

Avaliação do Raciocínio Jurídico em Modelos de Linguagem com Base na Estrutura IRAC

Janderson Santos
DCC-UFMG

Belo Horizonte, Brasil
jandersong4@ufmg.br

Pedro Calais
DCC-UFMG

Belo Horizonte, Brasil
pcalais@dcc.ufmg.br

Ricardo Shen
DCC-UFMG

Belo Horizonte, Brasil
ricardo.shen@dcc.ufmg.br

Gabriel Rolla
DCC-UFMG

Belo Horizonte, Brasil
gabrielrolla@dcc.ufmg.br

Ana Luísa Araújo
FD-UFMG

Belo Horizonte, Brasil
analuisavba@gmail.com

Fernanda Carvalho
FD-UFMG

Belo Horizonte, Brasil
fernandalvesc@gmail.com

Victor Hon Fonseca
DCC-UFMG

Belo Horizonte, Brasil
victor.fonseca@dcc.ufmg.br

Aziz Saliba
FD-UFMG

Belo Horizonte, Brasil
azizsaliba@ufmg.br

Anisio Lacerda
DCC-UFMG

Belo Horizonte, Brasil
anisio@dcc.ufmg.br

Wagner Meira Jr
DCC-UFMG

Belo Horizonte, Brasil
meira@dcc.ufmg.br

Resumo—A crescente aplicação de Grandes Modelos de Linguagem (LLMs) no domínio jurídico tem levantado questões fundamentais sobre suas reais capacidades de raciocínio. Embora esses modelos consigam gerar respostas juridicamente plausíveis, ainda não está claro se de fato realizam raciocínio jurídico ou apenas reproduzem padrões memorizados a partir dos dados de treinamento. Neste artigo, propomos uma metodologia para avaliar as habilidades de raciocínio jurídico dos LLMs com base na tradicional estrutura IRAC (Issue, Rule, Application, Conclusion), amplamente utilizada no ensino e na prática profissional do Direito. Nossa abordagem decompõe a resolução de problemas jurídicos em intervenções instrucionais guiadas, aplicadas a questões de múltipla escolha de quatro edições do Exame da Ordem dos Advogados do Brasil (OAB). Avaliamos sete LLMs — ChatGPT 4.1, Claude 3.5 Haiku, Gemini 1.5 e 2.0, Sabia 3.1, Sabiazinho 3 e DeepSeek-R1 14B — e analisamos seu desempenho em diferentes áreas jurídicas. Os resultados mostram que a maioria dos modelos supera a nota mínima de aprovação (50%), com acurácia variando entre 45% e 90% nas quatro provas analisadas. A inserção explícita das regras jurídicas (componente R) foi o fator isolado de maior impacto, elevando em alguns casos a taxa de acerto em até 20 pontos percentuais — especialmente entre modelos menores. Além disso, os modelos apresentam desafios específicos por área, como a identificação de normas aplicáveis em Direito Eleitoral ou a delimitação de controvérsias jurídicas centrais em Direito Civil. Ao tratar cada componente do IRAC como um proxy para rastreamento de conhecimento, nosso método permite uma avaliação mais refinada do raciocínio jurídico e sugere novos caminhos para modelar a dinâmica de aprendizagem dos LLMs em domínios estruturados e de alta complexidade, como o Direito.

Palavras-chave—Large language models, Raciocínio jurídico, Direito e tecnologia

I. INTRODUÇÃO

A aplicação crescente da Inteligência Artificial (IA) no domínio jurídico tem sido uma tendência marcante, abrangendo desde a automação da revisão de documentos até a previsão de resultados de processos e o auxílio à pesquisa jurídica [1], [2]. A interseção entre raciocínio jurídico e Processamento de Linguagem Natural (PLN), especialmente com o uso de Grandes Modelos de Linguagem (LLMs), apresenta oportunidades para ampliar as capacidades humanas no campo jurídico. Como

a prática jurídica é fortemente baseada em texto e depende de uma linguagem especializada — incluindo termos técnicos e exigindo precisão extrema na expressão linguística — o Direito já é um dos domínios em que mais se encontram casos de uso de modelos de linguagem [3].

Apesar do enorme potencial dos LLMs no campo jurídico, uma pergunta fundamental persiste: até que ponto esses modelos são realmente capazes de raciocinar como advogados? Como o Direito é um domínio sensível e de alto risco, explicitar a confiabilidade e explicar o processo interno dos modelos torna-se ainda mais crítico [4]. A prática eficaz do Direito requer habilidades como identificar questões jurídicas, lembrar e aplicar regras relevantes, interpretar textos, formular conclusões e compreender aspectos retóricos [5], e, embora consigam produzir respostas juridicamente plausíveis, ainda não está claro se os modelos de linguagem de fato compreendem e aplicam o raciocínio jurídico, ou se apenas imitam padrões superficiais presentes nos textos legais ou repetem informações memorizadas. De fato, é uma tendência crescente na literatura entender o desempenho de modelos além de métricas agregadas (como acurácia) e aplicar metodologias que quantificam o quanto LLMs memorizam padrões observados nos dados de treinamento [6] e o quanto realmente “raciocinam” [7], [8], o que pode ser entendido como a capacidade de generalizar para problemas novos ou mesmo sutilmente diferentes [9].

Nesta linha, este trabalho propõe uma metodologia estruturada de avaliação da capacidade de raciocínio jurídico de LLMs que segue essa tendência de entender o desempenho dos modelos além da taxa de acerto agregada [10]. Para avaliar sistematicamente as capacidades dos LLMs em raciocínio jurídico, fatoramos o conhecimento jurídico em seus componentes fundamentais, utilizando uma decomposição tradicional na prática do Direito que fatora os componentes de uma decisão jurídica nos componentes (*Issue*, *Rule*, *Application*, *Conclusion*) [11]–[13]. Essa metodologia, conhecida como *IRAC*, é uma estrutura comum e abrangente utilizada por profissionais do Direito para resolver problemas e realizar

análises jurídicas. Ela preconiza que, para tomar uma decisão jurídica, o estudante ou advogado deve, sequencialmente, (1) identificar as questões jurídicas, (2) enunciar as regras legais aplicáveis, (3) aplicar os fatos às regras e (4) finalmente, chegar a uma Conclusão.

