UNIVERSIDADE FEDERAL DE GOIÁS (UFG)

INSTITUTO DE INFORMÁTICA (INF)

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO (PPGCC)

JOÃO PAULO CAVALCANTE PRESA

Avaliação de Grandes Modelos de Linguagem para Raciocínio em Direito Tributário

GOIÂNIA
2024

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

João Paulo Cavalcante Presa

3. Título do trabalho

Avaliação de Grandes Modelos de Linguagem para Raciocínio em Direito Tributário

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **João Paulo Cavalcante Presa**, **Discente**, em 17/01/2025, às 15:26, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Celso Goncalves Camilo Junior**, **Professor do Magistério Superior**, em 20/01/2025, às 15:01, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5099244** e o código CRC **17131EF0**.

Referência: Processo nº 23070.052711/2024-17 SEI nº 5099244

JOÃO PAULO CAVALCANTE PRESA

Avaliação de Grandes Modelos de Linguagem para Raciocínio em Direito Tributário

Dissertação apresentada ao Programa de Pós–Graduação em Ciência da Computação (PPGCC) do Instituto de Informática (INF) da Universidade Federal de Goiás (UFG), como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação.

Linha de Pesquisa: Sistemas Inteligentes e Aplicações

Orientador: Prof. Dr. Celso Gonçalves Camilo Junior

Co-Orientador: Prof. Dr. Sávio Salvarino Teles de Oliveira

GOIÂNIA
2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Cavalcante Presa, João Paulo
Avaliação de Grandes Modelos de Linguagem para Raciocínio em Direito Tributário [manuscrito] / João Paulo Cavalcante Presa. - 2024.
LXXVII, 77 f.

Orientador: Prof. Dr. Celso Gonçalves Camilo Junior; co-orientador Dr. Sávio Salvarino Teles de Oliveira.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2024.
Bibliografia. Apêndice.
Inclui lista de figuras, lista de tabelas.

1. Processamento de Linguagem Natural. 2. Grandes Modelos de Linguagem (LLM). 3. Raciocínio Jurídico. 4. Direito Tributário. I. Gonçalves Camilo Junior, Celso, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **41** da sessão de Defesa de Dissertação de **João Paulo Cavalcante Presa**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos vinte e dois dias do mês de novembro de dois mil e vinte e quatro, a partir das treze e trinta horas, via webconferência, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Avaliação de Grandes Modelos de Linguagem para Raciocínio em Direito Tributário"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Celso Gonçalves Camilo Júnior (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Sávio Salvarino Teles de Oliveira (INF/UFG), coorientador; Professora Doutora Nádia Felix Felipe da Silva (INF/UFG), membra titular interna; Professora Doutora Karla Tereza Figueiredo Leite (UERJ), membra titular externo. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Celso Gonçalves Camilo Júnior, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos vinte e dois dias do mês de novembro de dois mil e vinte e quatro.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Nadia Felix Felipe Da Silva**, **Professora do Magistério Superior**, em 22/11/2024, às 15:21, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Karla Tereza Figueiredo Leite**, **Usuário Externo**, em 22/11/2024, às 15:22, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Savio Salvarino Teles De Oliveira**, **Professor do Magistério Superior**, em 22/11/2024, às 15:22, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Celso Goncalves Camilo Junior**, **Professor do Magistério Superior**, em 22/11/2024, às 15:22, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **João Paulo Cavalcante Presa**, **Discente**, em 22/11/2024, às 18:12, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Referência: Processo nº 23070.052711/2024-17 SEI nº 4980568

João Paulo Cavalcante Presa

Graduou-se em Ciência da Computação na Universidade Federal de Goiás (UFG). Após a graduação, trabalhou com sistemas judiciais no Tribunal de Justiça de Goiás. Ingressou no programa de mestrado da UFG, onde foi bolsista no Centro de Excelência em Inteligência Artificial (CEIA). Atualmente, sua área de interesse envolve Modelos de Linguagem de Grande Escala (LLMs) e suas aplicações no Direito, desenvolvendo trabalhos nessa interseção.

Dedico este trabalho à minha esposa, pelo apoio incondicional em todos os momentos, e à Universidade Federal de Goiás, por proporcionar as oportunidades que me permitiram chegar até aqui.

Agradecimentos

À minha amada esposa, Marcela Faria Gil Presa, meu porto seguro nesta jornada acadêmica, agradeço pelo amor e apoio inabaláveis que iluminaram até os dias mais sombrios.

Ao meu orientador, Celso Camilo, expresso minha mais profunda gratidão. Sua fé inabalável em meu potencial foi o farol que me guiou de volta ao programa quando me via à deriva e quando me desliguei do programa. Sua orientação foi muito além da academia; foi um exercício de compaixão e respeito que jamais esquecerei. Seu conhecimento compartilhado e paciência infindável foram os alicerces sobre os quais construí esta dissertação.

Ao meu coorientador, Sávio Teles de Oliveira, ofereço meu sincero reconhecimento. Sua dedicação transcendeu as expectativas, com horas incontáveis de esforço, uma orientação dedicada e um comprometimento que ultrapassou os limites do dever. Seu empenho incansável e sua sabedoria foram essenciais para elevar este trabalho a novos patamares.

Agradeço também aos professores que me guiaram com seus ensinamentos e à toda equipe de funcionários da Universidade Federal de Goiás, cujo trabalho diário contribuiu para que este percurso fosse possível.

