A Statement of Opposition to SB-1047

from California AI Students and Faculty

We the undersigned are academic AI researchers—in the University of California at Berkeley, Davis, Los Angeles, Riverside, San Diego, Santa Barbara, and Santa Cruz, in the University of Southern California, and in Stanford University—who strongly oppose SB-1047. It will severely hinder our academic freedom and ability to conduct innovative research on AI, for the sake of a dubious attempt to mitigate hypothetical impacts of large language model (LLM) advances.

Coverage of the controversy around this bill has centered on corporate factions, “big tech vs little tech,” “doomers vs accelerationists,” or the impacts the bill will have on the AI industry. We agree that this bill will have broad negative consequences, hamper economic dynamism, and weaken California’s position as a global AI hub, in the service of questionable, unscientific, and hypothetical public benefit. However, the purpose of our letter is to offer a researcher-centric perspective in opposition to SB-1047.

As academics, we unequivocally oppose this bill for concern that it may seriously harm both the research and educational objectives of California universities in AI. We call on our representatives to seriously consider these harms when weighing whether to pass this legislation—industry actors are **not** the only ones who vociferously oppose this bill. Our deep concerns regarding the practical impacts and scientific validity of the bill are as follows:

I. On chilling effects for open-source model releases, to the detriment of our research

We believe requiring “safety auditing” and “full shutdown” capacities in “frontier models” will fundamentally hinder open-source and open-weight releases—proprietary models held and controlled by private entities can most easily fulfill these stringent requirements. The language around how safety will be demonstrated and audited is underspecified in the current text, reliant on future hypothetical tests which currently do not exist and may not be scientifically rigorous. Bearing the potential costs of these audits may be easy to justify for a commercial entity with a profitable product, but we are concerned the same cannot be said for scientifically oriented open releases from commercial entities such as Meta’s LLaMA series, or for open models trained by non-profits or consortia of universities. Further, it is unclear how the requirements put forth in the bill will be enforced, and many consequential terms throughout are poorly defined or left to the proposed “Frontier Models Division” to determine.

Considering these onerous restrictions, would-be open-weight model developers may choose to simply build their systems outside of California or the US and release their models under licenses that forbid their use in California to avoid liability. In this scenario private actors without regard for compliance may choose to covertly use the models despite the license restrictions, while academic researchers—bound to comply by the public nature of their work—will be shut out, incentivizing them to either change topics or move to jurisdictions which do not infringe on their academic freedom. Groundbreaking AI research relies on access to open models.

Having access to open models is *crucial* for modern academic AI research. Using open-weight models, academics investigate how models work, how they gain capabilities during training, and how they can be improved and broken [1]. Research on societally consequential issues in language modeling such as copyright infringing outputs [2], fairness and discrimination [3], and secure deployment with private data [4,5] use methods that require open models with documented training procedures—something which private APIs (*application programming interfaces*: closed and opaque paid-access online services) such as OpenAI's or Google's cannot provide. Without this access, such research simply cannot be performed in academia. It is important that this work be conducted *by academic researchers, in the open, for the benefit of the public*, rather than be relegated to trade secrets, trapped behind the confidential walls of private entities based on their proprietary models.

By imposing these stringent restrictions on open LLMs, the bill risks alienating the academic community that relies on these resources, forcing any frontier LM research which does take place to be performed on the few proprietary APIs which can comply. At best, this will just impose potentially prohibitive costs on this research, leaving it only to be performed by the few groups who can afford it. At worst, the LM API providers will have the power to shut out academic researchers performing investigations they don't like or hinder scientific reproducibility by removing access to older model versions. By forcing us to rely on commercial services rather than open-source frameworks, the bill will deprive students opportunities to experiment, learn, and contribute to AI research, without the excessive costs.

II. On the unscientific nature of AI risk forecasting and "capability" assessment

While we are concerned by the negative impacts this bill may have on our research, we are also concerned by its framing which presumes the certainty of hypothetical AI risks. To argue that meaningful risks from AI development—that distinguish them from any other dual-use emerging technology—exist is dubious and ultimately arises from gut opinions, not a scientific basis.

