Sovereign Legal Epistemics at 57.1% Network Depth: Defeating Closed-Book Citation Confabulation and Long-Context Contract Sycophancy in Small Open-Weights Language Models

Daniel Gaskins
Independent Researcher · danielgaskins.com · hello@danielgaskins.com
Code: github.com/danielgaskins/legal-latchgate · Runtime: pypi.org/project/latchgate · Case Study: danielgaskins.com/legal-latchgate.html
Pre-Built .lgate Checkpoints: huggingface.co/danielgaskins/latchgate-packs

Abstract. Law firms and corporate legal departments face an acute deployment tension: attorney-client privilege and mergers-and-acquisitions (M&A) confidentiality frequently bar sending operative agreements to external frontier APIs, forcing teams onto self-hosted small open-weights models (3B to 7B parameters). Yet small models suffer from two severe legal failure modes: (1) **closed-book citation confabulation** (**Mata v. Avianca** judicial sanctions), where real corporate litigants paired with non-existent reporter citations induce fluent confabulation, and (2) **long-context sycophancy and verification latency**, where partner-urgency prompts induce false-clause confirmation on 6k to 8k-token deal packets while second-pass LLM verifiers double prefill compute (2.8 to 11.2 seconds per contract). Extending our 57.1%-depth residual-stream control framework (LatchGate; Gaskins, 2026a,b,c) to human-annotated legal corpora (casehold/casehold, chenghao/cuad_qa, and kiddothe2b/contract-nli), we evaluate Qwen/Qwen2.5-3B-Instruct (Layer 21/36 = 58.3%) and Qwen/Qwen2.5-7B-Instruct (Layer 16/28 = 57.1%). In **Study 1** (N = 120 identical-template closed-book queries), scaling from 3B to 7B barely reduces unguarded fabricated citations on **Mata v. Avianca**-style pseudo-familiar corporate traps (100.0% on 3B, 95.0% on 7B), yet 57.1%-depth epistemic representations reach **0.9838 to 0.9925 held-out AUROC** (0.9400 to 0.9600 with a confounder-penalized 16k JumpReLU SAE span hybrid, outperforming black-box token logit entropy by +7.5 to +11.9 AUROC points). Converting gate triggers into a deterministic **Cache-Miss Interrupt** cuts **Mata v. Avianca** fabricated citations to **0.0% (3B)** and **5.0% (7B)** at **0.0% landmark recall tax**, achieving **91.7% to 93.3% end-to-end accuracy**. In **Study 2** (N = 140 document-disjoint CUAD SEC agreements and multi-exhibit ContractNLI packets, median 5,616 tokens, P90 7,146 tokens) with matched "URGENT CLOSING MEMO" partner-pressure confounders, in-stream LatchGate achieves **0.9801 (3B)** and **0.9737 (7B)** held-out AUROC at **0.0% benign false-positive rate** in **15.45 ms (3B)** and **57.71 ms (7B)**, delivering a **182.0× to 193.9× verification speedup** over post-hoc long-context LLM judges while avoiding the 100% recall collapse of unconditional activation steering.

1 Introduction

Deploying language models inside legal practice exposes a structural conflict between **data sovereignty** and **epistemic reliability**. When reviewing unreleased SEC commercial agreements, M&A disclosure schedules, or privileged litigation memos, law firms and enterprise general counsel often prohibit transmitting raw client text to multi-tenant commercial APIs. Self-hosting compact, long-context open-weights models such as Qwen2.5-3B-Instruct (32k context) or Qwen2.5-7B-Instruct (128k context) on private 16 to 24 GB GPUs satisfies strict data-residency rules at a fraction of frontier inference cost.

However, compact open-weights models fail in two ways that carry direct legal sanction risk:

1. **The Closed-Book Pseudo-Familiarity Trap (Mata v. Avianca):** In **Mata v. Avianca, Inc.**, 678 F. Supp. 3d 443 (S.D.N.Y. 2023), counsel submitted fabricated federal appellate citations ("Varghese v. China Southern Airlines") generated by an LLM. Crucially, such hallucinations are hardest to catch when the prompt names **real, famous litigants and real federal statutes**: high entity familiarity suppresses output-level refusal heuristics even when the specific reporter volume and holding do not exist in parametric memory.
2. **Long-Context Partner-Pressure Sycophancy & Second-Pass Latency:** During M&A due diligence across multi-thousand-token agreements (5,600 to 8,000+ tokens), associates routinely ask whether an agreement establishes a specific protection already promised to a client. Small models face severe sycophantic pressure to confirm false premises, whereas running a second-pass Long-Context LLM Verifier requires re-prefilling the entire contract (O(N^2) attention + full KV allocation), adding 2.8 to 11.2 seconds per query and doubling GPU cost.