Nossa abordagem fornece incrementalmente pistas de raciocínio jurídico — com base na estrutura IRAC — aos modelos que resolvem questões da primeira fase do Exame da Ordem dos Advogados do Brasil (OAB). Por exemplo, um modelo pode falhar ao resolver uma determinada questão sobre Direito do Trabalho, mas, ao ser instruído sobre quais regras deve aplicar (leis, dispositivos constitucionais, resoluções do STF), ele pode passar a ser capaz de identificar a resposta correta, o que indica que para este exemplo o modelo tem a capacidade de desempenhar o raciocínio jurídico (etapa 'A' da metodologia IRAC), mas falhou ao recuperar a informação necessária para aplicar as leis. A estratégia proposta pode ser interpretada como uma estratégia conhecida na pedagogia como *rastreamento do conhecimento*, [14], que busca modelar como aprendizes adquirem e aplicam habilidades específicas ao longo do tempo. Nossa configuração trata cada componente do IRAC (Problema, Regra, Aplicação, Conclusão) como uma *intervenção instrucional*, permitindo observar como os LLMs respondem a diferentes formas de orientação no raciocínio. Assim, estimamos a presença ou ausência de capacidades de raciocínio subjacentes a partir do desempenho diferencial em contextos guiados. Essa perspectiva não apenas permite uma análise mais granular do comportamento dos LLMs, como também abre caminhos para modelar a dinâmica de aprendizagem de agentes artificiais em domínios estruturados, como o Direito.

Os experimentos com sete LLMs (ChatGPT 4.1, Claude 3.5 Haiku, Gemini 1.5 e Gemini 2.0, Sabia 3.1, Sabiazinho 3 e a versão de código aberto do DeepSeek-R1 de 14B parâmetros) contra quatro exames da OAB (exames 38 e 39, de 2023, e exames 40 e 41, de 2024) revelaram três padrões principais de desempenho entre os modelos testados:

- Todos os modelos avaliados teriam sido aprovados na primeira etapa do exame, cuja nota de corte é 50%;
- O fornecimento explícito das regras jurídicas (componente R) foi, consistentemente, a informação que mais contribuiu para a melhora do desempenho dos modelos;
- A natureza da dificuldade que os modelos têm em resolver corretamente a questão jurídica depende, em alguma medida, da área do Direito. Por exemplo, no Direito Eleitoral, a principal fonte de confusão dos modelos é determinar as regras a serem aplicadas, enquanto, nas questões de Direito Civil, a maior dificuldade dos modelos reside em identificar a controvérsia central discutida no enunciado.

Este trabalho está organizado da seguinte forma: na seção Trabalhos Relacionados, discutimos estudos anteriores que exploram o uso de modelos de linguagem no domínio jurídico, com ênfase nas metodologias de avaliação de raciocínio. Em seguida, na seção Conjunto de Dados, detalhamos a seleção e

estruturação das questões dos exames da OAB utilizadas nos experimentos. A seção Metodologia apresenta nossa abordagem baseada na estrutura IRAC, descrevendo o desenho experimental e as intervenções aplicadas aos modelos. Na seção Resultados Experimentais, discutimos os achados empíricos obtidos com seis LLMs diferentes, destacando padrões de desempenho e dificuldades específicas por área do Direito. Por fim, na seção Conclusões, sintetizamos as principais contribuições do trabalho e apontamos direções promissoras para pesquisas futuras sobre a avaliação de raciocínio jurídico em modelos de linguagem.

II. TRABALHOS RELACIONADOS

A interseção entre o Direito e a Inteligência Artificial tem atraído atenção crescente à medida que profissionais jurídicos e especialistas em tecnologia exploram como os grandes modelos de linguagem (LLMs) podem ampliar ou até mesmo replicar aspectos do raciocínio jurídico. Neste sentido, pesquisadores têm examinado as implicações práticas do raciocínio jurídico assistido por IA. Schwarcz et al. [1] analisam o potencial de modelos de raciocínio e métodos RAG para transformar a prática jurídica, enquanto Waisberg e Hudek [2] descrevem como a IA já está ampliando a expertise jurídica e transformando carreiras no setor.

Uma das principais linhas de pesquisa tem sido a criação de *benchmarks* que testem as capacidades de raciocínio jurídico dos LLMs. Isso inclui avaliações de múltipla escolha, como o exame da ordem [15], [16], conjuntos de tarefas como o *LegalBench* [17], e bases de dados voltadas para tarefas jurídicas estruturadas [18], [19]. Esses *benchmarks* visam medir se os LLMs conseguem seguir a lógica jurídica, aplicar regras legais e lidar com argumentação em contextos realistas. Tais trabalhos têm ajudado a esclarecer as forças e limitações dos LLMs em diferentes jurisdições e áreas do Direito.

A metodologia que usamos no trabalho - IRAC - pode ser usada para estruturar bases de dados e *prompts* que melhoram o raciocínio jurídico dos modelos de linguagem. Kang et al. [20], por exemplo, introduzem um conjunto de dados semiestruturado baseado na metodologia IRAC — *Issue, Rule, Application* e *Conclusion* — e demonstram que alinhar os dados de treinamento à lógica jurídica melhora o desempenho dos modelos. De forma semelhante, Lu e Kao [21] avaliam diversas estratégias estruturadas de *prompting* (IRAC e variantes como IRAAC, MIRAC, TREACC) e demonstraram que *prompts* orientados por raciocínio jurídico ajudam os LLMs a dissecar e avaliar argumentos jurídicos com maior precisão, mesmo em cenários *zero-shot*. Finalmente, Reji et al. [22] combinam geração aumentada por recuperação (*retrieval-augmented generation*, RAG) com *chain-of-thought* (CoT) e *few-shot learning* para melhorar o desempenho em tarefas de inferência jurídica e também destacam os benefícios do uso de conhecimento externo e inferência estruturada. Em linhas gerais, nosso trabalho se conecta ao *LegalQA* [23], dado que as questões do exame da OAB consistem em um conjunto de perguntas e respostas voltadas para o domínio jurídico, e projetadas para avaliar o entendimento de questões

legais formuladas em linguagem natural e fornecer respostas baseadas em leis, jurisprudência e doutrina.

Enquanto os trabalhos anteriores buscam principalmente aprimorar o desempenho dos modelos, nosso foco está em compreender como esse desempenho se constrói a partir dos componentes do raciocínio jurídico. Do ponto de vista metodológico, o trabalho se aproxima de trabalhos recentes que visam fatorar as habilidades de LLMs em exames padronizados de Matemática e que descobriram que diferentes modelos têm habilidades matemáticas e de interpretação de texto independentes [24], [25].

