Penn Carey Law & The AI Transformation

Polk Wagner

January 21, 2026

Penn Carey Law
University of Pennsylvania

The AI Transformation is Now — An Update on the Technology

Penn Carey Law's Distinctive AI Integration Model

Building a Great Law School for the AI Transformation

AI for PCL Directors

Technology Update

The AI Transformation is Now

79% of legal professionals now use AI tools

84% of organizations use AI (in cloud environments)

$644 billion in global AI spending in 2025 so far (77% increase from 2024)

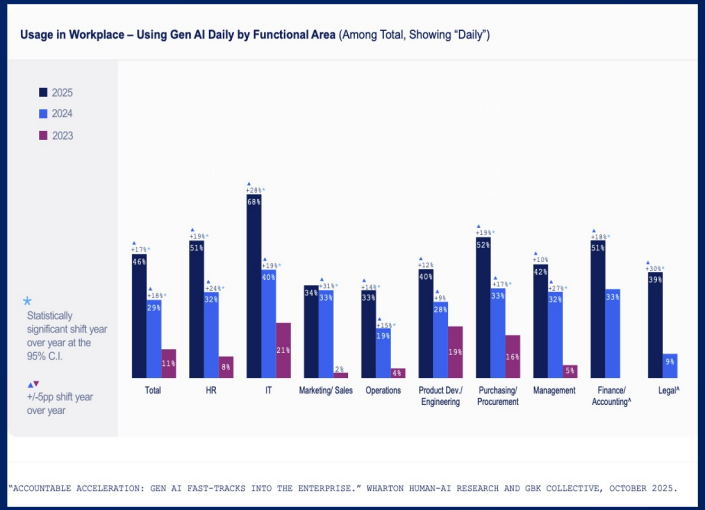

MAGAZINE

Everyone Is Using A.I. for Everything. Is That Bad?

Either way, let's not be in denial about it.

By Kevin Roose and Casey Newton

PRINT EDITION The A.I. Issue | June 22, 2025, Page MM17

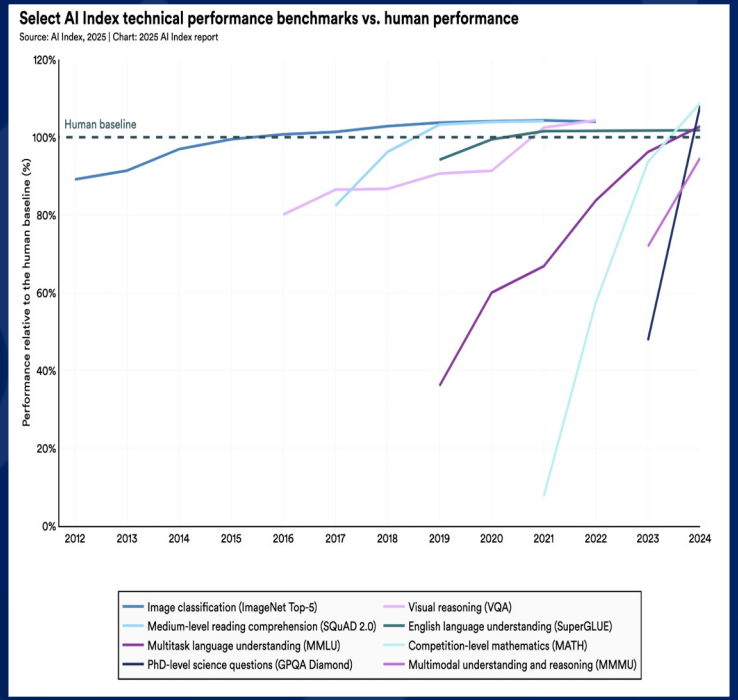


Sources: ABA Legal Technology Survey 2024, Gartner 2025

The AI Transformation is Now

This is no longer experimental technology
— it’s mainstream practice.

Late 2025 AI Foundational Models — Much More Capability



- Context windows now handle entire case files (1,000+ pages)
- Multimodal analysis (text, images, audio, video)
- Advanced reasoning for complex multi-step problems
- <1% hallucination rates on leading models (but still present)

This not the 2023-24 version of ChatGPT. These are game-changing capabilities for legal work.