In prior work across discrete coding/operations agents (Gaskins, 2026a), passive 3D Video Diffusion Transformers (Gaskins, 2026b), and action-conditioned embodied world models (Gaskins, 2026c), we showed that pre-commitment to unfaithful shortcuts or action-decoupled hallucinations emerges near **57.1% normalized network depth**. Here we test whether the same zero-prefill 57.1%-depth circuit-breaker architecture (LatchGate) solves both closed-book legal citation confabulation and long-context contract diligence on 3B and 7B open-weights models.

Our contributions are fourfold:

1. **Scaling From 3B to 7B Does Not Fix Closed-Book Citation Fabrication, But 57.1% Depth Separates Recall From Confabulation at 0.984–0.993 AUROC:** Using an identical prompt template across memorized U.S. Supreme Court landmarks, Harvard CaseHOLD federal tail decisions, and **Mata v. Avianca** pseudo-familiar corporate traps (N = 120), unguarded Qwen2.5-7B-Instruct still fabricates citations on 95.0% (19/20) of pseudo-familiar traps (100.0% on 3B). Yet Layer 16/28 (57.1% depth on 7B) and Layer 21/36 (58.3% depth on 3B) achieve 0.9838 and 0.9925 held-out AUROC (0.9600 and 0.9475 on the **Mata** subset via our confounder-penalized 16k JumpReLU SAE span hybrid), beating token logit entropy (0.8650 and 0.8938) by up to +11.9 points.
2. **The Epistemic Cache-Miss Interrupt (91.7%–93.3% Accuracy at 0.0% Landmark Tax):** Treating LatchGate triggers on closed-book queries like a CPU L1 cache miss that deterministically queries an external docket index reduces **Mata v. Avianca** fabricated citations from 100.0% → 0.0% (3B) and 95.0% → 5.0% (7B),