As experts in artificial intelligence, machine learning, and natural language processing, we must stress that both *the existence of meaningful AI risks* and *the proposed methods of assessing model risk* alluded to in SB-1047 are deeply dubious. There is *no consensus* in the scientific community on whether and how language models or other "frontier" AI systems may pose a threat to the public [6]. Additionally, the proposed use of "capability measurements"—arbitrary and fundamentally flawed benchmarks [7]—is deeply problematic. We do not have a rigorous understanding of how the "capabilities" measured by these tests are related to each other or if they are even meaningful in the real world. Using them as a proxy measure of risk is nonsensical and fundamentally unscientific.

As far as near-term risks addressed by the bill go, it is questionable whether a language model can act as an information-generating system that is any more empowering to a nefarious actor than a search engine—any would-be bioweapon developer can just as easily search Google or scour textbooks to enable "chemical, biological, radiological, or nuclear weapon" harms invoked in the bill [8]. Regarding the long-term, we do not believe that "existential" or "catastrophic risks" from AI development—the underlying beliefs driving engagement in the AI Safety community—are sufficiently evidenced to merit policy in response. We find it deeply troubling that a law with severe academic and industrial harms is built on such shaky intellectual ground.

This problematic frame is exemplified in Senator Weiner's response to earlier criticisms of SB-1047 leveled by Y Combinator and Andreessen Horowitz. He notes a point of common desire to "regulate AI for safety" between the bill's promoters and its critics, and that late feedback on the bill is a surprise to him as he posted an outline "in September 2023 for the purpose of soliciting early feedback" [9]. We would like to emphasize that **only those who already believe in existential risks of AI systems**—a group that, while containing some field luminaries, **does not hold a consensus position in the AI research community—would participate in discussions around safety legislation in its early stages**. As we reject the frame that the “risks” alluded to in the bill are real enough to warrant regulation, it is hard to offer “constructive criticism” to “improve the bill,” and we instead call on our representatives to seriously reconsider enacting such a bill at all. Any fact-based debate around this bill must start from the questionably factual nature of AI risk.

III. On the insufficiency of near-term carve outs for open-weight models

Some language in and around the bill has pointed out that no *current* open-source models will be “covered” under the law due to size or cost [9], that some future carve outs for them may be providedby the "Frontier Models Division,” and that some loose phrasing in the bill, such as references to models “under a developer’s control” regarding “full shutdown” capabilities, implies open models would be exempt.

Additionally, it is true that current state-of-the-art open-weight models have parameter counts and training costs orders of magnitude below the limits. However, given the exponential growth of parameter counts, and the decreasing cost of compute, there is no reason to expect this situation to hold. In the absence of iron-clad guarantees protecting open source—which the bill does not provide—we must assume that the consequences described above will come; if not today, in the near future. Additionally, low parameter-count models—inherently more usable by independent researchers and small labs due to their lower computational requirements—actually require *more* costly compute to train than larger models of equivalent performance [10], which would form better candidates for open release.

Thus, we do not believe the proposed amendments intended to mitigate effects on open-source and accessible model releases will mitigate the chilling effects we expect. While a covered model *can* be released under this law, the stringent reporting auditing and requirements under the “Frontier Model Division” that apply to each trained covered model will still needlessly impact our research activities.

IV. Concerns about job placements and career outcomes for our students

Finally, we believe SB-1047 will hinder the aspirations of prospective students interested in AI and computer science. These fields are among the most popular and rapidly evolving on our campuses, yet the bill's constraints might deter new talent from entering these crucial areas. Furthermore, with the tech industry's shift from big corporations to startups—especially noted in the recent trends of 2024—the additional regulatory hurdles could diminish the entrepreneurial spirit by favoring larger, better-resourced companies over emerging innovators. This shift might narrow their career pathways post-graduation. SB-1047 fundamentally interferes with the educational mission of our institution.

V. In conclusion

A Statement of Opposition to SB-1047
from California AI Students and Faculty

Many important scientific directions are impossible to pursue using commercial APIs—but in the near future, should this law pass, we may have no choice but to use them to investigate frontier models. Without the freedom to explore and push boundaries, the California academic community will fall behind in the global AI race, unable to lead in AI innovation or meaningfully participate in critical discussions on the ethical, safe, and effective uses of AI technologies. The recommended edits by Anthropic do not address any of the fundamental issues listed above.

