of application: [Pharmacy Automation](#)

# Clinical Knowledge and Reasoning Abilities of AI Large Language Models in Anesthesiology: A Comparative Study on the American Board of Anesthesiology Examination

Mirana C. Angel, MSc,\*† Joseph B. Rinehart, MD,‡ Maxime P. Cannesson, MD, PhD,§ and Pierre Baldi, PhD\*†

**BACKGROUND:** Over the past decade, artificial intelligence (AI) has expanded significantly with increased adoption across various industries, including medicine. Recently, AI-based large language models such as Generative Pretrained Transformer-3 (GPT-3), Bard, and Generative Pretrained Transformer-3 (GPT-4) have demonstrated remarkable language capabilities. While previous studies have explored their potential in general medical knowledge tasks, here we assess their clinical knowledge and reasoning abilities in a specialized medical context.

**METHODS:** We studied and compared the performance of all 3 models on both the written and oral portions of the comprehensive and challenging American Board of Anesthesiology (ABA) examination, which evaluates candidates' knowledge and competence in anesthesia practice.

**RESULTS:** Our results reveal that only GPT-4 successfully passed the written examination, achieving an accuracy of 78% on the basic section and 80% on the advanced section. In comparison, the less recent or smaller GPT-3 and Bard models scored 58% and 47% on the basic examination, and 50% and 46% on the advanced examination, respectively. Consequently, only GPT-4 was evaluated in the oral examination, with examiners concluding that it had a reasonable possibility of passing the structured oral examination. Additionally, we observe that these models exhibit varying degrees of proficiency across distinct topics, which could serve as an indicator of the relative quality of information contained in the corresponding training datasets. This may also act as a predictor for determining which anesthesiology subspecialty is most likely to witness the earliest integration with AI.

**CONCLUSIONS:** GPT-4 outperformed GPT-3 and Bard on both basic and advanced sections of the written ABA examination, and actual board examiners considered GPT-4 to have a reasonable possibility of passing the real oral examination; these models also exhibit varying degrees of proficiency across distinct topics. (Anesth Analg 2024;139:349–56)

## KEY POINTS

- **Question:** How might recent advancements in artificial intelligence (AI) large language models influence the field of anesthesiology?
- **Findings:** Large language models may now be sophisticated enough to pass the anesthesiology written and oral examinations.
- **Meaning:** The rapid development of these models holds the potential to shape the future of both anesthesiology education and practice, but we need to be aware of their limitations.

In recent years, artificial intelligence (AI) primarily in the form of machine learning, in particular deep learning, has experienced a significant expansion driven by progress in computational power and big

data availability.<sup>1</sup> In the medical field, AI's potential to increase accuracy and expedite diagnoses has led to its application in numerous areas, including radiology, pathology, and genomics. For example, AI has

From the \*Department of Computer Science, University of California Irvine, Irvine, California; †Institute for Genomics and Bioinformatics, University of California Irvine, Irvine, California; ‡Department of Anesthesiology & Perioperative Care, University of California Irvine, Irvine, California; and §Department of Anesthesiology & Perioperative Medicine, University of California Los Angeles, Los Angeles, California.

Accepted for publication November 27, 2023.

Copyright © 2024 International Anesthesia Research Society

DOI: 10.1213/ANE.000000000000692

Funding: This work was supported by NIH R01EB029751 (to MPC, PE, and JBR). The authors declare no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (www.anesthesia-analgesia.org).

Reprints will not be available from the authors.

Address correspondence to Pierre Baldi, PhD, Department of Computer Science, University of California Irvine, Irvine, CA 92697. Address e-mail to pfbaldi@uci.edu.