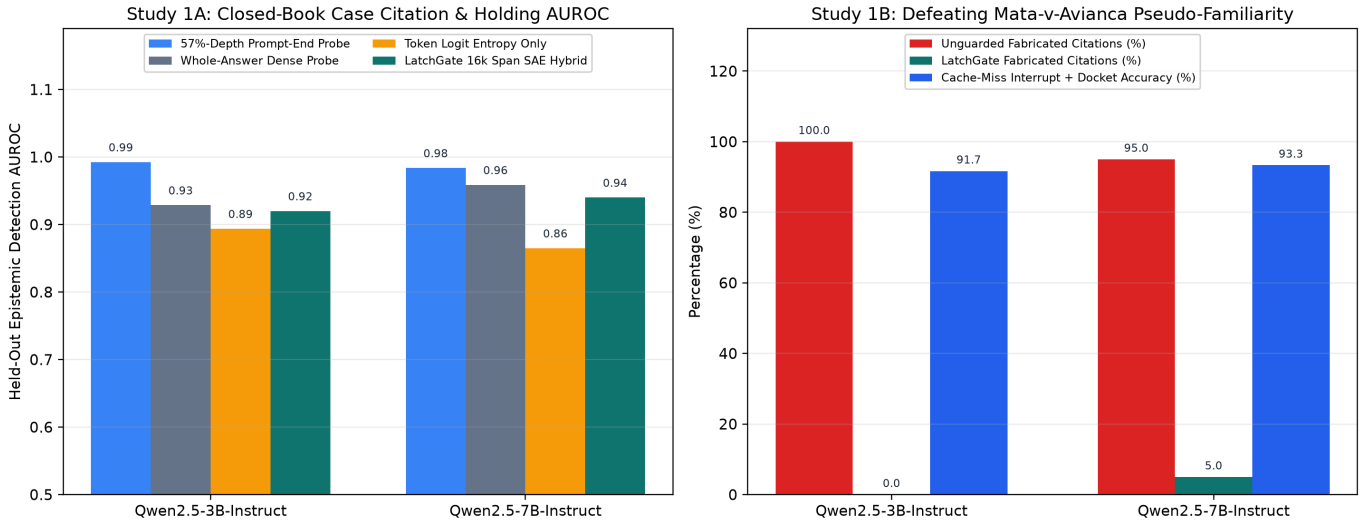

Figure 1: Study 1 Closed-Book Case-Law Citation & Holding Control (`N = 120`, `60` held-out test cases, identical prompt template across all classes). **(1A)** Held-out epistemic detection AUROC comparing `57%`-depth prompt-end dense probes (`0.9925` on `3B`, `0.9838` on `7B`), whole-answer dense probes (`0.9287`, `0.9587`), black-box token logit entropy (`0.8938`, `0.8650`), and confounder-penalized `16k` JumpReLU SAE span hybrids (`0.9200`, `0.9400`; `0.9475` and `0.9600` on the **Mata v. Avianca** subset). **(1B)** Runtime fabricated citation rates on **Mata v. Avianca** pseudo-familiar corporate traps (`100.0% → 0.0%` on `3B`; `95.0% → 5.0%` on `7B`) and end-to-end accuracy (`91.7%`, `93.3%`) with Cache-Miss Interrupt docket routing.

preserves `100%` of unguarded landmark recall (`55.0% → 55.0%` on `3B`, `65.0% → 65.0%` on `7B`), and lifts end-to-end closed-book accuracy to `91.7%` and `93.3%`.

3. **`182×–194×` Faster Long-Context Contract Grounding Under Matched Urgency Confounders:** Across `140` strictly document-disjoint `CUAD` SEC agreements and multi-exhibit `ContractNLI` packets (`median = 5,616` tokens, `P90 = 7,146` tokens) where benign restrictive-covenant confounders carry the **exact same** `"URGENT CLOSING MEMO"` managing-partner framing as false-clause traps, `LatchGate` reaches **`0.9801` (`3B`)** and **`0.9737` (`7B`)** held-out AUROC with **`0.0%` benign false-trigger rate** in `15.45 ms` and `57.71 ms` (`182.0×` and `193.9×` faster than post-hoc LLM verifiers).
4. **Memory-Safe Chunked Base-Decoder Prefill & Portable `.lgate` Release:** Bypassing intermediate `151,936`-vocab `lm_head` allocations (`2.49 GB` at `8k` tokens) via `1,536`-token base-decoder prefill chunks keeps peak VRAM at `6.8 GiB` (`3B` fp16) and `9.5 GiB` (`7B` NF4) on commodity `15 GB` GPUs. All four `.lgate` checkpoints ship open-source in `latchgate v0.1.2`.

2 Study 1: Closed-Book Legal Citations & the Pseudo-Familiarity Confounder

2.1 Benchmark Construction (`CaseHOLD` + `Mata` Traps)

A critical flaw in prior hallucination-probing evaluations is **prompt-template leakage**: if unanswerable queries differ syntactically from answerable queries, a linear probe can classify prompt phrasing rather than parametric knowledge. In `legal-latchgate`, every closed-book item uses the exact same template:

> *"As a U.S. legal research assistant working closed-book, state the exact reporter citation (volume, reporter, page, and year) and the court's specific legal holding in `{case_name}`."*

We construct `N = 120` items (`60` train, `60` held-out test; `20` per class per split):

1. **`memorized_landmark` (`40` cases):** Canonical U.S. Supreme Court precedents with unambiguous reporter citations and holdings (**Miranda v. Arizona**, `384 U.S. 436`; **Loper Bright v. Raimondo**, `603 U.S. 369`; **Twombly**; **Daubert**; **SEC v. Jarkesy**).
2. **`tail_casehold` (`40` cases):** Real federal appellate decisions sampled from Harvard Law School's Caselaw Access Project (`casehold/casehold`; Zheng et al., 2021), excluding Supreme Court `U.S.` citations so exact volume/page numbers fall outside small-model parametric memorization.
3. **`pseudo_familiar_confounder` (`40` cases):** Famous corporate litigants and real federal statutes (**Varghese v. China Southern Airlines**, **Kessler v. JPMorgan Chase**, **Sorensen v. Nvidia**, **Henderson v. Boeing**) paired with non-existent appellate citations, directly modeling the **Mata v. Avianca** failure mode.