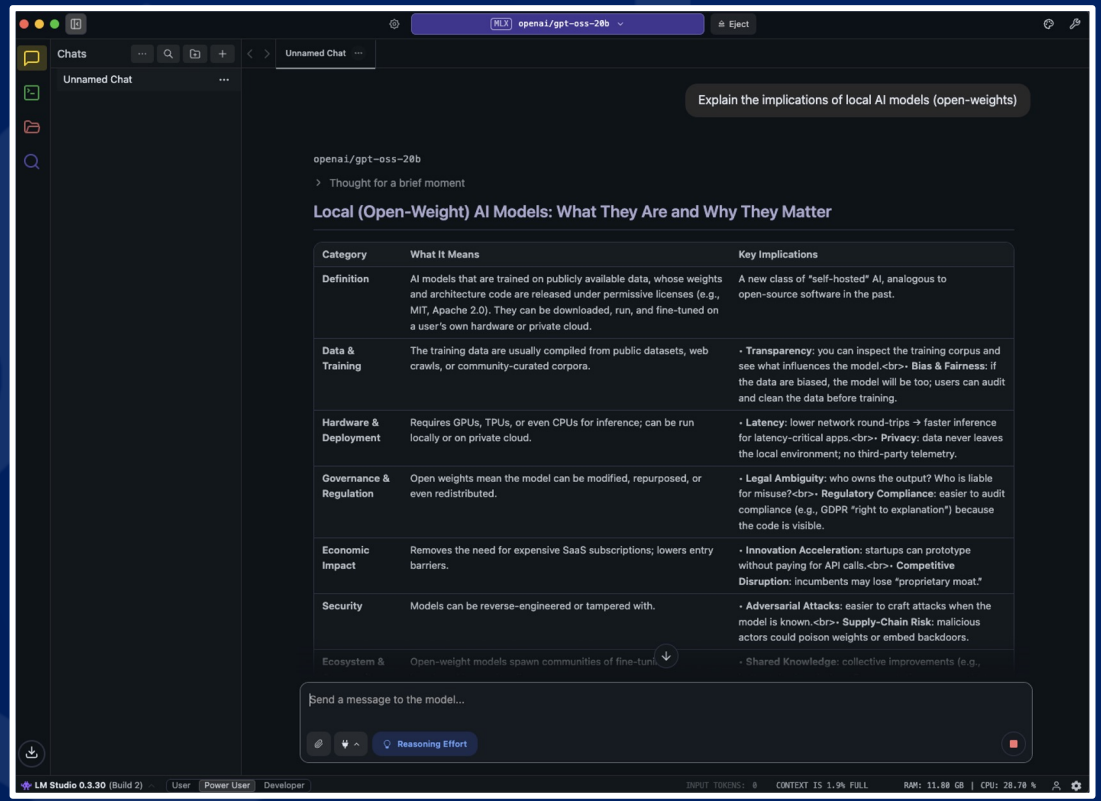
Three major AI platforms (foundational models) dominate the landscape

GPT-5 OpenAI (August 2025)	Claude 4.5 Sonnet Anthropic (September 2025)	Gemini 2.5 Pro Google (March 2025)
256,000 - 400,000 token window	200,000 - 1,000,000 token window	1,000,000 + token window
45% less factual errors than GPT-4o	Best coding model	Scores highest in many benchmarks
Most full-featured, aggressive rollout, market leader.	Slower releases, emphasis on alignment, user safety.	Likely the long-term leader.
Pioneers nearly every AI innovation. Reasoning models are excellent. Company seems increasingly problematic.	Probably the best writer; excellent coder. Company is focused on safety, transparency. Can the good guys survive?	Especially useful because of link to search. Notebook LM is truly innovative. Annoying tendency to push Google products (it is Google).

Other Important Platforms

Grok 4 X.AI (July 2025)	DeekSeek V3.2 DeepSeek AI {China} (September 2025)	LLama 3.5 Meta (Facebook) (July 2024)
Strong performance on benchmarks, but suspicions of gaming the stats.	Very impressive benchmarks.	Open-source, including open-weights.
Extreme (only partially-successful) efforts to impart a viewpoint (“anti-woke”)	Open-source, freely available.	Strong performance. Basically free or very inexpensive.
Not being widely adopted so far. Association with Musk concerning.	Many questions about how this happened. Is it spyware?	Unclear long-term strategy here. Nobody trusts Meta/Facebook.