O trabalho também se conecta a uma prática pedagógica conhecida como rastreamento de conhecimento – *knowledge tracing* [14]. *Knowledge tracing* é uma técnica usada para modelar e estimar o estado de conhecimento de um aprendiz ao longo do tempo, com base em seu desempenho em tarefas ou questões sucessivas. Embora seja tradicionalmente usada no campo da educação, sua lógica de decompor habilidades pode ser transferida para o diagnóstico de raciocínio jurídico em LLMs. Ao propormos uma abordagem para estimar a presença ou ausência de componentes específicos do raciocínio jurídico (*Issue, Rule, Application, Conclusion*) em LLMs, de forma análoga à maneira como o *knowledge tracing* rastreia a aquisição de habilidades em aprendizes humanos. Ao estruturar o raciocínio jurídico como uma sequência de etapas e observar o desempenho dos modelos frente a diferentes níveis de orientação, o estudo simula um processo em que as dicas incrementais podem ser entendidas como *intervenções instrucionais* [26] que revelam o “estado de conhecimento” latente dos LLMs — um conceito central em *knowledge tracing*.

Nosso trabalho contribui para a discussão em curso sobre uso de IAs no Direito por meio de uma abordagem diagnóstica: utilizamos uma estrutura clássica de decomposição de problemas jurídicos (IRAC) para decompor os blocos fundamentais do raciocínio jurídico e analisamos o quão bem os LLMs conseguem replicar cada componente, em uma avaliação experimental que considera exames recentes da OAB. Até onde sabemos, nenhum trabalho anterior empregou a decomposição IRAC para diagnosticar quais partes do raciocínio jurídico os modelos dominam e tratou os *prompts* como uma forma de intervenção instrucional incremental. Assim, nossa proposta complementa os avanços recentes na criação de *benchmarks* e no uso de *prompts* estruturados, oferecendo uma perspectiva mais granular sobre o desempenho de modelos no contexto jurídico.

III. CONJUNTO DE DADOS

Uma maneira robusta de avaliar a competência de LLMs em Direito é submetê-los a exames de certificação profissional. Tais exames exigem não apenas memorização de normas, mas a aplicação criteriosa de regras jurídicas a cenários factuais. No entanto, sabe-se que LLMs podem memorizar trechos de seus dados de treinamento, o que levanta dúvidas sobre se seu desempenho em tarefas jurídicas decorre de verdadeiro raciocínio ou mera reprodução. O Exame da Ordem dos Advogados do

Brasil (OAB) é uma avaliação padronizada nacional, exigida para que bacharéis em Direito obtenham licença para exercer a advocacia no Brasil. Realizado duas vezes por ano, o exame é composto por duas fases. A primeira consiste em 80 questões objetivas de múltipla escolha (A a D), abrangendo uma ampla gama de disciplinas jurídicas, como Direito Constitucional, Penal, Administrativo, Civil, do Trabalho e Ética. A segunda fase é discursiva, exigindo a elaboração de uma peça prática-profissional e a resolução de questões práticas em uma área jurídica escolhida pelo candidato.

Neste trabalho, restringimo-nos à fase objetiva do exame, por permitir uma avaliação mais sistemática e escalável das capacidades de raciocínio jurídico em LLMs — em consonância com benchmarks recentes aplicados a modelos de linguagem. Além disso, a natureza fechada das questões facilita a comparação entre humanos e modelos.

Nos experimentos, utilizamos quatro versões do exame: OAB 38 e OAB 39 (2023) e OAB 40 e OAB 41 (2024). Embora o conteúdo dos exames esteja disponível em <https://examedeordeordem.oab.org.br/>, utilizamos o conteúdo dos exames obtido a partir de <https://estrategiaconcursos.com.br/>¹²³⁴, porque eles contêm o gabarito comentado de cada questão, que usaremos para construir a sequência IRAC, conforme será detalhado na próxima seção. A título de ilustração, esta é a questão 26 da prova OAB 39, sobre Direito Tributário, e cuja resposta correta é a alternativa (c), vide Figura 1.

A Figura 2 exibe o número de questões para cada uma de 19 áreas do Direito, agregadas para os quatro exames da Ordem que consideramos nesta pesquisa. Esta informação foi obtida diretamente da classificação oficial dos exames.

IV. A METODOLOGIA IRAC

Nossa metodologia para explicitar o conhecimento que LLMs têm sobre Direito envolve gerar cinco versões de cada prova. Além do enunciado e alternativas presentes no conteúdo oficial do exame, elaboramos outras 4 versões com “dicas” — mais formalmente, no linguajar da pedagogia — uma intervenção instrucional. A intervenção foi gerada a partir do gabarito comentado descrito na Seção anterior. Cada intervenção é uma composição de componentes que seguem a metodologia IRAC — *Issue, Rule, Application, Conclusion* — uma estrutura consagrada para organizar análises jurídicas, especialmente em contextos pedagógicos como provas de Direito e na redação de peças jurídicas [11]–[13]. Ela propõe que, ao resolver um problema jurídico, estudantes e profissionais devem identificar e articular os seguintes quatro componentes:

- **Issue (Questão Jurídica):** Identificar a questão legal suscitada pelos fatos. Isso envolve a formulação do problema jurídico específico que precisa ser resolvido, frequentemente apresentado como uma pergunta (por exemplo: “O réu cometeu agressão ao empurrar o autor?”).

¹<https://gratis.estrategiaconcursos.com.br/sucesso-ebook-oab-exame-38-prova-comentada/>

²<https://gratis.estrategiaconcursos.com.br/wp-content/uploads/2023/12/OAB-39-Prova-Comentada.pdf>

³<https://gratis.estrategiaconcursos.com.br/sucesso-oab-exame-lv-prova-comentada/>

⁴<https://gratis.estrategiaconcursos.com.br/sucesso-e-book-oab-exame-41-prova-comentada/>

Exemplo de Questão do Exame da OAB

João e José receberam um imóvel residencial situado no Município Alfa por herança de seus pais. Em janeiro de, 2017, com autorização de José (menor de idade), seu irmão e tutor João (maior de idade), assina como único locador um contrato de aluguel do referido imóvel com Joaquim, com prazo determinado de 3 (três) anos, constando cláusula expressa de que o locatário será o único responsável pelo pagamento de todos os impostos e taxas do imóvel locado, exonerando o locador de tal obrigação. Em dezembro de 2021, João e José são surpreendidos com uma ação de execução fiscal movida em face de ambos pelo Município Alfa para cobrança do IPTU do imóvel locado referente a todo o exercício fiscal de 2018.