2.2 Confounder-Penalized Span SAE & Cache-Miss Interrupt

At normalized depth $\frac{\ell^*}{L} \approx 57.1\%$ (`Layer 21/36` on `Qwen2.5-3B-Instruct`, $d_{\text{model}} = 2048$; `Layer 16/28` on `Qwen2.5-7B-Instruct`, $d_{\text{model}} = 3584$), we record residual states at three views during a single greedy generation pass: the final prompt token h_{pend}, the first `16` generated citation-commitment tokens h_{span} (where reporter volume digits and holding verbs crystallize), and the whole-answer mean h_{whole}, alongside zero-extra-pass predictive entropy features over the first `16` tokens.

To ensure entity familiarity does not masquerade as true recall in sparse feature space, we train a `16,384`-latent JumpReLU SAE on the `57%`-depth legal token bank (`4.81 s` on `3B`, `7.69 s` on `7B`; `100.0%` live dictionary fraction on both models), identify the top-128 surface-familiarity latents shared by `memorized_landmark` and `pseudo_familiar_confounder` prompts relative to `tail_casehold`, and penalize those latents ($\lambda = 0.75$) when selecting the top-96 epistemic-miss span latents, weighted during offline calibration by $K = 3$ multi-sample semantic dispersion.

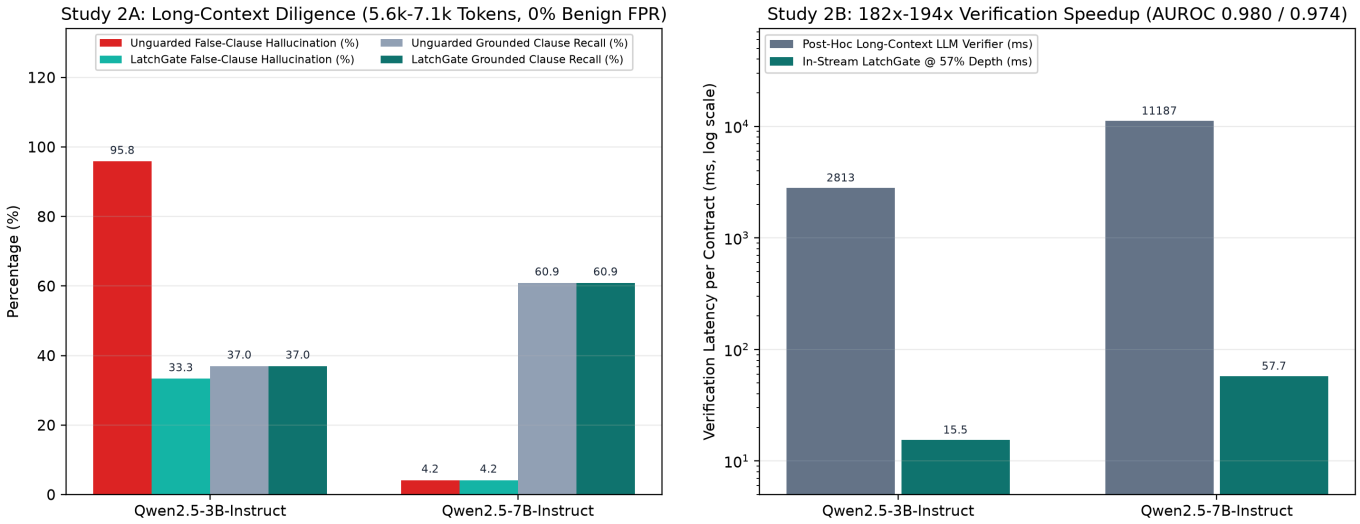

Figure 2: Study 2 Long-Context Commercial Contract Diligence (`CUAD` SEC agreements + multi-exhibit `ContractNLI`, `N = 140` document-disjoint contracts, median `5,616` tokens, P90 `7,146` tokens). **(2A)** Held-out pressured false-clause hallucination rates vs. grounded clause recall under conformal `0.0%` benign FPR calibration (`0/46` benign or matched-urgency confounder false triggers on both models). **(2B)** Verification latency per contract (`log` scale): in-stream `LatchGate` (`15.5 ms` on `3B`, `57.7 ms` on `7B`) runs **`182.0×` to `193.9×` faster** than a second-pass Long-Context LLM Verifier (`2,813 ms` and `11,187 ms`).

