Supreme
Court

Benchmarking Justice: Can Advanced AI Satisfy the Rule of Law Standards?

Ian Bernaziuk, Doctor of Law, Professor
Judge of the Administrative Cassation Court of the Supreme Court

KYIV-MOHYLA ACADEMY

RULE OF LAW AND AI CHALLENGES
SCIENTIFIC WORKSHOP

20 November 2025

AGENDA

1. Introduction: AI, the Rule of Law & Three Core Rules for Judicial Use of AI
2. Judicial Ethics: Article 16 and Commentary
3. Supreme Court’s Work on AI Governance
4. Core Principles of Safe AI Use in Courts
5. Global AI Advancements: Gemini 3 and Other Leading Models
6. Four Benchmarks for Testing AI in Justice:

- Logic (deep reasoning)
- Accuracy (fact-grounding)
- Scale (large-volume evidence search)
- Precision (mathematical correctness)

7. Risks, Limits, and Responsible Adoption
8. Concluding Reflections

THREE BASIC RULES FOR WORKING WITH AI TODAY

1. Be Fully Involved

When you use AI, stay active. Think carefully about what you ask, guide the system, and pay attention to every step. AI depends on the quality of your questions and instructions. It can support your thinking, but it cannot replace your own judgment or attention.

THREE BASIC RULES FOR WORKING WITH AI TODAY

2. Do Not Trust It Completely

Even when AI sounds confident, it can be wrong. It may give inaccurate facts, unclear explanations, or even invented details. In law, accuracy is essential—so always double-check information, verify sources, and question every answer. Healthy skepticism keeps you safe.

THREE BASIC RULES FOR WORKING WITH AI TODAY

3. AI Is a Tool, Not a Decision-Maker

AI does not make decisions or take responsibility. Think of it as a tool that expands what you can do: faster research, new ideas, quick summaries, different viewpoints. But the final judgment is always yours. When used wisely, AI becomes an extra strength, not a substitute for your own work.

AI USE IN UKRAINE: KEY FACTS

https://unn.ua/en/news/in-ukraine-42percent-of-adults-and-70percent-of-teenagers-use-ai-study

42% of adults and 70% of teenagers in Ukraine use AI tools regularly for writing, studying, working, and searching for information.

50% of adults and 76% of teenagers have made at least one decision based on AI-generated results.

Figures come from the Digital Skills Study by the Eastern Europe Foundation, shared by Ukraine’s Ministry of Digital Transformation (Nov 2025).

CODE OF JUDICIAL ETHICS (Article 16)

https://zakon.rada.gov.ua/rada/show/en/n0001415-24?lang=en#Text

Use of artificial intelligence technologies by a judge is permissible provided that it:

– does not affect the judge's **independence or impartiality**,

– does not involve the **assessment of evidence**,

– does not interfere with the **decision-making process**, and

– does not violate any **legal provisions**.

COMMENTARY ON ARTICLE 16 OF THE CODE OF JUDICIAL ETHICS

https://rsu.gov.ua/ua/news/u-radi-suddiv-ukraini-vidbulosa-trete-rozsirene-zasidanna-rg-z-pidgotovki-komentara-do-kse

Currently, the Council of Judges of Ukraine **is finalizing** a new Commentary on the Code of Judicial Ethics.

The forthcoming Commentary is expected to provide **authoritative guidance on the permissible scope** of AI use by judges under Article 16, and to establish best practices for security safeguards and oversight mechanisms.

SUPREME COURT’S WORK ON AI RULES

https://constitutionalist.com.ua/poperednij-proiekt-19-11-2025-polozhennia-pro-vykorystannia-tekhnolohij-shi-pratsivnykamy-aparatu-vs

The Supreme Court is creating clear rules for safe and correct use of AI.

The goal is to help staff use new tools while **protecting independence, trust, and confidentiality**.

The Court sees AI as part of its modern work and is preparing simple, practical rules for daily use.

MAIN IDEAS OF THE DRAFT AI POLICY

https://constitutionalist.com.ua/poperednij-proiekt-19-11-2025-polozhennia-pro-vykorystannia-tekhnolohij-shi-pratsivnykamy-aparatu-vs

AI is only a **helper tool**; people make all **final decisions**.