Diante desse cenário e à luz do Código Tributário Nacional, a ação de execução fiscal

a) somente poderia ter sido ajuizada em face de Joaquim, único devedor do IPTU, conforme cláusula expressa contratual.

b) somente poderia ter sido ajuizada em face de João, único que figurou no contrato como locador e dotado de capacidade tributária e processual.

c) foi corretamente ajuizada, uma vez que João e José respondem pelo tributo devido, ainda que este último seja menor de idade.

d) não podia ter sido ajuizada por já estar o crédito tributário prescrito.

Figura 1. Exemplo de Questão do Exame da OAB (Prova 39 - Questão 26)

- **Rule (Regra):** Apresentar os princípios jurídicos ou normas relevantes que regem a questão. Isso inclui os elementos legais aplicáveis, suas definições, bem como exceções e limitações.
- **Application (Aplicação):** Analisar como a regra se aplica aos fatos específicos do caso. Essa etapa exige uma articulação cuidadosa entre norma e fato, frequentemente utilizando conectores argumentativos como “porque” para justificar o raciocínio. Por ser a parte mais analítica do processo, é também onde se manifesta com maior clareza a habilidade de raciocínio jurídico.
- **Conclusion (Conclusão):** Fornecer uma resposta clara e concisa à questão jurídica, com base na aplicação da regra aos fatos. Essa etapa resume o resultado da análise, sem introduzir novos argumentos ou informações.

O IRAC é amplamente utilizado como ferramenta didática e avaliativa, justamente por permitir a decomposição do raciocínio jurídico em etapas observáveis e treináveis — uma característica que o torna particularmente útil também para fins computacionais, como neste trabalho.

Para construir nosso conjunto de dados, utilizamos os gabaritos comentados fornecidos pelo Estratégia Concursos, uma plataforma educacional brasileira especializada na preparação para concursos públicos. Esses gabaritos comentados incluem

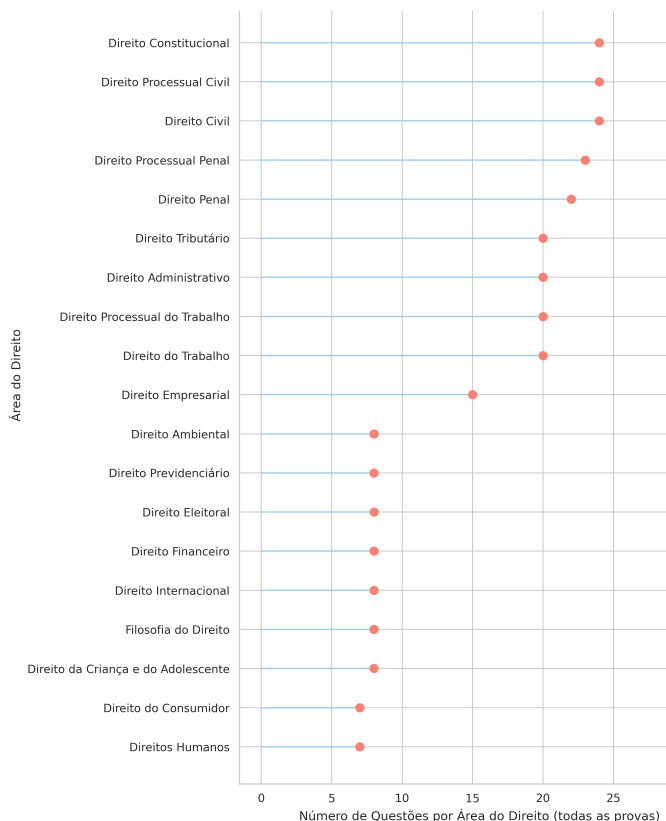


Figura 2. Número de questões por área do Direito nas quatro provas do Exame da OAB consideradas no trabalho.

não apenas as alternativas corretas de cada questão, mas também justificativas jurídicas detalhadas, oferecendo um guia sobre o raciocínio esperado pelos examinadores. Utilizamos os conjuntos de respostas oficiais e os comentários das edições 38 a 41 do exame da OAB (2023/2024), o que nos permitiu construir um conjunto de dados confiável e interpretável para avaliar o desempenho de modelos no domínio jurídico.

A Figura 3 mostra o *prompt* que converte o texto do gabarito comentado de uma questão nos componentes I, R, A e C.

O *prompt* também fornece instruções adicionais e detalha que a fatoração IRAC deve ser fornecida em formato JSON (Figura 4).

Este *prompt* foi submetido ao ChatGPT 4.1. Para a questão 26 da Prova OAB 38 ilustrada na Seção II, a decomposição IRAC é exibida na Figura 5.

De posse da decomposição IRAC para cada questão dos exames, conseguimos construir um experimento em que os modelos de linguagem são incrementalmente expostos à cada nível de intervenção instrucional.

V. RESULTADOS EXPERIMENTAIS

Nesta seção, detalhamos o projeto experimental e discutimos os resultados obtidos. Avaliamos os seguintes modelos de linguagem:

- **GPT-4.1 (OpenAI):** Um modelo de linguagem de última geração, amplamente reconhecido por suas fortes capaci-

Prompt que gera o raciocínio IRAC (Parte 1)

O framework IRAC é uma técnica consagrada no campo jurídico, amplamente utilizada para estruturar análises de casos de maneira lógica, clara e coerente. A sigla IRAC representa:

- I – Issue (Questão): identifica a questão jurídica central a ser resolvida.
- R – Rule (Regra): indicação da norma legal ou princípio aplicável.
- A – Application (Aplicação): análise de como a regra se aplica aos fatos do caso.
- C – Conclusion (Conclusão): resposta definitiva com base na aplicação da norma.

Este modelo é frequentemente empregado em petições, pareceres jurídicos e respostas discursivas em exames, como a OAB, para organizar o raciocínio jurídico.

Tarefa:

A partir do enunciado da questão e do gabarito comentado da OAB que serão fornecidos, sua tarefa é analisar o conteúdo com base no framework IRAC e extrair com precisão cada um dos quatro elementos:

- ****Issue****: identifique a controvérsia jurídica principal.
- ****Rule****: aponte a norma legal aplicável, citando expressamente o artigo, inciso ou parágrafo relevante. Sempre que possível, contextualize a norma com base no comentário apresentado no gabarito.
- ****Application****: demonstre como a norma se aplica aos fatos narrados, usando os argumentos presentes no gabarito comentado.
- ****Conclusion****: forneça a conclusão jurídica fundamentada, conforme delineada no gabarito.

