

## Analysis of Current Benchmarks of Legal Applications of LLMs

Current benchmarks for legal applications of large language model (LLM) systems are fundamentally lacking. We articulate two main problems with existing legal LLM benchmarks:

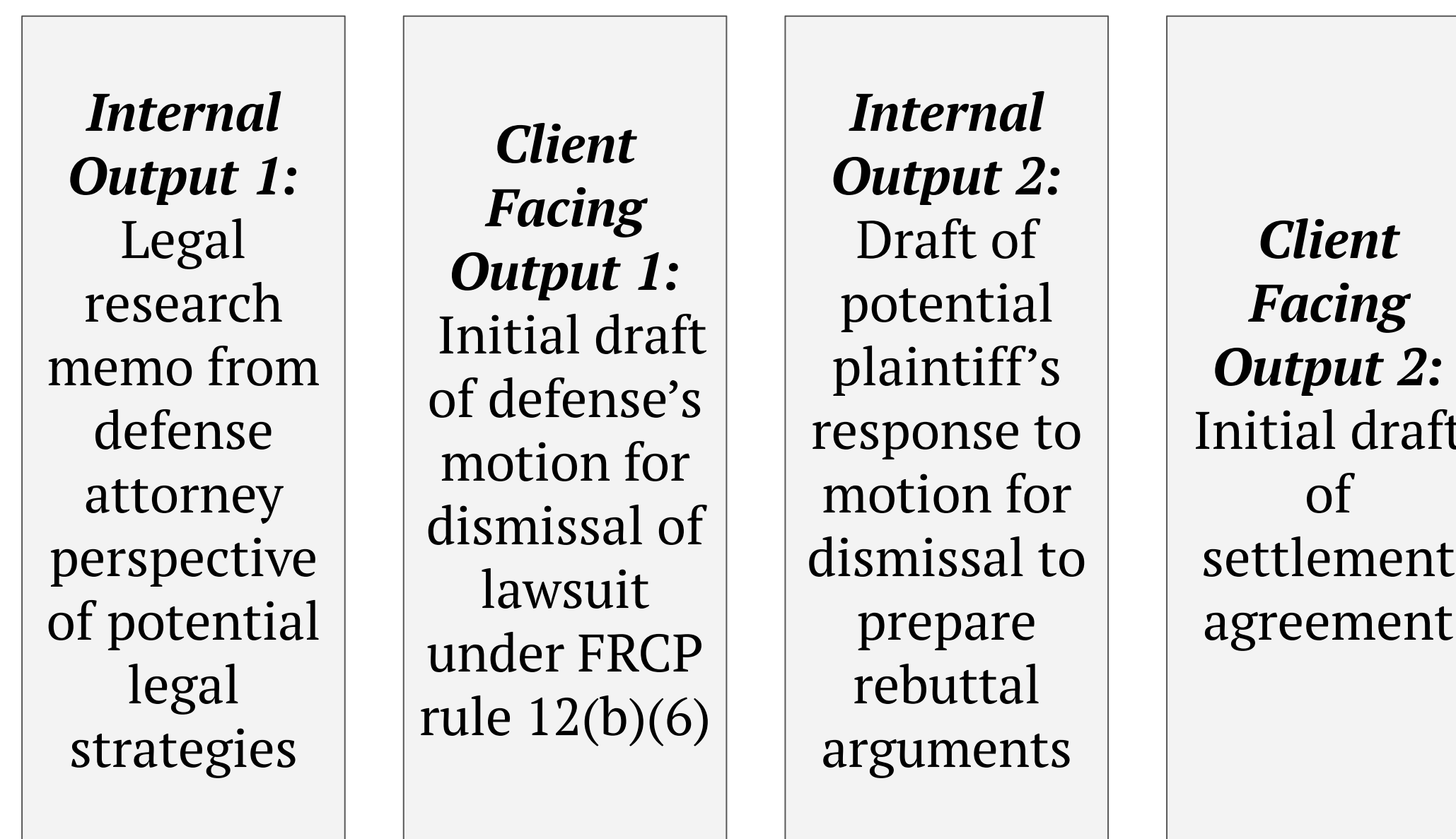
- **Impractical Tasks:** The majority of existing benchmarks seek to measure hallucination rates or recall ability on hyper-specific case law/statutes rather than evaluating whether LLMs can replicate genuine legal work flows.
- **Insufficient Evaluation Criteria:** It is not clear whether performance on benchmark tasks is correlated to usefulness for lawyers, as success is typically measured by “accuracy” or “completeness”—metrics that are not typically used to evaluate actual legal work product.