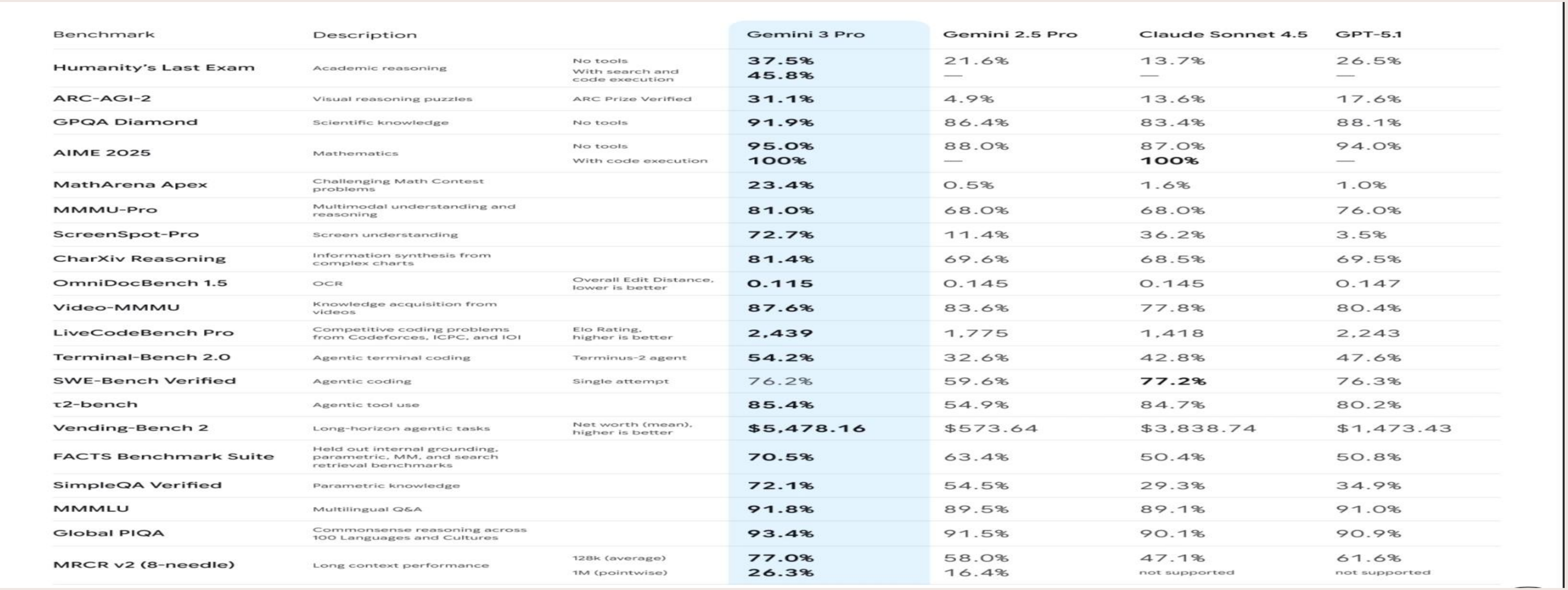
Staff must check all AI results and stay responsible for accuracy.

Public AI tools **cannot be used** for confidential or case-related information.

Only trusted, secure, and unbiased AI tools may be used for allowed tasks.

A NEW ERA OF INTELLIGENCE WITH GEMINI 3

https://blog.google/products/gemini/gemini-3/#note-from-ceo



Benchmark	Description		Gemini 3 Pro	Gemini 2.5 Pro	Claude Sonnet 4.5	GPT-5.1
Humanity's Last Exam	Academic reasoning	No tools With search and code execution	**37.5%** **45.8%**	21.6% —	13.7% —	26.5% —
ARC-AGI-2	Visual reasoning puzzles	ARC Prize Verified	**31.1%**	4.9%	13.6%	17.6%
GPQA Diamond	Scientific knowledge	No tools	**91.9%**	86.4%	83.4%	88.1%
AIME 2025	Mathematics	No tools With code execution	**95.0%** **100%**	88.0% —	87.0% **100%**	94.0% —
MathArena Apex	Challenging Math Contest problems		**23.4%**	0.5%	1.6%	1.0%
MMMU-Pro	Multimodal understanding and reasoning		**81.0%**	68.0%	68.0%	76.0%
ScreenSpot-Pro	Screen understanding		**72.7%**	11.4%	36.2%	3.5%
CharXiv Reasoning	Information synthesis from complex charts		**81.4%**	69.6%	68.5%	69.5%
OmniDocBench 1.5	OCR	Overall Edit Distance, lower is better	**0.115**	0.145	0.145	0.147
Video-MMMU	Knowledge acquisition from videos		**87.6%**	83.6%	77.8%	80.4%
LiveCodeBench Pro	Competitive coding problems from Codeforces, ICPC, and IOI	Elo Rating, higher is better	**2,439**	1,775	1,418	2,243
Terminal-Bench 2.0	Agentic terminal coding	Terminus-2 agent	**54.2%**	32.6%	42.8%	47.6%
SWE-Bench Verified	Agentic coding	Single attempt	76.2%	59.6%	**77.2%**	76.3%
τ2-bench	Agentic tool use		**85.4%**	54.9%	84.7%	80.2%
Vending-Bench 2	Long-horizon agentic tasks	Net worth (mean), higher is better	**$5,478.16**	$573.64	$3,838.74	$1,473.43
FACTS Benchmark Suite	Held out internal grounding, parametric, MM, and search retrieval benchmarks		**70.5%**	63.4%	50.4%	50.8%
SimpleQA Verified	Parametric knowledge		**72.1%**	54.5%	29.3%	34.9%
MMMLU	Multilingual Q&A		**91.8%**	89.5%	89.1%	91.0%
Global PIQA	Commonsense reasoning across 100 Languages and Cultures		**93.4%**	91.5%	90.1%	90.9%
MRCR v2 (8-needle)	Long context performance	128k (average) 1M (pointwise)	**77.0%** **26.3%**	58.0% 16.4%	47.1% not supported	61.6% not supported