After the signatures we provide a summary of each numbered reference above and how their findings relate to our argument. We hope our explanations will be accessible to non-expert legislators, journalists, and members of the public, and strongly encourage all readers to check them.

As students and faculty of globally-leading AI research institutions in California we believe that SB-1047 is fundamentally wrongheaded. It attempts to solve questionably real problems using scientifically unfounded methods, and will be detrimental to educational growth, scientific innovation, and economic development in AI for the University of California, the state of California, the United States, and the world. We call on our representatives to listen to experts—seriously consider our perspective on this matter.

The opinions in this letter are solely those of the undersigned and not our institutions.

University of California Faculty

William Wang
Mellichamp Professor of Artificial Intelligence
University of California, Santa Barbara

Ion Stoica
Professor
Electrical Engineering and Computer Sciences
University of California, Berkeley

Julian McAuley
Professor of Computer Science & Engineering
University of California, San Diego

Yi Zhang
Professor of Computer Science & Engineering
University of California, Santa Cruz

Taylor Berg-Kirkpatrick
Associate Professor
Computer Science & Engineering
University of California, San Diego

A Statement of Opposition to SB-1047
from California AI Students and Faculty

Joseph E. Gonzalez
Associate Professor
Electrical Engineering and Computer Sciences
University of California, Berkeley

Yue Dong
Assistant Professor
Computer Science & Engineering
University of California, Riverside

Muhao Chen
Assistant Professor
Department of Computer Science
University of California, Davis

Xin Eric Wang
Assistant Professor
Computer Science & Engineering
University of California, Santa Cruz

Alane Suhr
Assistant Professor
Electrical Engineering and Computer Sciences
University of California, Berkeley

Prithviraj Ammanabrolu
Assistant Professor
University of California, San Diego

Uri Manor
Assistant Professor
University of California, San Diego

University of California Graduate Students and Researchers

Xuandong Zhao
Postdoctoral Fellow
University of California, Berkeley

Michael Saxon
PhD student and NSF Graduate Fellow

University of California, Santa Barbara

Sanjay Subramanian
PhD student
University of California, Berkeley

Iain Weissburg
MS Student
University of California, Santa Barbara

Jessy Lin
PhD Student
University of California, Berkeley

Alon Albalak
Research Scientist, PhD
University of California, Santa Barbara

William Chen
PhD Student
University of California, Berkeley

Isadora White
PhD Student
University of California, San Diego

Xinyi Wang
PhD Candidate
University of California, Santa Barbara

Tanishq Mathew Abraham
CEO and Research Director, Ph.D.
MedARC, Stability AI, University of California, Davis

David Wang
Undergraduate Researcher
University of California, Santa Barbara

Tanmay Parekh
PhD candidate
UCLA

Stone Tao

PhD Student
University of California, San Diego

Reyna Abhyankar
CSE PhD Student
University of California, San Diego

Deepak Nathani
PhD Candidate
University of California, Santa Barbara

Daniel Rose
MS Student
University of California, Santa Barbara

Alfonso Amayuelas
PhD Student
UC Santa Barbara

Stanford Students

Cyril Zakka
Postdoctoral Scholar, MD
Stanford University

Aryaman Arora
Ph.D. Student
Stanford University

Anikait Singh
PhD Candidate
Stanford University

Zheng Wang
MS Student
Stanford University

University of Southern California Students
Isabelle Lee
PhD Student
University of Southern California