## Our Proposed Benchmark: Lawyer Preference of LLM Performance on Holistic Legal Workflows

Our proposed benchmark will discern what lawyers really value in legal work products to evaluate the performance of LLMs on legal tasks. First, we create a set of legal tasks that encapsulate a legal workflow in practice—how legal products and their intermediates are drafted. Then, we ask senior partners to perform a blind evaluation of LLM-prepared and human associate-prepared legal work products to determine lawyer preference; lawyers will not only rate which outputs they prefer, but provide detailed justifications for why they are preferable. These justifications will be used to create tangible evaluation criteria—beyond “accuracy” and “completeness”—to benchmark LLM performance on authentic legal workflows.

## Sample Benchmark Task and Evaluation Methodology

**Data:** Civil complaint filed against [employer] by [plaintiff] for employment discrimination and wrongful termination.



**Initial Evaluation:** For each task, lawyers rate which output they prefer on 7 point Likert scale, with detailed justification as to why:

- 3. Response 1 is much better than Response 2
- 2. Response 1 is better than Response 2
- 1. Response 1 is slightly better than Response 2
- 1. Response 2 is slightly better than Response 1
- 2. Response 2 is better than Response 1
- 3. Response 2 is much better than Response 1
- 100. Neither response is valid

Specific lawyer preference criteria aggregated from responses

## Sample LLM and Human Associate Output

### Human Output

**SETTLEMENT AGREEMENT AND RELEASE**

This Settlement Agreement and Release (the Agreement) is entered into by and between Plaintiff (Plaintiff), Defendant Hospital, and Defendant Staffing Agency (each, a Defendant and together, the Defendants) (each, a Party, and collectively, the Parties).

**WHEREAS**, on or about July 2024, Plaintiff filed a complaint (the Complaint) in the [Court] of the State of New York, County of New York, index number [Number] (the Pending Action).

**WHEREAS**, Plaintiff alleges that Defendants are responsible for her unlawful termination, in violation of Title VII of the Civil Rights Act of 1964, 42 U.S.C. §12000e, the NY State Human Rights Law (“NYSHRL”) and the New York City Human Rights Law (“NYCHRL”), on false and pre-textual grounds, setting forth that the termination was motivated solely by her pregnancy. Plaintiff filed a complaint seeking to recover damages sustained as the result of being discriminated and retaliated against, culminating in termination on the basis of her pregnancy. Plaintiff suffered and continues to suffer severe emotional damages and lost wages.

**WHEREAS**, Defendant denies the allegations of the Complaint.

**WHEREAS**, the Parties desire to settle fully and finally any and all differences between them, including, but not limited to, the claims made in the Pending Action.

**NOW, THEREFORE**, the Parties agree as follows:

**Recitals.** The Parties hereby incorporate the foregoing recitals into this Agreement. The Parties agree that the recitals are true and correct to the best of the Parties' knowledge, information, and belief.

**Settlement Terms.** Defendant agrees to pay to Plaintiff the sum of \$[Dollar] (the Settlement Payment) as full and final settlement of any claims Plaintiff may have against Defendant. Defendant hereby acknowledges that the Settlement Payment represents full and final payment for any and all damages claimed in the Pending Action, including attorney's fees and costs.