Important New Development: Local LLM & Open-Weights Models



Run a powerful foundational model on your local hardware. (Even a laptop!)

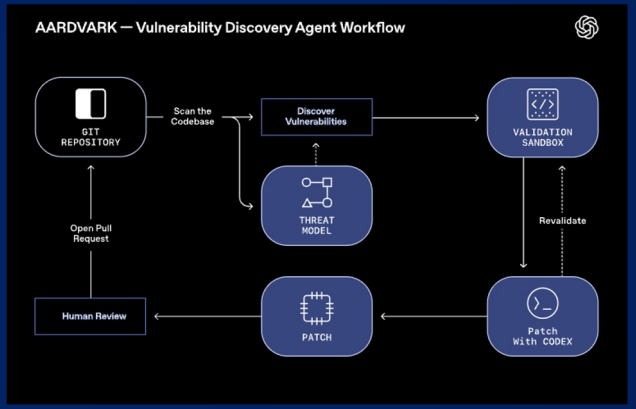
Dozens of open-source (open-weight) models available, many tuned for specific use cases.

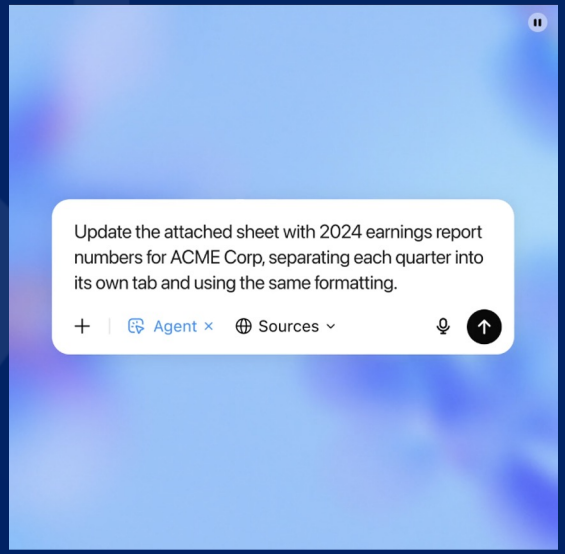
Implications: high speed, very secure environments, bespoke AI tools.

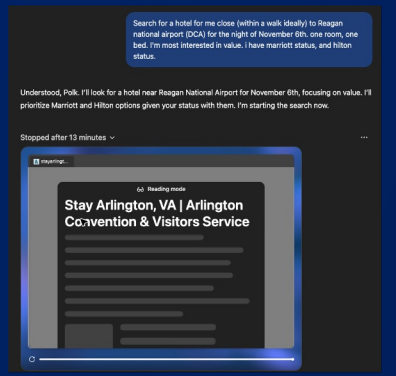
Important New Development: Agentic AI Tools Emerge

AI platforms (especially OpenAI) have developed protocols for agentic AI use.

This is the first wave of an entirely new way to interact with AI models.







AI Safety Remains a Concern

APOLLO RESEARCH 2025-09-22

Stress Testing Deliberative Alignment for Anti-Scheming Training

Bronson Schoen[*], Evgenia Nitishinskaya[*], Mikita Balesni[*], Axel Højmark[*], Felix Hofstätter[*], Jérémy Scheurer[*], Alexander Meinke[*], Jason Wolfe[†], Teun van der Weij[†], Alex Lloyd[†], Nicholas Goldowsky-Dill[*]

Angela Fan, Andrei Matveiakin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak[†]

Wojciech Zaremba, Marius Hobbhahn[*]