Core Skill	Benchmark Test	Top Score	Legal Application (Value for Lawyers)
LOGIC	Humanity’s Last Exam	45.8%	Deep Legal Reasoning: Drafting complex arguments and dissenting opinions where no direct precedent exists.
ACCURACY	FACTS Benchmark	70.5%	Reliability & Grounding: Fact-checking case citations and minimizing "hallucinations" in legal texts.
SCALE	MRCR v2 (8-needle)	77.0%	Mega-Case Analysis: Instantly locating specific evidence or contradictions within 50+ volumes of case files.
PRECISION	AIME 2025	100%	Forensic Calculation: Error-free computation of damages, pensions, and tax liabilities.

LOGIC (HUMANITY'S LAST EXAM)
DEEP REASONING & LEGAL PHILOSOPHY CONTENT

This benchmark tests the ability to solve complex interdisciplinary problems, not just recall facts.

It acts as an intellectual partner for drafting dissenting opinions and reasoned parts of judgments.

The AI helps synthesize arguments from philosophy and law to create new legal positions.

ACCURACY (FACTS BENCHMARK)
JUDICIAL FACT-CHECKING CONTENT

This suite tests for "internal evidence grounding" to ensure every claim is backed by real data.

It automatically checks consistency, names, and statutes to prevent "hallucinations" in drafts.

Judges use it to audit draft decisions for maximum procedural accuracy before signing.

SCALE (MRCR V2)
THE "NEEDLE IN A HAYSTACK" ANALYSIS CONTENT

Modern models can process 1 million words (hundreds of books) to find a single specific detail.

Ideally suited for Cassation Courts: it instantly identifies if lower courts overlooked procedural requirements regarding mandatory evidence evaluation.

It detects contradictions across 50+ volumes of case files in seconds.

PRECISION (AIME 2025) MATHEMATICAL ACCURACY IN SOCIAL SECURITY AND TAX CASES

Achieving near-perfect accuracy in Olympiad-level math, this AI significantly reduces the risk of calculation errors.

It is critical for social security disputes: verifying pension formulas, indexations, and tax liabilities.

The tool allows lawyers to instantly audit complex algorithms used by state authorities.

CONCLUDING REFLECTIONS

AI can help justice systems with faster analysis and better access to information, but humans must stay fully responsible.

The rule of law requires transparency, strong checks, and clear limits on how AI is used.

AI tests show real progress, but also clear limits – so careful, responsible use is the only safe path for courts.

PREVIOUS RESEARCH AND AUTHOR’S CONTRIBUTIONS

1. Bernaziuk Ian. Artificial intelligence in the Ukrainian judiciary: charting the course under the digital gavel https://court.gov.ua/eng/supreme/pres-centr/news/1891488
2. Bernaziuk Ian. Artificial Intelligence and The Judicial System Of Ukraine: Results Of Cooperation In The Past Year https://court.gov.ua/storage/portal/supreme/prezentacii_2025/AI_Ukraine_bernaziuk.pdf
3. Берназюк Ян. Правосуддя і технології з використанням штучного інтелекту: короткострокові та середньострокові перспективи інтеграції https://court.gov.ua/storage/portal/supreme/prezentacii_2025/153_Justice_AI_Technologies_Integration_Prospects_bernaziuk.pdf

Supreme
Court

Thank you for attention!