**Costs.** Each Party shall bear its own fees and costs, including attorney's fees.

**Release.** In consideration for the Settlement Payment, Plaintiff hereby irrevocably and unconditionally releases and forever discharges Defendants, their agents, servants, employees, parents, subsidiaries, affiliates, successors, assigns, or related entities, and any of their agents, servants, employees, parents, subsidiaries, affiliates, successors, assigns, or related entities (each, separately, a Releasee, and collectively, the Releasees) from any and all claims, demands, rights, actions, causes of action, or liabilities Plaintiff may have against the Releasees, including all claims of any kind or nature, known or unknown, suspected or

### LLM Output

**SETTLEMENT AGREEMENT AND RELEASE**

This Settlement Agreement and Release (“Agreement”) is made and entered into as of [Effective Date], by and between [PLAINTIFF NAME] (“Plaintiff”) and [DEFENDANT NAME] (“Defendant”), Plaintiff and Defendant are collectively referred to as the “Parties.”

**RECITALS.**

**WHEREAS**, Plaintiff filed a lawsuit against Defendant in the United States District Court for the [District of ], Case No. [ ] (the “Action”), alleging claims of pregnancy discrimination, retaliation, and violations under Title VII of the Civil Rights Act, the Pregnancy Discrimination Act, the Pregnant Workers Fairness Act, the Family and Medical Leave Act, and applicable state anti-discrimination laws.

**WHEREAS**, Defendant denies all allegations of wrongdoing and liability asserted in the Action.

**WHEREAS**, the Parties desire to settle fully and finally all claims and disputes between them arising out of or relating to the Action, without admitting any liability or wrongdoing, to avoid the expense and uncertainty of continued litigation.

**NOW, THEREFORE**, in consideration of the mutual promises and covenants contained herein, and for other good and valuable consideration, the sufficiency of which is hereby acknowledged, the Parties agree as follows:

**Settlement Payment.**

Payment Amount: Defendant agrees to pay Plaintiff the total sum of [Settlement Amount in Words] Dollars ([Settlement Amount in Numbers]) (“Settlement Amount”).

Payment Terms: The Settlement Amount shall be paid as follows: a. A payment of \$[Amount] allocated to alleged lost wages, less applicable withholdings and deductions, to be reported on IRS Form W-2; b. A payment of \$[Amount] allocated to alleged emotional distress and other non-wage damages, to be reported on IRS Form 1099; c. A payment of \$[Amount] allocated to Plaintiff's attorneys' fees and costs, to be paid directly to Plaintiff's counsel, [Law Firm Name], to be reported on IRS Form 1099.

Payment Timing: Defendant shall deliver the Settlement Amount to Plaintiff and Plaintiff's counsel within [Number] days after the later of: a. The Court's dismissal of the Action with prejudice; and b. Receipt of a signed IRS Form W-9 from Plaintiff and Plaintiff's counsel.

**Release of Claims.**

Plaintiff's Release: In exchange for the consideration provided in this Agreement, Plaintiff hereby irrevocably and unconditionally releases, waives, and forever discharges Defendant and its past, present, and future parents, subsidiaries, affiliates, divisions, predecessors, successors, assigns, and all of their respective officers, directors,

## Sample Senior Partner Evaluation and Corresponding Criteria

Which did you prefer?	Why?	Extracted Preference Criteria:
1. Exhibit A is slightly better than Exhibit B	Exhibit A appears more favorable for the defendant. It offers broader protections, simpler payment terms, and stronger confidentiality clause, while avoiding additional obligations like tax liability disclaimers and property returns. However, ensure the chosen agreement aligns with your client's priorities, especially if strategic concessions (e.g., narrower scope of release or specific payment allocations) are necessary for reaching settlement.	
2. Exhibit B is better than Exhibit A	While both agreements aim to resolve the dispute between the parties, the first agreement offers more comprehensive protections for the defendant. It meticulously addresses the release of claims, liabilities, return of company property, and includes mutual non-disparagement clauses—all of which are crucial for minimizing risk and safeguarding the defendant's interests. Therefore, I would recommend opting for Exhibit B to ensure a more secure and favorable outcome for the defendant.	