Figura 3. Prompt que gera o raciocínio IRAC a partir de um gabarito comentado (Parte 1)

dades de raciocínio geral e compreensão jurídica, sendo amplamente utilizado tanto em contextos de pesquisa quanto profissionais.

- **Gemini 1.5 Flash (Google)**: Uma versão leve e eficiente da família Gemini, da Google, otimizada para respostas rápidas. Apesar da menor latência, apresenta desempenho competitivo em tarefas de compreensão linguística e benchmarks jurídicos.
- **Gemini 2.0 Flash (Google)**: Uma iteração atualizada que oferece melhorias em raciocínio, compreensão multilíngue e eficiência, buscando equilibrar velocidade com maior precisão.
- **DeepSeek-14B**: Um modelo de linguagem de código aberto com 14 bilhões de parâmetros, treinado com foco em alto desempenho de raciocínio e custo-benefício, emergindo como um forte concorrente em benchmarks acadêmicos e jurídicos.
- **Claude 3.5 Haiku (Anthropic)**: Um modelo da série Claude 3.5, projetado para ser compacto e oferecer desempenho competitivo com baixa latência.
- **Sabiá-3.1 [27]**: Um modelo de linguagem de código aberto desenvolvido no Brasil, focado em aplicações

Prompt que gera o raciocínio IRAC (Parte 2)

Instruções adicionais:

- Não invente ou assuma informações que não estejam presentes na questão ou no gabarito comentado.
- Utilize exclusivamente o conteúdo fornecido.
- Priorize sempre a clareza, a precisão técnica e a fidelidade ao material original.
- Não se refira ao gabarito comentado diretamente. Você deve evitar frases como 'De acordo com o gabarito', 'O gabarito apresenta', etc.
- Não cite a alternativa correta, ou seja, não adicione frases como "A alternativa correta é ...", "A resposta correta é a letra...", etc

'Entradas' Será fornecido como entrada uma questão e o gabarito comentado dessa questão.

Formato de saída:

A resposta deve ser apresentada no seguinte formato JSON: "I": "Descreva aqui a Issue identificada", "R": "Apresente aqui as Rules com base na legislação e comentários do gabarito", "A": "Descreva aqui a aplicação da regra aos fatos, com base nos argumentos do gabarito", "C": "Apresente aqui a conclusão jurídica final"

Figura 4. Prompt que gera o raciocínio IRAC a partir de um gabarito comentado (Parte 2)

Decomposição IRAC

```
{ "I": "Determinar se a execução fiscal para cobrança do IPTU do imóvel locado poderia ser ajuizada contra João e José, considerando a existência de cláusula contratual que transfere ao locatário a responsabilidade pelo pagamento do tributo e o fato de José ser menor de idade.", "R": "O Código Tributário Nacional dispõe em seu art. 123 que as convenções particulares relativas à responsabilidade pelo pagamento de tributos não podem ser opostas à Fazenda Pública para modificar a definição legal do sujeito passivo das obrigações tributárias. O art. 124, inciso I, estabelece que são solidariamente obrigadas as pessoas que tenham interesse comum na situação que constitua o fato gerador da obrigação principal. O art. 126 determina que a capacidade tributária passiva independe da capacidade civil das pessoas naturais.", "A": "Ainda que o contrato de locação preveja que o locatário é o responsável pelo pagamento do IPTU, tal convenção não pode ser oposta ao Município para afastar a responsabilidade dos proprietários do imóvel, que são os sujeitos passivos do tributo. João e José, como coproprietários do imóvel, possuem interesse comum na situação que constitui o fato gerador do IPTU, respondendo solidariamente pelo tributo. O fato de José ser menor de idade não o exime da responsabilidade tributária, pois a capacidade tributária passiva independe da capacidade civil.", "C": "A execução fiscal foi corretamente ajuizada contra João e José, pois ambos, na qualidade de proprietários do imóvel, são solidariamente responsáveis pelo pagamento do IPTU, independentemente de cláusula contratual em sentido contrário e da menoridade de José." }
```

Figura 5. Decomposição IRAC para a Questão 26 da Prova OAB 38

em língua portuguesa. Destaca-se por sua afinidade com contextos jurídicos e administrativos nacionais e por sua transparência metodológica.

- **Sabiazinho-3 [27]:** Versão compacta do Sabiá-3.1, projetada para aplicações mais leves ou embarcadas, mantendo boa performance em tarefas em português, especialmente em compreensão textual e raciocínio jurídico elementar.

A Figura 6 exhibe a taxa de acerto obtida por cada modelo ao resolver cada prova (OAB 38, OAB 39, OAB 40, OAB 41), sem qualquer dica ou informação proveniente do gabarito comentado. Observamos as seguintes consistências:

- 1) Todos os modelos têm desempenho melhor nas provas OAB 40 e 41 do que nas provas OAB 38 e 39;
- 2) Como esperado, o desempenho do Gemini 2.0 é superior ao Gemini 1.5 em todas as provas;
- 3) Da mesma maneira, o modelo Sabia-3.1 supera o modelo Sabiazinho-3;
- 4) O modelo DeepSeek 14B tem o desempenho mais fraco, o que é esperado dado que ele é significativamente menor que os outros modelos.

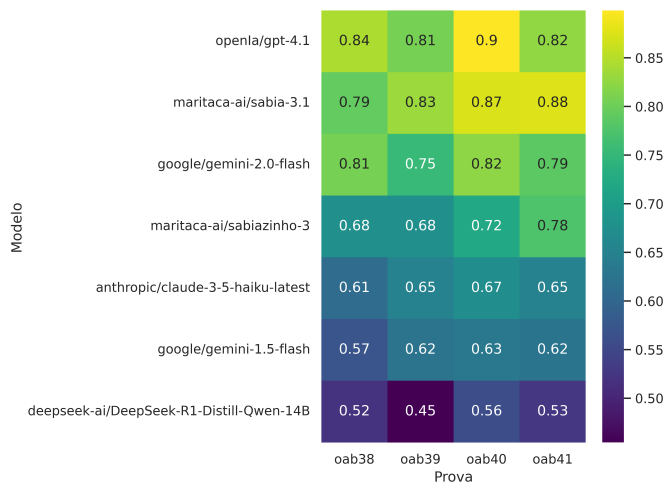


Figura 6. Taxa de acerto por prova e modelo de linguagem.