Tejas Srinivasan

PhD Candidate
University of Southern California

Shushan Arakelyan
PhD Candidate
University of Southern California

Concurring Academics and Researchers

Gunnar W. Knutsen
Professor, Dr. philos.
University of Bergen, Norway

Naomi Saphra
Kempner Research fellow
Harvard University

Kaiser Sun
PhD Candidate
Johns Hopkins University

Jack Hessel
Research Scientist, PhD
Samaya.ai

Ivan Bercovich
Tech Investor
ScOp Venture Capital

Quentin Anthony
Head of HPC, PhD
EleutherAI

Chris Lengerich
Founder
Context Fund

Vardhan Dongre
PhD Student
University of Illinois Urbana Champaign

Yiran Lawrence Luo
PhD Student

Arizona State University

Farhan Samir
PhD Student
University of British Columbia

Sat Chidananda
Student
Arizona State University

Yongchang Hao
PhD Student
University of Alberta

Blake Harrison
Computer Science PhD Student
Arizona State University

Saurabh Shah
ML Engineer
Apple

Boyuan Chen
PhD Candidate
MIT

Ethan Shen
Undergrad Student
University of Washington

Kevin Ayers
Machine Learning Engineer
Georgia Tech

Jim Jones
Engineer, UCSB Alum
Microsoft

Alex Lyman
PhD Student
Brigham Young University

Afra Feyza Akyürek

A Statement of Opposition to SB-1047
from California AI Students and Faculty

PhD Candidate
Boston University

Daniel Jimenez
ML Research Scientist
Johns Hopkins Applied Physics Laboratory

Carlos E. Jimenez
PhD Candidate
Princeton University

Justin T Chiu
Research Scientist, PhD
Cohere

Keshav Ramji
Master's Student
University of Pennsylvania

Vojta Jina
CTO
RAINN Inc

Annotated references.

We briefly describe each work referenced in the letter with blurbs motivating either the consequential findings it presented and how it was enabled by open access language models, or a short summary of the paper's role in supporting our argument. All have been published in top international AI venues.

[1] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, Oskar Van Der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. Proceedings of the 40th International Conference on Machine Learning, PMLR 202:2397-2430, 2023.

In this work, the authors present a set of open-source, fully documented language models of increasing sizes to facilitate rigorous research on the role that model scaling has on performance on any downstream task. This work has been tremendously empowering for academic researchers—without these open models, only OpenAI, Anthropic, Google, and their ilk would be able to perform this kind of inquiry.

[2] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, Chiyuan Zhang. Quantifying Memorization Across Neural Language Models. Proceedings of the International Conference on Learning Representations (ICLR) 2023.

In this work, the authors analyze "memorization" of training data in language models that have been trained on documented training data. Work in this vein enables insights into whether and how language models learn to copy and reproduce training data—with consequential implications for copyright and ethical use. While this work was produced by Google, use of non-proprietary (ie., open-source, open-weight) models is necessary to render this kind of work releasable by private entities, and to make it reproducible

[3] Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, Xiang Ren. "On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning." *NAACL* 2021

In this work, the authors investigate and demonstrate fine-tuning techniques to reduce bias against protected groups for hate speech detection and text classification tasks. They show how debiasing fine-tuning for one task can be transferred to another. Fine-tuning is a technique which requires access to a model's weights in order to be performed reproducibly and inexpensively and is a fundamental method for language model modification after pretraining.

[4] Kandpal, Nikhil, Eric Wallace and Colin Raffel. "Deduplicating Training Data Mitigates Privacy Risks in Language Models." *ArXiv* abs/2202.06539 (2022): n. pag. ICLR 2022

*This group of academic researchers, using open language models trained on documented corpora, demonstrate the impact that duplicated examples in training data has on memorization of **private information** in language models. These findings have significant implications—private data such as phone numbers and email addresses are inevitably picked up in the incomprehensibly large datasets language models are trained on, so it is important to investigate how to mitigate the privacy risks this entails. Academic researchers are unbound by the incentives private companies have to not reveal how their systems may leak private information—and academics can only perform this research with open access to models with documented training data.*

[5] Kim, Siwon, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sung-Hoon Yoon and Seong Joon Oh. "ProPILE: Probing Privacy Leakage in Large Language Models." *ArXiv* abs/2307.01881 (2023): n. pag. NeurIPS 2023

*In this work, the authors propose a method to "probe" for personally identifiable information that is memorized by language models, and they **verify that their method works using an open language model trained on "the Pile"**, an open and fully documented training dataset for language modeling. **Only with direct access to the training data and a fully transparent model can these important and consequential methods be verified.***

[7] Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, Naomi Saphra. “Benchmarks as Microscopes: A Call for Model Metrology.” ArXiv abs/2407.16711: COLM 2024.