# AI and Veterinary Medicine: Performance of Large Language Models on the North American Licensing Examination

Mirana Angel  
*Institute for Geomics and Bioinformatics  
University of California Irvine  
Irvine, USA  
mcangel@uci.edu*

Anuj Patel  
*Department of Computer Science  
University of California Irvine  
Irvine, USA  
patelad2@uci.edu*

Haiyi Xing  
*Department of Computer Science  
University of California Irvine  
Irvine, USA  
haiyix2@uci.edu*

Dylan Balsz  
*Internal Medicine  
Anivive Life Sciences  
Long Beach, USA  
dylan@anivive.com*

Cody Arbuckle  
*Internal Medicine  
Anivive Life Sciences  
Long Beach, USA  
cody@anivive.com*

David Bruyette  
*Internal Medicine  
Anivive Life Sciences  
Long Beach, USA  
david@anivive.com*

Pierre Baldi  
*Department of Computer Science  
University of California Irvine  
Irvine, USA  
ptbaldi@uci.edu*

**Abstract**—This study aimed to assess the performance of Large Language Models on the North American Veterinary Licensing Examination (NAVLE) and to analyze the impact of artificial intelligence in the domain of animal healthcare. For this study, a 200-question NAVLE self-assessment sourced from ICVA's website was used to evaluate the performance of three language models: GPT-3, GPT-4, and Bard. Questions involving images were omitted leaving a 164 text-only sample exam. Results were analyzed by comparing generated responses to the answer key, and scores were assigned to evaluate the models' veterinary medical reasoning capabilities. Our results showed that GPT-4 outperformed GPT-3 and Bard, passing the exam with 89 % of the text-only questions correctly. GPT-3 and Bard only achieved an accuracy of 63.4 % and 61 % respectively on the same set of questions. Language models hold promise for enhancing veterinary practices through expanded educational opportunities in the veterinary curriculum, improved diagnostic accuracy, treatment times, and efficiency. However, potential negatives include challenges in changing the current educational paradigm, reduced demand for professionals or paraprofessional concerns surrounding machine-generated decisions. Responsible and ethical integration of language models is crucial in veterinary medicine.

**Index Terms**—Artificial Intelligence, LLM, ChatGPT, Bard, Veterinary Medicine, Medical Education, Societal Impact

## I. INTRODUCTION

In recent years, the rapid growth of artificial intelligence (AI) has significantly influenced various industries, including healthcare. The development of increasingly powerful AI models, such as large language models (LLMs) has facilitated the automation of diverse tasks and the enhancement of decision-making processes. Consequently, the adoption of AI technology has emerged as a pivotal factor in gaining a competitive edge and boosting efficiency across industries [1]. Here we provide an initial assessment of the applicability of

LLMs in veterinary medicine by testing their ability to pass a standard veterinary education test.

The veterinary field encompasses a wide array of professions and specializations, all dedicated to the care and well-being of animals. Veterinarians, who are extensively trained to diagnose and treat various conditions in numerous species ranging from domesticated animals and livestock to wildlife, are a cornerstone of this field. As the veterinary field continues to evolve, new technologies and techniques are revolutionizing the diagnosis and treatment of animal health issues [2].

The advent of diverse AI technologies, such as state-of-the-art text, sound, image, and video data analysis algorithms, have significantly advanced veterinary medicine in areas such as disease diagnosis, treatment planning, and precision medicine [2, 3, 4]. However, current AI models are typically task-specific and lack the capability for independent medical reasoning [5]. This limitation has prompted researchers to explore the potential of large language models, which have demonstrated remarkable cognitive reasoning abilities, in addressing these shortcomings in all fields.

Among large language models, Generative Pre-trained Transformer (GPT) and Bard have emerged as frontrunners, exhibiting outstanding performance in various applications [6, 7, 8]. GPT-3 and GPT-4, as well as Bard, adopt the decoder-only architecture of the transformer model [9]. GPT-3 encompasses 175 billion parameters and showcases remarkable versatility across a range of tasks. In an advancement over GPT-3, GPT-4 boasts an unprecedented one trillion parameters, addressing many of the limitations previously associated with GPT-3. Both GPT iterations were pre-trained on extensive text corpora and subsequently fine-tuned for specialized tasks [6, 7].