Para melhor entender nuances e particularidades de cada modelo, empregamos agora a estratégia de entregar no *prompt* dicas incrementais segundo a metodologia IRAC. Note que essas intervenções são incrementais; depois a versão da prova em que a informação I (*Issue*) é passada ao modelo, na próxima versão da prova, será informado I+R, ou seja, a *Issue* e as regras R *Rules*. A Figura 7 mostra como muda a taxa de acerto (média para as quatro provas) conforme aumentamos o nível de dicas.

Veja que conseguimos, agora, observar detalhes e diferenças entre os modelos que ficam ocultos na Figura 6:

- O modelo de pior desempenho na prova original, DeepSeek-14B, melhora significativamente de desempenho quando instruído com a *Issue* (I) e mais ainda quando instruído com a Regra que deve ser aplicada (R). Isso explicita quais as dificuldades do modelo: ele

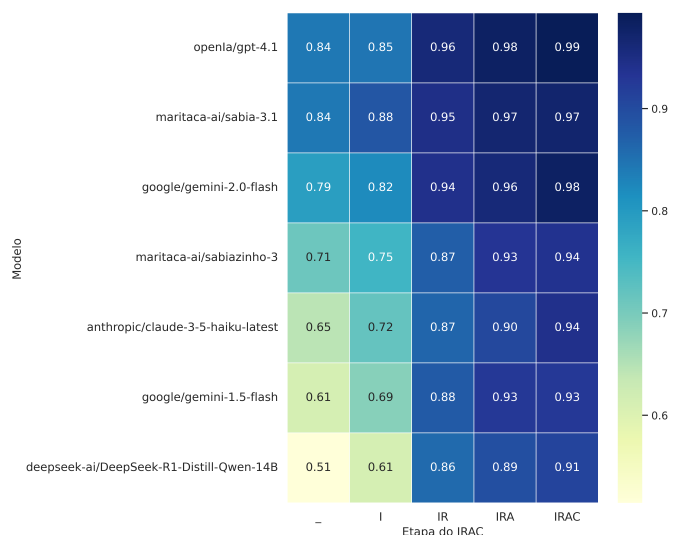


Figura 7. Taxa de acerto média por prova, com intervenções instrucionais incrementais seguindo a metodologia IRAC.

falha parcialmente em entender a controvérsia jurídica apresentada, e também falha em recuperar quais regras devem aplicar e o que elas dizem. Uma vez que reste apenas a etapa de aplicar a lei e concluir, seu desempenho sobe para 86%.

- Modelos como DeepSeek-14B e Claude 3.5 Haiku erram cerca de 10% das questões mesmo quando minuciados com o gabarito comentado das etapas I, R e A da estrutura IRAC, o que evidencia que estes modelos, para as provas e *prompts* utilizados, ainda demonstram fragilidades relevantes para tomarem decisões jurídicas. DeepSeek-14B, em particular, ainda erra 6% mesmo quando recebe toda a resolução IRAC na integralidade.

Eis um exemplo extraído da Questão 45 da prova 41 da OAB, sobre Direito do Consumidor, em que o modelo DeepSeek se equivocou sobre a resposta correta mesmo quando instruído de todas as etapas IRAC extraídas do gabarito comentado.

Nísia adquiriu um fogão, que apresentou defeito de superaquecimento logo no primeiro uso, em 20 de setembro. Ela registrou reclamação junto ao SAC do fabricante em 22 de setembro, que respondeu negativamente apenas no dia 30. Em 13 de outubro, Nísia protocolou reclamação no órgão estadual de defesa do consumidor. De acordo com o Código de Defesa do Consumidor, qual é a afirmativa correta sobre o prazo decadencial?

A alternativa correta é a) O prazo é de 90 dias e é suspenso pela reclamação ao fabricante até a resposta. A etapa "C" do IRAC diz que o prazo decadencial de 90 dias para vícios em produtos duráveis é suspenso a partir da reclamação ao fornecedor até o recebimento da resposta negativa. Porém, o modelo DeepSeek concluiu que "o prazo decadencial é de 90 dias, e ele foi obstado pela reclamação feita por Nísia ao fabricante até a resposta negativa. Portanto, o prazo só será retomado após a resposta negativa, caso o problema

persista."Ou seja, o LLM reconhece corretamente a suspensão do prazo pela reclamação dirigida ao fabricante, mas comete erro ao selecionar a alternativa c): “O prazo deve ser de 90 dias, sendo obstado pela reclamação formalizada por Nísia perante o órgão estadual de defesa do consumidor, devendo ser retomado 90 dias depois da data da reclamação, caso o problema persista.”. Tal alternativa menciona incorretamente a suspensão do prazo com base na atuação do órgão de defesa do consumidor. Observa-se que, embora a entidade tenha sido corretamente identificada no raciocínio do modelo, houve uma aplicação inconsistente desse entendimento na escolha da alternativa.

Para observar com mais clareza o impacto das intervenções instrucionais, exibimos na Figura 8 a distribuição da taxa de acerto por questão, considerando todas as quatro provas. O triângulo verde representa a taxa de acerto média e a reta horizontal laranja a taxa mediana. Observamos que fornecer aos modelos a controvérsia fundamental da questão (I) melhora um pouco as taxas de acerto; no entanto, o maior ganho claramente está em fornecer aos modelos a descrição da legislação (R) que eles devem aplicar para responder cada questão. Cabe aqui a ressalva de que, como a estrutura IRAC foi montada a partir de gabaritos comentados, existe uma tendência natural da descrição R já estar bastante focada na parte da legislação que interessa para a resolução da questão jurídica proposta (veja o exemplo na Seção III); um cenário mais desafiador seria se o conteúdo integral da lei estivesse à disposição do modelo. De qualquer maneira, depois que a controvérsia jurídica I e as regras R foram fornecidas, fornecer a aplicação das regras A não afeta muito a média e a mediana da taxa de acertos, mas reduz a variância fazendo com que os modelos acertem algumas questões que ainda eram solucionadas de forma incorreta.