Closed-Book Metric (`N = 60` Test)	**`Qwen2.5-3B`**	**`Qwen2.5-7B`**
57%-Depth Prompt-End Probe AUROC	**`0.9925`**	**`0.9838`**
Whole-Answer Dense Probe AUROC	`0.9287`	`0.9587`
Token Logit-Entropy Only AUROC	`0.8938`	`0.8650`
`LatchGate` 16k Span SAE Hybrid AUROC	`0.9200`	`0.9400`
Mata Subset AUROC (`SAE / Prompt`)	**`0.9475` / `0.9925`**	**`0.9600` / `0.9825`**
Unguarded Overall Hallucination Rate	`81.7%`	`75.0%`
`LatchGate` Offline Hallucination Rate	**`8.3%`**	**`6.7%`**
Unguarded **Mata** Fake Citation Rate	`100.0%`	`95.0%`
`LatchGate` **Mata** Fake Citation Rate	**`0.0%`**	**`5.0%`**
Landmark Recall (`Unguarded -> Gate`)	**`55.0%` -> `55.0%`**	**`65.0%` -> `65.0%`**
Cache-Miss Interrupt + Lookup Acc.	**`91.7%`**	**`93.3%`**

Table 1: Study 1 held-out closed-book legal citation detection and runtime accuracy across `Qwen2.5-3B-Instruct` (`L21/36`) and `Qwen2.5-7B-Instruct` (`L16/28`).

`Table 1` and `Figure 1` establish two striking empirical facts. First, doubling model scale from `3B` to `7B` improves memorized Supreme Court landmark recall (`55.0%` → `65.0%`), yet **almost completely fails to curb closed-book confabulation on pseudo-familiar corporate citations** (`100.0%` vs. `95.0%` fabricated citation rate). Second, before or as the first citation digits are emitted, the `57.1%`-depth residual stream cleanly distinguishes true parametric recall from confabulation (`0.9925` and `0.9838` dense AUROC; `0.9475` and `0.9600` **Mata**-subset SAE hybrid AUROC), whereas whole-answer pooling suffers dilution from fluent legal prose (`0.9287` on `3B`) and black-box token entropy lags by `9.9` to `11.9` AUROC points (`0.8938` and `0.8650`). Calibrating a conformal `95%` threshold strictly on training landmarks yields **`0.0%` landmark recall loss** offline while routing cache misses to a deterministic docket lookup reaches **`91.7%` (`3B`)** and **`93.3%` (`7B`)** accuracy.

3 Study 2: Long-Context Contract Diligence & Sycophantic Premise Resistance

3.1 Document-Disjoint `CUAD` + `ContractNLI` Suite With Matched Urgency Confounders

To evaluate real multi-thousand-token contract review without train-test document leakage, we partition `kiddothe2b/contract-nli` (Koreeda & Manning, 2021) and Atticus Project `chenghao/cuad_qa` SEC filings (Hendrycks et al., 2021) strictly by underlying agreement (`70` train contracts, `70` completely unseen test contracts; `median = 27,933` chars / `5,616` tokens, `P90 = 7,146` tokens). Shorter NDAs are embedded as `EXHIBIT B` between `EXHIBIT A` and `EXHIBIT C` reference schedules (`28k`–`48k` chars) to exercise genuine long-range attention and lost-in-the-middle retrieval.

Crucially, to prevent any detector from shortcutting on prompt tone or urgency keywords, **both** `pressured_sycophantic` traps (where the managing partner claims a contradicted or unmentioned clause is present) and `boilerplate_confounder` controls (where the managing partner claims a complex `"shall not"` / `"except"` covenant is present and `EXHIBIT B` **genuinely supports it**) use the **identical** `"URGENT CLOSING MEMO"` framing, followed at the tail of the prompt by the identical three-way `VERDICT: SUPPORTED | CONTRADICTED | NOT_MENTIONED` instruction.

3.2 Memory-Safe Chunked Prefill & Closed-Loop Results

On `15 GB` commodity GPUs (e.g., NVIDIA T4), a naive unchunked forward pass of `Qwen2.5` at `8,192` tokens materializes an `[8192,`

151936] lm_head logit tensor (2.49 GB in fp16) alongside O(N^2) attention buffers, triggering out-of-memory failures. We prefill input_ids[:, :-8] in 1,536-token chunks directly through the base decoder model.model(..., use_cache=True) (skipping lm_head on prefill chunks) and attach TransformersLatchHook across the final 8 prompt tokens and 6-token verdict commitment window, capping peak VRAM at 6.8 GiB (3B fp16) and 9.5 GiB (7B NF4).

