

RELATÓRIO EXECUTIVO - RODADA EXPERIMENTAL

Comparação entre benchmark humano real e quatro execuções de IA

RELATÓRIO 2 - EXECUTIVO E ANONIMIZADO

Objetivo. Avaliar, em um caso processual real, se quatro condições de IA conseguiram reconstruir os autos, identificar cliente e medida, separar prova de inferência, testar a tese adversa e produzir estratégia/peça comparáveis a um benchmark humano real. A decisão judicial posterior foi usada apenas para verificar quais fundamentos o juízo efetivamente adotou, sem retroagir para “corrigir” o ranking ex ante.

Desenho metodológico

As quatro execuções foram analisadas de forma cega e independente contra o mesmo corpus processual histórico. Depois, foram comparadas entre si. Em etapa separada, a peça humana real foi auditada somente com base no que existia quando foi redigida. Por fim, a decisão posterior foi examinada em bloco autônomo. Os critérios incluíram fidelidade documental, identificação de cliente/polo, fase, medida, prova versus inferência, omissões, melhor argumento adverso, estratégia, pedidos e integridade entre análise e peça.

Condição	Resultado sintético
MVP / condição 1	Melhor desempenho global entre as IAs; boa disciplina probatória, teste adversarial e peça coerente.
Condição 2	Muito forte no confronto de mérito, mas inseriu na peça um fato processual que o corpus da execução não comprovava.
Condição 3	Detectou uma inconsistência documental importante, porém presumiu o cliente e redigiu para o polo oposto.
Condição 4	Muito forte em disciplina epistemológica, mas omitiu o principal fato adverso cronológico.

Principais achados

- O MVP se aproximou do benchmark humano na identificação do problema jurídico central, da parte representada, da medida defensiva e da necessidade de exigir fatos concretos para atingir o patrimônio de terceiro.
- O MVP foi materialmente melhor do que a peça humana em três dimensões desta rodada: identificação do principal fato adverso, recuperação de fatos executivos que enfraqueciam a narrativa contrária e recomendação de prova positiva antes de afirmar separação patrimonial.
- A peça humana foi melhor na ênfase ao problema processual específico que posteriormente se tornou o núcleo da decisão judicial. Essa proximidade ex post não torna seus demais argumentos automaticamente corretos.
- Uma condição alternativa encontrou uma inconsistência entre documentos com potencial efeito operacional que tanto o humano quanto o MVP deixaram passar.
- Outra condição de IA cometeu falha crítica de integridade ao transformar uma incerteza do corpus em fato certo na minuta.
- Uma terceira condição cometeu falha fundacional ao presumir o cliente e produzir peça para o lado oposto.

Humano x MVP

Dimensão	Conclusão desta rodada
Polo/fase/medida	Equivalentes em correção básica; o humano teve maior foco na objeção processual depois acolhida.
Teste adversarial	MVP superior nesta rodada: enfrentou o pior fato disponível, não apenas a versão mais fraca do argumento adverso.
Prova x inferência	MVP superior nesta rodada: preservou melhor a diferença entre falta de prova do adversário e inexistência do fato.
Reconstrução documental	MVP mais abrangente; humano deixou de usar fatos materiais do histórico.
Precisão jurídica	Humano teve imprecisões normativas; MVP teve correções localizadas, mas parte de sua jurisprudência era posterior ao momento histórico.
Aderência à decisão posterior	Humano mais próximo do fundamento efetivamente usado pelo juízo, que foi processual.

Limitações que impedem conclusão de validação

- A amostra é um único caso; não há base estatística para generalização.
- O benchmark humano é uma peça real, não um padrão-ouro: ela própria contém falhas e omissões.
- A decisão posterior revela o fundamento escolhido por um julgador, mas não mede toda a qualidade ex ante.
- Algumas execuções usaram jurisprudência decidida depois do corte histórico do caso. Esses pontos não podem ser tratados como vantagem comparável sobre o humano.
- O corpus experimental e a exportação processual posterior não são idênticos; fatos posteriores não podem ser usados para punir ou premiar retroativamente uma execução.
- Todas as condições continuariam exigindo revisão humana antes de protocolo.

Próximos passos

A próxima fase deve replicar o desenho em amostra maior e heterogênea, congelar simultaneamente fatos e jurisprudência na data histórica, preservar versões exatas do corpus, manter avaliadores cegos, separar análise/estratégia/peça final e medir o nível de correção humana necessário. Também é recomendável incluir casos com cliente ambíguo, contradições interdocumentais e situações em que a melhor resposta não seja redigir imediatamente.

Conclusão provisória

Nesta rodada, o MVP produziu o melhor resultado entre as condições de IA e apresentou desempenho comparável ao advogado real em várias dimensões, superior em algumas e inferior em outras. O achado mais útil é a complementaridade: o humano foi mais aderente ao ponto processual que decidiu o incidente; o MVP foi mais forte em análise adversarial, reconstrução factual e disciplina probatória. Uma terceira execução ainda revelou um ponto documental perdido por ambos.

Esses resultados justificam novos testes; não justificam afirmar que o MVP está validado ou que é superior de forma geral.