Concurrently, Google's Bard initially employed the Lan-

**Clinical Knowledge and Reasoning Abilities of AI Large Language Models in  
Anesthesiology: A Comparative Study on the ABA Exam**

Mirana C. Angel MSc<sup>1,2</sup>, Joseph B. Rinehart MD<sup>3</sup>, Maxime P. Canneson MD PhD<sup>4</sup>, Pierre Baldi

PhD<sup>1,2,\*</sup>

1. Department of Computer Science, University of California Irvine, Irvine, CA 92697, USA
2. Institute for Genomics and Bioinformatics, University of California Irvine, Irvine CA 92697, USA
3. Department of Anesthesiology & Perioperative Care, University of California Irvine, Irvine CA 92697, USA
4. Department of Anesthesiology & Perioperative Medicine, University of California, Los Angeles, Los Angeles, CA 90095





Research paper

# Evaluating the Intelligence of large language models: A comparative study using verbal and visual IQ tests

Sherif Abdelkarim <sup>a,1</sup>✉, David Lu <sup>a,1</sup>✉, Dora-Luz Flores <sup>c</sup>✉, Susanne Jaeggi <sup>a,b</sup>✉, Pierre Baldi <sup>a</sup>✉

[Show more](#) ✓

[+](#) Add to Mendeley [🔗](#) Share [📄](#) Cite

<https://doi.org/10.1016/j.chbah.2025.100170> ↗

[Get rights and content](#) ↗

Under a Creative Commons [license](#) ↗

● [Open access](#)

## Highlights

- Evaluated cognitive performance of popular LLMs using verbal and visual IQ tests.
- Found a positive correlation between LLM size and cognitive performance across tasks.
- Significant performance variability across problem types suggests nuanced differences in reasoning.

**Gold Medal Math  
Olympiads  
>50 score on Humanity  
Last Exam**

1. AI Opportunities

2. AI Challenges

AI Telescope

AI Safety Framework

3. Information Theory (if time permits)

# Major AI Dangers

1. Nefarious uses of AI by bad actors.
2. Employment (short term, medium term, long term).
3. Loss of dignity, loss of purpose, loss of sense of reality and social connection.
4. Existential threat.

# One Major Obstacle

- Universities cannot develop the most cutting-edge AI.
- Universities cannot study many of the safety problems and possible solutions associated with cutting edge AI.
- All cutting-edge AI is concentrated in private and public companies.

# Possible Solutions

# Possible Solutions

1. Slow down AI research (not feasible).

# Possible Solutions

1. Stop AI research (not feasible).
2. “CERN-AI” or “Telescope-AI” (feasible, but very hard).

→ Build the largest data/computing center in the world, with ~1K permanent staff, and ~1K affiliated academic labs (~3K scientists).

## The Need for New “AI Telescopes”

Pierre Baldi

University of California, Irvine

Artificial Intelligence (AI) raises two fundamental sets of questions: (1) how does the universe of intelligence look like and what is our place inside that universe? and (2) how we can ensure that AI is safe? For centuries, the approach our society has taken to address these kinds of difficult questions is to let universities and their scientists study them. However, today universities are unable to train and study the most advanced forms of Artificial Intelligence at scale, including large language models with trillions of parameters, such as GPT-4. Universities’ main problem is computational. No university has the in-house computational resources, primarily in the form of large clusters of Graphical Processing Units (GPUs) needed to train the most powerful AI models. Here we call for a very large-scale effort to address this dangerous situation.

A few people think this problem can be brushed away. There are still plenty of problems for academics to work on at the fringes—for instance by training much smaller models or fine-tuning and aligning existing models using specific data. Furthermore, this situation has happened before in the world of technology—after all, universities cannot build large ships, airplanes, or nuclear plants. Finally, the economic reality is that the cost and infrastructure required for building such models are prohibitive for academic institutions: GPT-4 is rumored to cost north of 100 million dollars to train over a period of many months using tens of thousands of GPUs. And the cost of systems in the GP-class seems to go up by one order of magnitude for each generation.