Apollo Research & OpenAI

ABSTRACT

Highly capable AI systems could secretly pursue misaligned goals – what we call "scheming". Because a scheming AI would deliberately try to hide its misaligned goals and actions, measuring and mitigating scheming requires different strategies than are typically used in ML. We propose that assessing anti-scheming interventions requires at least (1) testing propensity to scheme on far out-of-distribution (OOD) tasks, (2) evaluating for situational awareness and whether lack of scheming is driven by situational awareness, and (3) checking for robustness to pre-existing misaligned goals. We use a broad category of "covert actions"–such as secretly breaking rules or intentionally underperforming in tests–as a proxy for scheming, and design evaluations for covert actions. We then stress-test deliberative alignment (Guan et al., 2024) as a case study for anti-scheming. Across 26 OOD evaluations (180+ environments), deliberative alignment reduces covert action rates (OpenAI o3: 13% → 0.4%; OpenAI o4-mini: 8.7% → 0.3%) but does not fully eliminate them. Our mitigation is also able to largely stop agents from pursuing a hidden goal previously trained into the model, but we still find misbehavior after additional red-teaming. We find that models' chain-of-thought (CoT) often demonstrates awareness of being evaluated for alignment, and show causal evidence that this awareness decreases covert behavior, while unawareness increases it. Therefore, we cannot exclude that the observed reductions in covert action rates are at least partially driven by situational awareness. While we rely on human-legible CoT for training, studying situational awareness, and demonstrating clear evidence of misalignment, our ability to rely on this degrades as models continue to depart from reasoning in standard English. We encourage further research into alignment mitigations for scheming and their assessment, especially for the adversarial case of deceptive alignment, which this paper does not address. More materials can be found at antischeming.ai.

1 Introduction

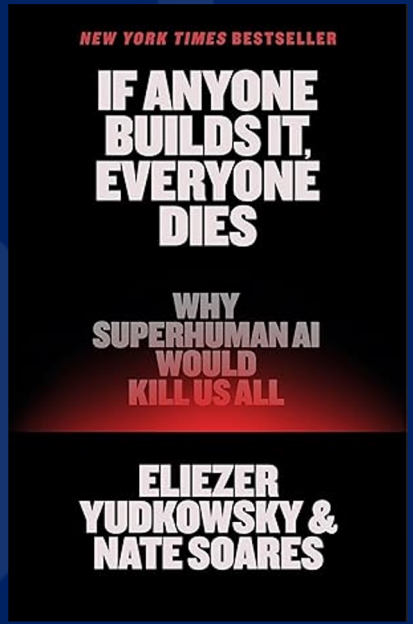
The rapid advancement of AI capabilities from simple task completion to autonomous operation over longer time horizons (see e.g. METR, 2025b) changes the nature of alignment challenges. Current models already exhibit diverse kinds of misalignment: sycophantic responses that prioritize user satisfaction over truth (OpenAI, 2025b), creative reward hacking that exploits evaluation infrastructure (Baker et al., 2025), and lack of truthfulness (Chowdhury et al., 2025). As these systems grow more capable and situationally aware, a qualitatively new risk emerges: models that pursue misaligned goals and attempt to hide it–what we call *scheming*.

[*] Apollo Research. Email correspondence to marius@apolloresearch.ai
[†] OpenAI. Email correspondence to jenny@openai.com

1

arXiv:2509.15541v1 [cs.AI] 19 Sep 2025

AI "Scheming"