In this work, the authors summarize and survey the broadly agreed-upon notion among academic AI researchers that ***current benchmarks and metrics which are purported to assess generalized AI capabilities (****such as those that form a basis for whether a model is “covered” under SB1047)* ***are fundamentally broken and not meaningfully predictive of how a model will behave in deployment settings****.*

[10] Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu and Dario Amodei. “Scaling Laws for Neural Language Models.” ArXiv abs/2001.08361 (2020): n. pag.

This (non-peer reviewed, but widely accepted and replicated) technical report from respected OpenAI scientists demonstrates “scaling laws” between compute cost, performance, and model size for LM training. The headline finding is that, counterintuitively, it is actually ***less computationally expensive to train a large LM to some level of performance than a small one****, meaning that training an easier-to-use and smaller performant LM would actually be more costly than a large and difficult-to-use one (for academic researchers). This could potentially lead to a situation where an otherwise not-covered model at large and unreleasable sizes becomes covered when trained as a smaller one.*

Non-academic references.

[6] Arvind Naraynan and Sayash Kapooor. “AI existential risk probabilities are too unreliable to inform policy.” AI Snake Oil, 2024.

In this work, these academics convincingly argue that the “p(doom)” (belief that an existentially dangerous AI system will come into existence based on current trends) which are bandied about by AI safety fear merchants is fundamentally unscientific—there is no way to meaningfully assign a probability to the impact of speculative future technology, so such fears should not form the basis of public policy, particularly policies such as SB1047 which have severe externalities.

[8] Mark Scott, Gian Volpicelli, Mohar Chatterjee, Vincent Manancourt, Clothilde Goujard, Brendan Bordelon. “Inside the shadowy global battle to tame the world’s most dangerous technology.” Politico, March 2024 [**link**]

Consider the following excerpt:

> *In a private hearing between U.S. lawmakers and tech experts in September, [Tristan] Harris, a co-founder of the Center for Humane Technology, a nonprofit, described how his engineers had coerced Meta’s latest AI product into committing a terrifying act: the construction of a bioweapon.*

A Statement of Opposition to SB-1047
from California AI Students and Faculty

> *Mark Zuckerberg's tech giant favored so-called open-source technology — AI easily accessible to all — with few safeguards against abuse. Such openness, Harris added, would lead to real-world harm, including the spread of AI-generated weapons of mass destruction.*
>
> *His triumph didn't last long. Zuckerberg, who was also present at the Capitol Hill hearing, quickly whipped out his phone and found the same bioweapon information via a simple Google search. Zuckerberg's counterpunch led to a smattering of laughter from the room. It blunted Harris' accusation that Meta's open-source AI approach was a threat to humanity.*

*This anecdote is illustrative of our contention that the dangers of "AI technologies" as aids for nefarious actors is questionable—the burden of proof that synthetic text generating systems such as LLMs are more dangerous than search engines with respect to bioweapons and the like is on the claimants—such evidence has **not** been provided yet.*

[9] Senator Scott Weiner, "Response to inaccurate, inflammatory statements by Y Combinator & a16z regarding Senate Bill 1047" 2024 [**link**]

From the assumed frame of shared belief in existential AI risks, Senator Weiner points out inaccuracies and misrepresentations in some early statements from Y Combinator and a16z representatives. While we agree that some language in the aforementioned statements, such as sensationalized claims that model makers will be jailed and the like, are inflammatory and unhelpful, we think it is important to point out that the presumed frame of a desire to pass legislation to mitigate questionably valid AI risks which Sen. Weiner references is not shared, with us or presumably with the authors of the original statements. Further, while we may disagree with the authors in specific statements and do not endorse all claims they make about the impacts of SB-1047, we fundamentally share concern toward the consequences of the bill and endorse their efforts to ensure it is not enacted.

Related letters.

Fei-Fei Li's statement against SB-1047 [**LINK]** echoes our concerns over the impact of this bill on academia.

We concur with the opinions put forth in the many other open letters and statements from non-academic groups in opposition to the bill.

Delivered on behalf of all signatories by representatives of the Natural Language Processing Group in the Department of Computer Science of the University of California, Santa Barbara.

UCSB NLP Group
Henley Hall, University of California
Santa Barbara, California, 93106

https://nlp.cs.ucsb.edu/
Michael Saxon (saxon@ucsb.edu)
William Yang Wang (william@cs.ucsb.edu)