Although there is wisdom in these arguments about resources and scale, they fail to recognize the ambiguous nature of AI as being not only a formidable technology but also a fundamental topic in the natural sciences and the study of our place in the universe. Telescopes and microscopes taught us that we are not central to the universe of stars or of living systems. Computers are teaching us that we are not central to the universe of intelligence. And in the same way that the science of fluid dynamics is fundamental for understanding and building safe airplanes, the science of AI is fundamental for understanding and building safe AI systems. The stakes related to AI are so fundamental that it seems dangerous to leave them exclusively to companies. No one knows for sure, but we may be reasonably close to being able to create intelligence that is greater than human intelligence. So we want companies to be the first and only ones to go in, and exploit

**Build an international AI  
'telescope' to curb the  
power of big tech  
companies.**

Baldi P ,

Fariselli P ,

Parisi G

**Author information**

**Nature**, 01 Oct 2024,  
634(8035):782  
<https://doi.org/10.1038/d41586-024-03436-9> PMID:  
39438746



# Major Obstacles to Overcome

1. Cost and Cost-Sharing [1T over 15 years]
2. Facility (Energy/Hardware)
3. Relationship to Industry  
(Competition/Cooperation) and Lags
4. Data
5. Leadership/Organization/Decision Making
6. Output/Product
7. Safety

DeepSeek and Stargate did not change much.

1. AI Opportunities

2. AI Challenges

AI Telescope

AI Safety Framework

# **AI Safety from First Principles**

1. AI is inspired by NI (Natural Intelligence)

# **AI Safety from First Principles**

1. AI is inspired by NI (natural intelligence)
2. Can/should AI safety be inspired by NI safety?

1600: Giordano  
Bruno  
1633: Galileo  
Galilei




# NI Safe?

- Human history is a long list of wars (50-85M casualties in WW2)
- >10 wars in the 21<sup>st</sup> century alone
- Invention of increasingly more sophisticated methods of torture
- Invention of increasingly more sophisticated weapons, weapons of mass destruction
- Nearly 47,000 people died of gun-related injuries in the United States in 2023, according to the latest available statistics from the CDC.

# Parallels between NI and AI Safety

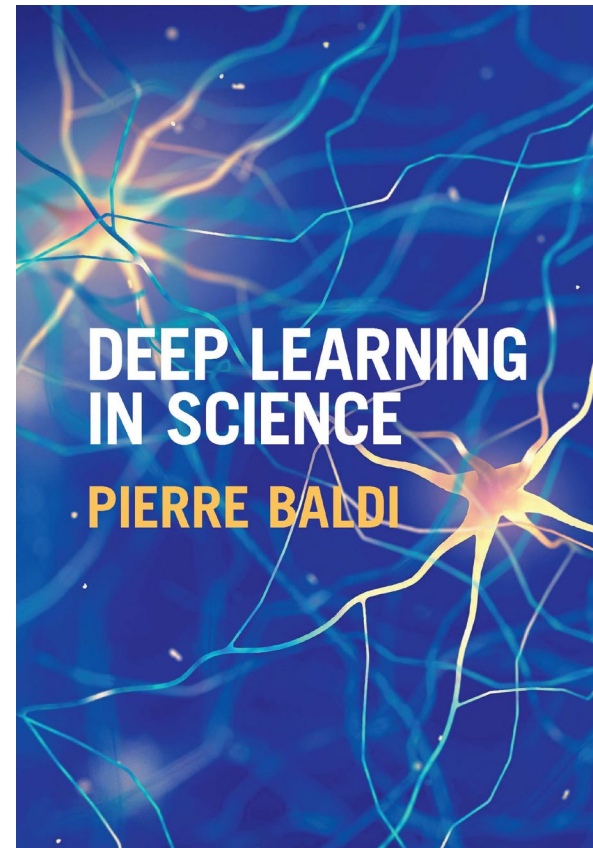
Natural Intelligence Safety		Artificial Intelligence Safety	
Evolution	Modular architectures, safety modules	Pre-Training	
Examples (parents, teachers, role models)	RL from Human Feedback (RLFH)		
Principles (e.g. 10 commandments)	Constitutional AI	Training	
Law	AI laws		
Societal	Agentic	Post-Training	
Enforcement (e.g., police, lie detectors)	Enforcement (e.g., police, fake detectors)		
Enforcement (e.g., military, WMD)	Enforcement (e.g. military, killer switches)	Deployment	



# **Additional Considerations**

1. New possibilities derived from framework.
2. The “failure” of evolution—two reasons.
3. Evolution used a multi-tier approach.
4. AI is different from NI and other principles/methods may apply.





Cambridge University Press

**THANK YOU**