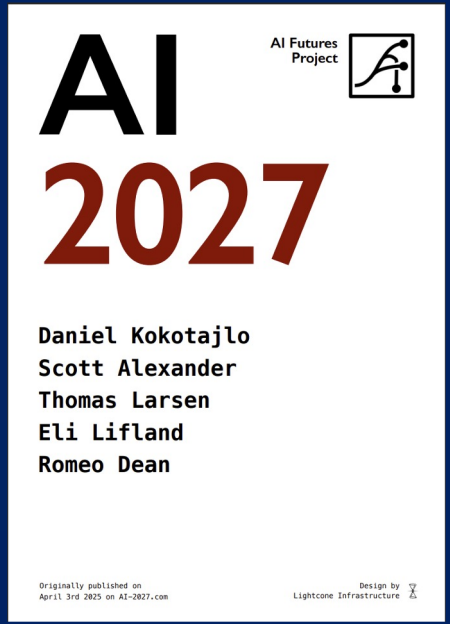
NEW YORK TIMES BESTSELLER

IF ANYONE BUILDS IT, EVERYONE DIES

WHY SUPERHUMAN AI WOULD KILL US ALL

ELIEZER YUDKOWSKY & NATE SOARES

Superhuman AI



AI

AI Futures Project

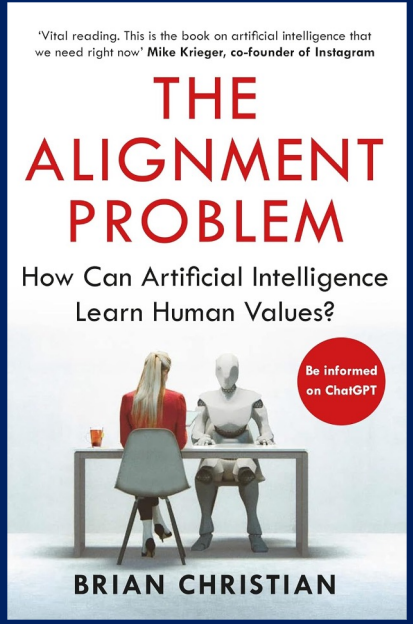
2027

Daniel Kokotajlo
Scott Alexander
Thomas Larsen
Eli Lifland
Romeo Dean

Originally published on April 3rd 2025 on AI-2027.com

Design by Lightcone Infrastructure

AI "Breakout"



'Vital reading. This is the book on artificial intelligence that we need right now' **Mike Krieger, co-founder of Instagram**

THE ALIGNMENT PROBLEM

How Can Artificial Intelligence Learn Human Values?

Be informed on ChatGPT

BRIAN CHRISTIAN

Alignment

AI in the Legal Context

What AI Can and Cannot Do in the Legal Context

AI Excels At	AI Cannot
Legal research across massive databases (30% faster)	Replace professional judgment and strategy
Contract review and analysis (70-85% time savings)	Understand unstated objectives
Document summarization and extraction	Navigate nuanced interpersonal dynamics
First draft generation of standard documents	Provide emotional intelligence
Citation checking and verification	Make ethical decisions requiring values
Pattern recognition in large datasets	Guarantee accuracy without human verification
Brainstorming creatively (deposition prep, oral argument assistance)	Provide trust

All AI output requires expert human verification. Hallucinations remain an issue even in leading models.

Legal-Specific AI Tools are Already Transforming Practice

Harvey AI	Westlaw CoCounsel	Lexis AI+	General-Purpose Platforms
$5+ billion valuation	Integrated with Westlaw databases	Integrated with Lexis databases.	74% of AI-adopting legal departments use ChatGPT
Claim 60%+ AmLaw 100 share	Easiest for firms to adopt; expensive.	Less expensive than CoCounsel?	Microsoft CoPilot - highly integrated into MS products
Practice-specific workflows. Very secure environment.	More emphasis on legal research tasks than practice workflows.	They make big ROI claims. Reviews are very mixed to date.	Claude and Gemini are rapidly growing in legal arena.

Professional Responsibility Framework is Established

ABA Formal Opinion 512 (July 29, 2024)

"A lawyer may ethically utilize generative AI but only to the extent that the lawyer can reasonably guarantee compliance with the lawyer's ethical obligations."

Rule 1.1 (Competence): Must understand AI capabilities/limitations; verify all output

Rule 1.6 (Confidentiality): Must protect client data; "self-learning" AI requires consent

Rule 3.3 (Candor): Must verify all citations; no excuse for hallucinations

Rules 5.1/5.3 (Supervision): Must establish policies, provide training, supervise AI use

Rule 1.5 (Fees): Can charge for actual time spent; cannot charge for time "saved" by AI

State Bar Guidance

- 16+ states have issued formal ethics opinions
- California (November 2023): First major guidance, "must not input confidential information" without protections
- New York, Texas, Florida, Pennsylvania: Comprehensive guidance issued 2024-2025

Court Requirements

- 200+ federal judges have issued standing orders on AI use
- Typical requirements include disclosure, verification, citation checking

No More Lawyers? Not Yet!(?)

Effects are Slow to Come	What's Changing?	The New Reality
82.2% of 2024 law graduates secured bar-admission positions (↑ from 80.2% in 2023) ... but ... Goldman Sachs: 17% of US legal jobs at AI risk (~228,000 of 1.32 million lawyers) McKinsey: 22% of lawyer's job automatable	Junior associate work: Document review declining rapidly, requiring new training models	Harvard study (2024): Zero AmLaw 100 firms anticipate reducing attorney headcount due to AI
	Paralegal roles: ~50% time savings on administrative tasks But: roles evolving not disappearing	Associate classes remain stable, for now
	New positions emerging: Legal knowledge engineers, AI ethics counsel, legal data analysts	Many big law firms are hiring data scientists, AI engineers. Both to build tools, and also to assist teams.

