Motivations for Reframing Legal Large Language Model Benchmarking

Riya Ranjan[1]   Megan Ma[2]

[1]Stanford Department of Computer Science   [2]Stanford Law School

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:
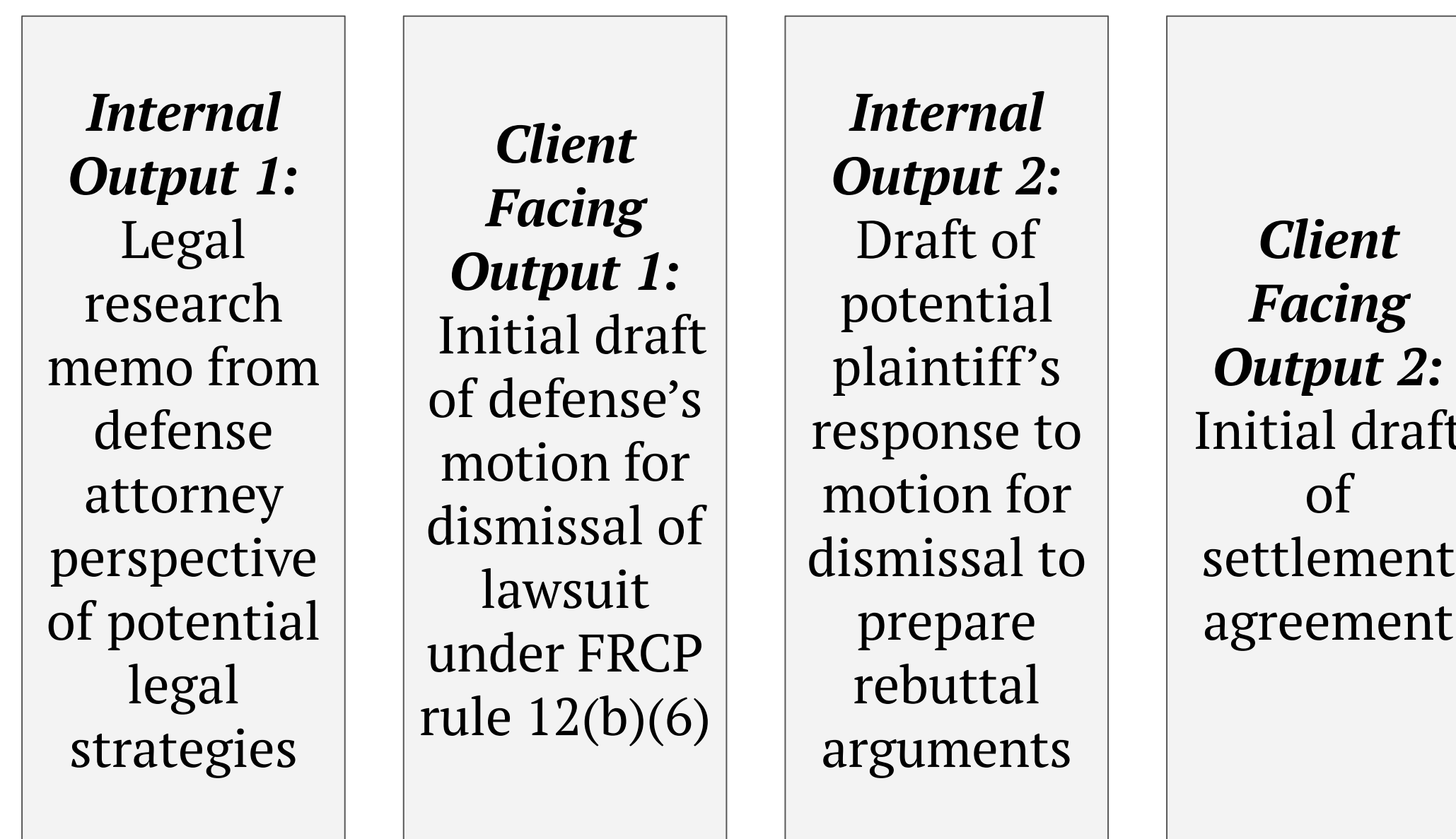
- *Impractical Tasks:* The majority of existing benchmarks seek to measure hallucination rates or recall ability on hyper-specific case law/statutes rather than evaluating whether LLMs can replicate genuine legal work flows.
- *Insufficient Evaluation Criteria:* It is not clear whether performance on 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.

Internal Output 1: Legal research memo from defense attorney perspective of potential legal strategies

Client Facing Output 1: Initial draft of defense's motion for dismissal of lawsuit under FRCP rule 12(b)(6)

Internal Output 2: Draft of potential plaintiff's response to motion for dismissal to prepare rebuttal arguments

Client Facing Output 2: Initial draft of settlement agreement

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

LLM Output

Sample Senior Partner Evaluation and Corresponding Criteria

Extracted Preference Criteria: Broad protections, simple payment terms, lack of additional obligations, minimizing risk