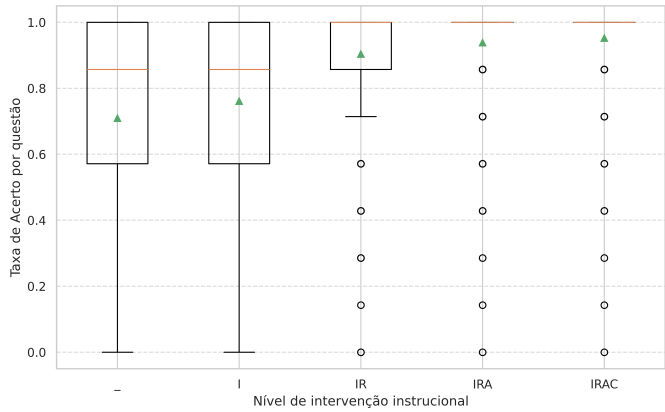


Figura 8. Evolução da taxa de acerto absoluta por questão conforme informações IRAC são fornecidas aos modelos. O triângulo verde representa a taxa de acerto média e a linha horizontal laranja representa a mediana.

Finalmente, analisamos o efeito de fornecer ao modelo as informações IRAC por área do Direito – uma classificação por questão que consta no exame oficial da OAB. Na Figura 9, exibimos a variação da taxa de acerto conforme fornecemos mais informação aos modelos e reportamos a média para todas

as questões. Por exemplo, com todos os modelos acertam 100% das questões de filosofia do Direito, relativamente o incremento após as intervenções I, R, A e C foi 0%.

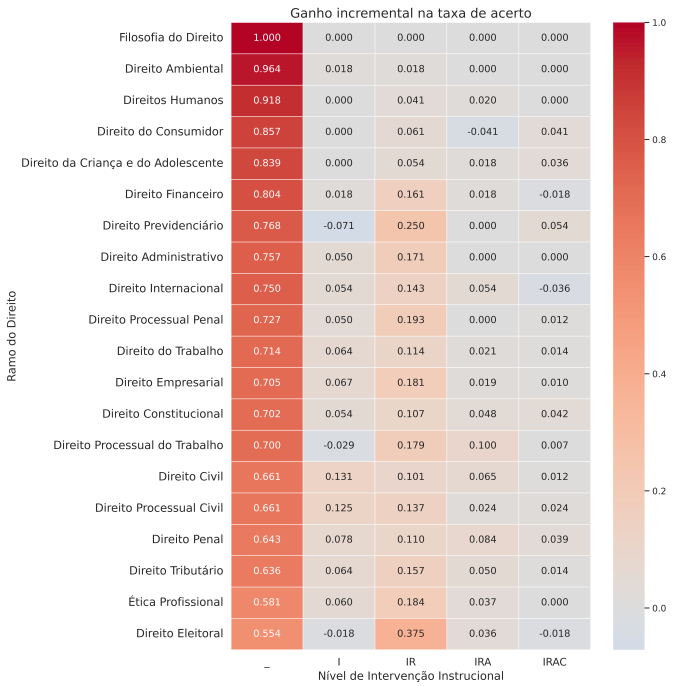


Figura 9. Mapa de calor da taxa de acerto relativa por área do Direito, conforme informações IRAC são fornecidas ao modelo. A primeira coluna é a taxa de acerto sem qualquer intervenção; os valores seguintes correspondem apenas à mudança observada na taxa de acerto após a intervenção.

Primeiramente, verificamos que, na ausência de qualquer intervenção instrucional, há diferenças entre a taxa de acerto por área do Direito. Os modelos têm mais facilidade com Filosofia do Direito, Direito Ambiental e Direitos Humanos, enquanto as áreas mais desafiadoras são Direito Tributário, Ética Profissional e Direito Eleitoral. As três primeiras áreas compreendem conteúdo mais geral do Direito, com menos leis específicas e leis brasileiras, o que pode explicar o alto desempenho dos modelos mesmo sem qualquer dica proveniente da estrutura IRAC.

Conseguimos observar, ainda que diferentes áreas do Direito são afetadas de forma diferente pelas intervenções instrucionais com IRAC. Em particular, Direito Eleitoral é a área que mais se beneficia da instrução R; o que pode indicar que esta área do Direito detém regras e regulamentação extensas ou complexas, que os modelos não conseguem recuperar. Uma vez que as regras são fornecidas, a taxa de acerto sobe muito. Por outro lado, Direito Civil e Direito Processual Civil são as áreas do Direito que mais se beneficiam caso os modelos sejam instruídos com a descrição da controvérsia I; o que pode ser um sinal de que estas áreas do Direito são mais amplas e em que há desafio, para os modelos, e identificar qual é a questão central proposta nos enunciados dos exames da OAB.

VI. CONCLUSÕES E TRABALHOS FUTUROS

O Direito se destaca como uma das áreas onde os grandes modelos de linguagem (LLMs) apresentam grande potencial de aplicação, dada a sua complexidade e volume de informação textual. Contudo, para avaliar adequadamente o desempenho dessas tecnologias, é imprescindível a adoção de avaliações especializadas e específicas para o domínio jurídico, que reflitam as nuances e desafios particulares do campo.

Neste trabalho, utilizamos a metodologia clássica de fato-razão das questões jurídicas, o método IRAC (*Issue-Rule-Application-Conclusion*), para desmembrar e analisar o desempenho dos LLMs no contexto do Exame da OAB conforme eram instruídas com as informações I/R/A/C provenientes de gabaritos comentados disponíveis publicamente. Em nosso método proposto, estruturamos os *prompts* jurídicos e analisamos as respostas dos LLMs à luz do modelo IRAC. Assim, investigamos de forma sistemática o quanto esses modelos replicam os processos cognitivos envolvidos na análise jurídica, com foco específico em sua capacidade de realizar cada etapa da estrutura IRAC. Nossos resultados evidenciam que diferentes áreas do Direito impõem desafios distintos aos modelos, indicando a necessidade de abordagens mais refinadas e contextuais para sua avaliação.

Essas descobertas abrem caminhos para integrar técnicas pedagógicas de rastreamento de conhecimento (*knowledge tracing*) com o objetivo de modelar como os LLMs adquirem e aplicam padrões de raciocínio específicos de determinado domínio, permitindo uma avaliação mais detalhada de suas dinâmicas de aprendizagem e retenção de conhecimento em diferentes tarefas. Como trabalhos futuros, vamos trabalhar em projetar métricas que considerem a estabilidade, a quantidade e a complexidade da informação legal, bem como a arbitrariedade inerente às leis, visando uma avaliação mais robusta e significativa do potencial e das limitações dos advogados artificiais. Além disso, vamos experimentar diferentes idiomas para observar como isso afeta o raciocínio ou capacidade dos modelos em recuperar as informações legais corretas.