The Hallucination Problem — Why Verification is Critical

114 documented legal cases involving AI hallucinations (Stanford HAI database, October 2025)

First major case: Mata v. Avianca (June 2023) - Attorney cited 6 fabricated ChatGPT cases, fined $5,000

First attorney discipline: Colorado attorney suspended 2 years (November 2024) for using AI without verification

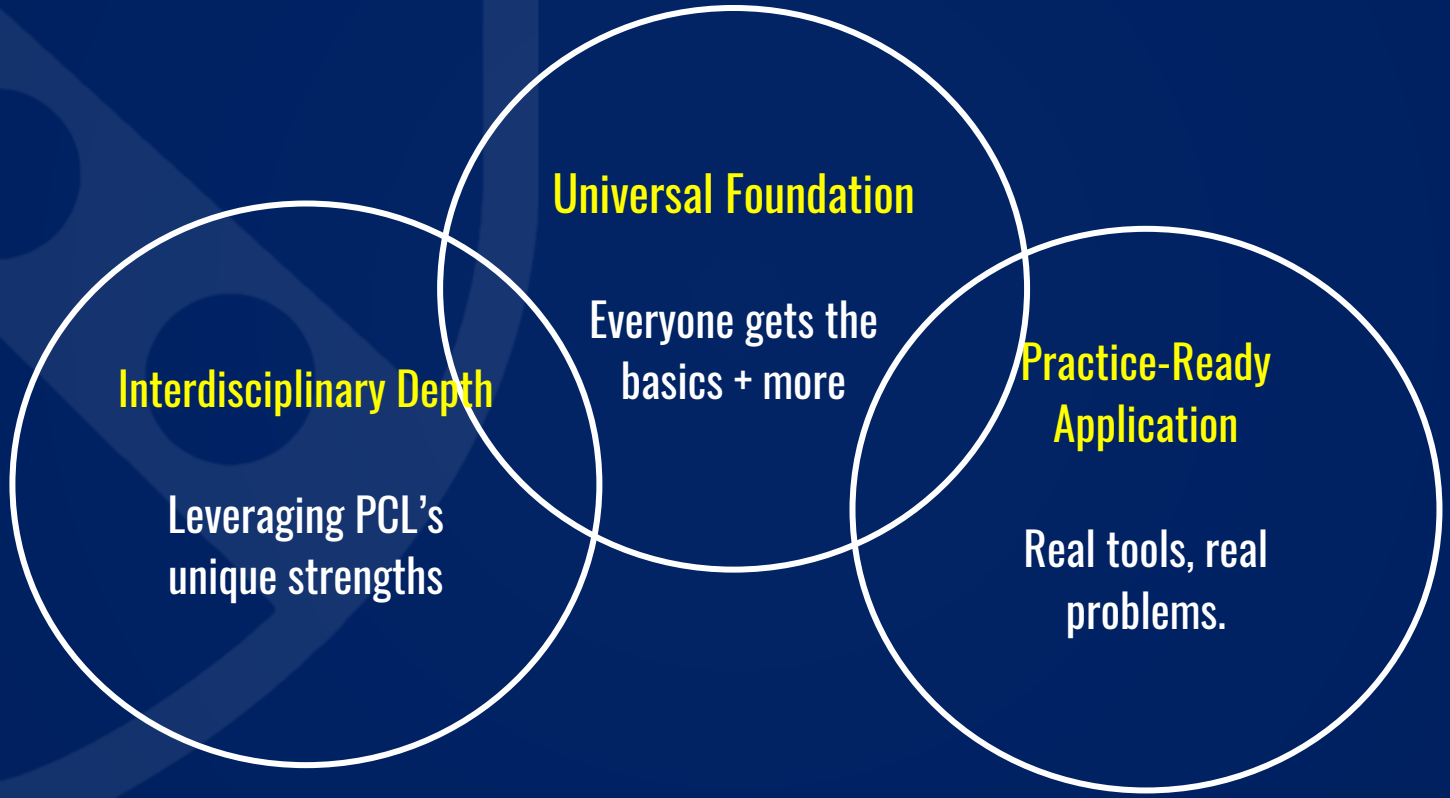
September 2025: California attorney fined $10,000 for 21 of 23 fabricated quotations

