Can AI Hold Office Hours? — A Summary

A summary of Ouellette, Motomura, Reinecke & Masur, Can AI Hold Office Hours? (draft of Mar. 4, 2026; forthcoming Journal of Legal Education). Read the full paper at ssrn.com/abstract=5166938.

The headline number

Not reassuring.

A substantial fraction of AI-generated responses were *harmful for learning* — answers that a student could reasonably rely on but that were actually wrong, misleading, or incomplete in ways that would undermine the student's understanding. The rates: **26% for GPT-4o, 14% for Claude 3.5, and 31% for NotebookLM**.

Another large fraction of responses failed to fully answer the question or contained minor errors: **25% for GPT, 31% for Claude, and 32% for NotebookLM**.

Add those together and the picture is that roughly half of the responses — across all three tools — had problems a student would have to already know enough to catch.

The authors' conclusion: they would not recommend these tools to their students as office-hours substitutes at this time.

The study that produced it

Four law professors — Lisa Larrimore Ouellette (Stanford), Amy R. Motomura (LMU Loyola), Jason D. Reinecke (Wisconsin), and Jonathan S. Masur (Chicago) — took the Masur & Ouellette *Patent Law: Cases, Problems, and Materials* casebook and fed it to three leading AI tools: OpenAI's GPT-4o, Anthropic's Claude 3.5 Sonnet, and Google's NotebookLM. They then posed 185 patent-law questions a typical student might ask during office hours, asking each tool to answer using only the casebook's content.

The design is the right one for testing whether AI can serve as a reliable teaching supplement. The tools are constrained to a specific text — so hallucinations "beyond the source" are out of scope, at least in principle. The questions are real ones students ask. The graders are faculty who actually teach the course.

Why this matters for the course

Three reasons this paper matters, including a specific one tied to how the course is built.

It disciplines the "AI can do law school" enthusiasm. Module II's other empirical pieces (Schwarcz, Cope) show real gains. This one shows real reliability problems. Any honest reading of the module has to hold both. The gains don't erase the reliability problems, and the reliability problems don't erase the gains.

It specifically tests the design pattern this course uses. NotebookLM, the tool I seeded as the companion notebook for this very course, was one of the tools Ouellette and colleagues tested. Its reliability rate was the lowest of the three. That fact should be fully visible to the reader. I've published the companion

NotebookLM with a caveat on the page — and the caveat is not boilerplate. It is informed by this paper.

It shows that RAG-style constraint is not a reliability panacea. A common assumption is that if you ground the AI in a specific set of sources, hallucinations disappear. They don't. The Ouellette study fed the models a single, bounded source (a casebook) and asked only questions answerable from that source. The unacceptable-response rates were still 14–31%. That is a finding.

The caveats

The study used early-2025 models. GPT-4o, Claude 3.5 Sonnet, and the then-current NotebookLM are all one to two generations behind what's available in mid-2026. It would be a mistake to assume the current-generation versions perform the same; they almost certainly perform better. It would also be a mistake to assume they are now reliable; the 2025 benchmarks suggest how much distance there still is.

A replication with Opus 4.5, GPT-5.x, and current NotebookLM is the natural next study. Until someone runs it, the 2025 numbers are what we have.

Also: the study tested the casebook alone, not the casebook plus the supplementary materials a typical patent-law professor would hand students. In a well-designed course, the AI office-hours tool would have access to the course's full canon, and reliability would presumably improve. Presumably. Whether it *would* is an empirical question no one has tested yet.

The line that stays with you

Roughly half of AI-generated answers — across three leading tools, constrained to the casebook — had problems a student would have to already know enough to catch. The companion NotebookLM on this very course page is exactly this kind of tool. Treat it as a study aid, not a study substitute.

Full citation: Lisa Larrimore Ouellette, Amy R. Motomura, Jason D. Reinecke & Jonathan S. Masur, Can AI Hold Office Hours? *(draft of Mar. 4, 2026, forthcoming* Journal of Legal Education*),* *ssrn.com/abstract=5166938*.