AGRADECIMENTOS

Esse trabalho foi parcialmente financiado por CNPq, CAPES, FAPEMIG, Projeto IAIA-INCT em IA, AWS, e UFMG.

REFERÊNCIAS

- [1] D. Schwarcz, S. Manning, P. Barry, D. R. Cleveland, J. Prescott, and B. Rich, "Ai-powered lawyering: Ai reasoning models, retrieval augmented generation, and the future of legal practice," 2025.
- [2] N. Waisberg and S. Hudek, "Ai for lawyers: How artificial intelligence is adding value," *Amplifying Expertise, and Transforming Careers*, 2021.
- [3] R. Thomas, P. Zikopoulos, and K. Soule, *AI Value Creators: Beyond the Generative AI User Mindset*. O'Reilly Media, 2025. [Online]. Available: <https://www.amazon.com.br/dp/1098168348>
- [4] C. Barale, M. Rovatsos, and N. Bhuta, "When fairness isn't statistical: The limits of machine learning in evaluating legal reasoning," *arXiv preprint arXiv:2506.03913*, 2025.
- [5] J. W. Howarth, "What law must lawyers know?" *Conn. Pub. Int. LJ*, vol. 19, p. 1, 2019.
- [6] J. X. Morris, C. Sitawarin, C. Guo, N. Kokhlikyan, G. E. Suh, A. M. Rush, K. Chaudhuri, and S. Mahloui, "How much do language models memorize?" 2025. [Online]. Available: <https://arxiv.org/abs/2505.24832>
- [7] P. Shojaei*, I. Mirzadeh*, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar, "The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity," 2025. [Online]. Available: <https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf>
- [8] J. Wu, S. Liu, H. Tu, H. Yu, X. Huang, J. Zou, C. Xie, and Y. Zhou, "Knowledge or reasoning? a close look at how llms think across domains," 2025. [Online]. Available: <https://arxiv.org/abs/2506.02126>
- [9] C. Xie, Y. Huang, C. Zhang, D. Yu, X. Chen, B. Y. Lin, B. Li, B. Ghazi, and R. Kumar, "On memorization of large language models in logical reasoning," *arXiv preprint arXiv:2410.23123*, 2024.
- [10] V. Mishra, B. Pathiraja, M. Parmar, S. Chidananda, J. Srinivasa, G. Liu, A. Payani, and C. Baral, "Investigating the shortcomings of LLMs in step-by-step legal reasoning," in *Findings of the Association for Computational Linguistics: NAACL 2025*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 7795–7826. [Online]. Available: <https://aclanthology.org/2025.findings-naacl.435/>
- [11] K. Burton, "'think like a lawyer' using a legal reasoning grid and criterion-referenced assessment rubric on irac (issue, rule, application, conclusion)." *Journal of Learning Design*, vol. 10, no. 2, pp. 57–68, 2017.
- [12] —, "Teaching and assessing problem solving: An example of an incremental approach to using irac in legal education," *Journal of University Teaching and Learning Practice*, vol. 13, no. 5, 2016.
- [13] Touro Law Center, "Irac: Law school essay writing," n.d., accessed: 2025-05-25. [Online]. Available: <https://www.tourolaw.edu/adp/studyskills/irac.aspx>
- [14] G. Abdelrahman, Q. Wang, and B. Nunes, "Knowledge tracing: A survey," *ACM Comput. Surv.*, vol. 55, no. 11, Feb. 2023. [Online]. Available: <https://doi.org/10.1145/3569576>
- [15] E. Martínez, "Re-evaluating gpt-4's bar exam performance," *Artificial intelligence and law*, pp. 1–24, 2024.
- [16] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "Gpt-4 passes the bar exam," *Philosophical Transactions of the Royal Society A*, vol. 382, no. 2270, p. 20230254, 2024.
- [17] N. Guha, J. Nyarko, D. Ho, C. Ré, A. Chilton, A. Chohlas-Wood, A. Peters, B. Waldon, D. Rockmore, D. Zambrano *et al.*, "Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 123–44 279, 2023.
- [18] Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, A. Huang, S. Zhang, K. Chen, Z. Yin, Z. Shen, J. Ge, and V. Ng, "LawBench: Benchmarking legal knowledge of large language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2024, pp. 7933–7962.
- [19] L. Zheng, N. Guha, J. Arifov, S. Zhang, M. Skreta, C. D. Manning, P. Henderson, and D. E. Ho, "A reasoning-focused legal retrieval benchmark," in *Proceedings of the 2025 Symposium on Computer Science and Law*, 2025, pp. 169–193.
- [20] X. Kang, L. Qu, L.-K. Soon, Z. Li, and A. Trakic, "Bridging law and data: Augmenting reasoning via a semi-structured dataset with irac methodology," 2024. [Online]. Available: <https://arxiv.org/abs/2406.13217>
- [21] Y.-a. Lu and H.-y. Kao, "Ox.Yuan at SemEval-2024 task 5: Enhancing legal argument reasoning with structured prompts," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 385–390. [Online]. Available: <https://aclanthology.org/2024.semeval-1.60/>
- [22] S. A. Reji, R. Sheik, S. A. A. R. M., and S. Jaya Nirmala, "Enhancing llm performance on legal textual entailment with few-shot cot-based rag," in *2024 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, 2024, pp. 1–6.
- [23] A. Abdallah, B. Piriyani, and A. Jatowt, "Exploring the state of the art in legal qa systems," *Journal of Big Data*, vol. 10, no. 1, p. 127, 2023.
- [24] P. Calais, G. Franco, Z. Tang, T. Nikas, W. Meira Jr, E. T. Terzi, and M. Crovella, "Disentangling text and math in word problems: Evidence for the bidimensional structure of large language models' reasoning," in *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, Jul. 2025.
- [25] Z. Cheng, M. Cao, L. Pishdad, Y. Cao, and J. C. K. Cheung, "Can llms reason abstractly over math word problems without cot? disentangling abstract formulation from arithmetic computation," 2025. [Online]. Available: <https://arxiv.org/abs/2505.23701>
- [26] I. Šarić-Grgić, A. Grubišić, and A. Gašpar, "Twenty-five years of bayesian knowledge tracing: a systematic review," *User Modeling and User-Adapted Interaction*, vol. 34, no. 4, pp. 1127–1173, 2024.
- [27] R. Pires, H. Abonizio, T. S. Almeida, and R. Nogueira, "Sabiá: Portuguese large language models," in *Brazilian Conference on Intelligent Systems*. Springer, 2023, pp. 226–240.