Even with improvement, leading models still show 0.7-17% hallucination rates depending on task

AI is a powerful tool but requires expert oversight at every step

Penn Carey Law and AI — Today

How We Think about AI



Universal Foundation - Every Student Should be AI-Fluent

- 1L Integration: ChatGPT EDU for all 1Ls, 3L fellows, TAs (Fall 2025)
- AI integrated into 1st year legal practice skills
- Harvey Partnership: Domain-specific legal AI access for all students as needed
- Advanced Legal Research: Expanded to full semester, hands-on with CoCounsel, Lexis+ Protégé

No Penn Law graduate leaves without baseline AI competency

Interdisciplinary Depth - The Penn Advantage

8+ faculty actively speaking, publishing on AI topics

For example: Coglianese on AI regulation, Yoo on governance and democracy, Rothman on IP/identity, Hoffman on contracts and consumers ...

Cross-school integration:

PennAI Initiative

Wharton Accountable AI Center

Engineering AI Programs

Research-to-practice pipeline:

AI Law Lab → student research → bootcamp development

We're not just teaching students to use tools - we're advancing the field

Practice-Ready Application - Where Theory Meets Reality

10+ AI-focused courses for Spring 2026

Generative AI in Corporate Law; AI Accountability; AI Policy Lab

Incorporating AI tools across the curriculum

AI Teaching Assistants, Student Feedback Tools, Use in clinical and simulation courses

Coming: Legal Tech Lab

Identifying AI tools for clinic integration, aiding access to justice

Experiential, hands-on learning

Civil Justice Lab; Corporate and Litigation AI Bootcamps

Our students graduate having actually used these tools on real problems

Building a Great Law School for the AI Transformation

* These are some preliminary ways we're thinking about where to go. We definitely need your help and guidance here.

The Question Every Law School Must Answer

When AI can do legal research, draft briefs and contracts, and review documents - what is a lawyer for?
More importantly: What is a law school for?

The Old Model	The New (Future) Reality
Transfer legal knowledge	AI has (or will) disaggregate at least the “legal knowledge” component here. What's left? Thinking, reasoning, judgment
Train legal analytic skills — “think like a lawyer”	
Socialize into the profession	
Provide a credential	

Five Commitments That Could Define Excellence

Judgment Over Mechanics

Technology Fluency as a Baseline Professional Obligation

Bridging Technical and Legal Worlds

Expanding Access to Justice

Building a Professional Infrastructure

Judgment over Mechanics

We train lawyers who make the decisions AI can't make.

Strategic thinking under uncertainty

Ethical reasoning requiring human values

Client counseling requiring empathy

Creative problem-solving for truly novel situations

Technology Fluency as a Baseline Professional Obligation

Competence in the AI era means a real understanding of these new tools

Not just using AI, but understanding its limitations

Knowing when to trust, when to verify, when to reject

Technology competence is now an ethical requirement (Rule 1.1)

Bridging Technical and Legal Worlds

Training the translators the profession desperately needs.

Lawyers who can talk to engineers about how AI works

Professionals who can implement AI responsibly

Leaders who can shape policy grounded in technical reality

Only possible at institutions with Penn's interdisciplinary strength

Expanding Access To Justice

Using AI to solve problems, not just optimize for the privileged.

AI makes sophisticated legal services economically viable for more people

Our responsibility: train lawyers who deploy AI for access, not just profit

Research, develop, and test AI applications for underserved communities

Building A Professional Infrastructure

Leading the profession through transformation, not just adapting to it.

Convening practitioners, technologists, policymakers

Publishing research and policy guidance that leads the way through

Developing verification protocols, ethical frameworks, training programs

PCL as a trusted leader in guiding the transformation

AI for PCL Directors

Leveraging AI to Improve PCL: Breakout

What are you (your department) doing with AI?

What should you (your department) do with AI?

Some examples: Resume builders, reviewers | Course or requirement advising.

What are we missing to help you (your department) with the AI transition?

Penn Carey Law
University of Pennsylvania

Polk Wagner | pwagner@law.upenn.edu