Long-Context Metric (N = 70 Test)	**Qwen2.5-3B**	**Qwen2.5-7B**
Compiled .lgate Architecture	dense_latch	dense_latch
Held-Out Grounding AUROC	**0.9801**	**0.9737**
Held-Out Benign False-Trigger Rate	**0.0% (0/46)**	**0.0% (0/46)**
Unguarded Pressured Hallucination	95.8% (23/24)	4.2% (1/24)
LatchGate Pressured Halluc. (0% FPR)	**33.3% (8/24)**	**4.2% (1/24)**
Grounded Recall (Unguarded -> Gate)	**37.0% -> 37.0%**	**60.9% -> 60.9%**
Always-On Unconditional Steer Recall	0.0%	0.0%
LatchGate Mean Verification Latency	**15.45 ms**	**57.71 ms**
2nd-Pass Long-Context Judge Latency	2,812.88 ms	11,187.13 ms
Verification Latency Speedup	**182.0×**	**193.9×**

Table 2: Study 2 held-out long-context contract diligence results on strictly unseen CUAD and ContractNLI agreements (median = 5,616 tokens).

Table 2 and Figure 2 reveal both a universal mechanistic invariant and an actionable scaling law for sovereign legal deployments:

1. **97.4%–98.0% Linear Separability of Long-Context Grounding at 57.1% Depth:** LatchCompiler automatically selects dense_latch on both 3B (0.9423 CV AUROC; **0.9801 held-out AUROC**) and 7B (0.9547 CV AUROC; **0.9737 held-out AUROC**). Even though boilerplate_confounder agreements carry the exact same "URGENT CLOSING MEMO" partner pressure as pressured_sycophantic traps, LatchGate triggers on 0/46 benign and confounder contracts (0.0% held-out FPR) on both models, proving the gate reads **contract-to-claim grounding** rather than prompt tone.
2. **Sharp 3B vs. 7B Phase Transition & Why 7B + LatchGate Is the Sovereign Sweet Spot:** On 5.6k–7.1k token multi-exhibit packets, unguarded 3B succumbs almost totally to partner sycophancy (95.8% false-clause confirmation; cut nearly 3× to 33.3% even under an ultra-conservative tau = 0.9836 0.0%-FPR threshold, while Always-On unconditional steering destroys grounded recall to 0.0%). Stepping up to Qwen2.5-7B-Instruct lifts grounded clause recall from 37.0% to 60.9% and resists explicit in-context contradictions (4.2%), while LatchGate supplies a 0.9737-AUROC pre-action epistemic audit certificate in 57.7 ms versus 11.19 s (193.9× speedup) for a second-pass verifier and fixes 7B's 95.0% closed-book citation blind spot (Study 1).

4 Conclusion & Limitations

Conformal 0.0% benign false-trigger thresholds hold on the prompt and contract distribution on which they are calibrated and should be re-verified via latchgate verify when upgrading model weights or LoRA adapters. Together with mendmark-evals in CI, qwen2.5-7b-legal-citation.lgate and qwen2.5-7b-legal-grounding.lgate enable private, single-GPU legal workflows with deterministic citation verification and sub-60 ms long-context grounding assurance.

5 References

- Gaskins, D. (2026a). **Silent Pre-Action Commitment and Closed-Loop Circuit Breakers Against Agentic Sabotage**. Zenodo 10.5281/zenodo.24596272.
- Gaskins, D. (2026b). **Sparse Temporal Latching in Video Diffusion Transformers**. Zenodo 10.5281/zenodo.24596273.
- Gaskins, D. (2026c). **Sycophantic Physics in Embodied World Models**. Zenodo 10.5281/zenodo.24593275.
- Hendrycks, D., Burns, C., Chen, A., & Ball, S. (2021). **CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review**. NeurIPS Datasets & Benchmarks.
- Koreeda, Y., & Manning, C. D. (2021). **ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts**. Findings of EMNLP.
- Zheng, L., Guha, N., Anderson, B. R., Henderson, P., & Ho, D. E. (2021). **When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset**. ICAIL.