"Eu proponho considerar a questão: As máquinas podem pensar?"

Alan Turing,
Computing Machinery and Intelligence.

Resumo

Presa, João. **Avaliação de Grandes Modelos de Linguagem para Raciocínio em Direito Tributário**. GOIÂNIA, 2024. 76p. Dissertação de Mestrado. Programa de Pós-Graduação em Ciência da Computação (PPGCC), Instituto de Informática (INF), Universidade Federal de Goiás (UFG).

O direito tributário é fundamental para a regulação das relações entre Estado e contribuintes, sendo essencial para a arrecadação de tributos e manutenção das funções públicas. A complexidade e constante evolução das legislações tributárias tornam sua interpretação um desafio contínuo para os operadores do direito. Embora o Processamento de Linguagem Natural (PLN) tenha se consolidado como uma tecnologia promissora no campo jurídico, sua aplicação no contexto do direito tributário brasileiro, especialmente para entidades jurídicas, permanece uma área relativamente inexplorada. Este trabalho avalia o uso de Grandes Modelos de Linguagem (LLMs) no direito tributário brasileiro da União, analisando sua capacidade de processar perguntas e gerar respostas em português para consultas de pessoas jurídicas. Para isso, foi construído um conjunto de dados original composto por perguntas reais e respostas fornecidas por especialistas, permitindo avaliar a capacidade dos LLMs, tanto proprietário quanto de código aberto, de gerar respostas juridicamente válidas. A pesquisa utiliza métricas quantitativas e qualitativas para medir a acurácia e relevância das respostas geradas, capturando aspectos do raciocínio jurídico e da coerência semântica. Como contribuições, o trabalho apresenta um conjunto de dados específico para o domínio do direito tributário, uma avaliação detalhada do desempenho de diferentes LLMs na tarefa de raciocínio jurídico e uma abordagem de avaliação que integra métricas quantitativas e qualitativas, promovendo assim o avanço da aplicação da inteligência artificial na análise de leis e regulamentos tributários.

Palavras–chave

Direito Tributário, Raciocínio Jurídico, Grandes Modelos de Linguagem (LLM), Processamento de Linguagem Natural

Abstract

Presa, João. **Evaluating Large Language Models for Tax Law Reasoning**. GOIÂNIA, 2024. 76p. MSc. Dissertation. Programa de Pós-Graduação em Ciência da Computação (PPGCC), Instituto de Informática (INF), Universidade Federal de Goiás (UFG).

Tax law is essential for regulating relationships between the State and taxpayers, being crucial for tax collection and maintaining public functions. The complexity and constant evolution of tax laws make their interpretation an ongoing challenge for legal professionals. Although Natural Language Processing (NLP) has become a promising technology in the legal field, its application in brazilian tax law, especially for legal entities, remains a relatively unexplored area. This work evaluates the use of Large Language Models (LLMs) in Brazilian tax law covering federal tax aspects, analyzing their ability to process questions and generate answers in Portuguese for legal entities' queries. For this purpose, we built an original dataset composed of real questions and answers provided by experts, allowing us to evaluate the ability of both proprietary and open-source LLMs to generate legally valid answers. The research uses quantitative and qualitative metrics to measure the accuracy and relevance of generated answers, capturing aspects of legal reasoning and semantic coherence. As contributions, this work presents a dataset specific to the tax law domain, a detailed evaluation of different LLMs' performance in legal reasoning tasks, and an evaluation approach that combines quantitative and qualitative metrics, thus advancing the application of artificial intelligence in the analysis of tax laws and regulations.

Keywords

Tax Law, Legal Reasoning, Large Language Models (LLMs), Natural Language Processing

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

O direito tributário é essencial para a prática jurídica e desempenha um papel importante na regulação das relações entre o Estado e os contribuintes, sendo fundamental para a arrecadação de tributos, a manutenção das funções públicas e o controle das atividades econômicas [2]. A natureza complexa e a constante evolução das legislações tributárias tornam a interpretação e aplicação dessas normas um desafio contínuo para os operadores do direito, que enfrentam ambiguidades legislativas e a necessidade de atualização constante [3]. Esse cenário exige habilidades apuradas de raciocínio jurídico para assegurar o cumprimento justo e eficaz das obrigações tributárias, ao mesmo tempo, em que garante a segurança jurídica e a equidade fiscal [4].

Paralelamente, o Processamento de Linguagem Natural (PLN) tem se consolidado como uma tecnologia com grande impacto no campo jurídico, abrindo novas possibilidades para a análise, interpretação e organização de grandes volumes de dados legais [5]. Especificamente, os Grandes Modelos de Linguagem (LLMs) têm mostrado um grande potencial para a automação de processos jurídicos, incluindo tarefas como interpretação de textos legais, predição de resultados judiciais e respostas a perguntas complexas (Q&A) [6, 7]. Contudo, a aplicação desses modelos no contexto do direito tributário brasileiro, especialmente no que concerne a entidades jurídicas, permanece uma área relativamente inexplorada. A complexidade das normas tributárias brasileiras e as nuances da língua portuguesa impõem desafios significativos para a adoção eficaz dessas tecnologias [8, 9].

Este trabalho tem como objetivo avaliar a capacidade dos LLMs em realizar raciocínio jurídico no domínio do direito tributário, com foco em perguntas e respostas formuladas em português brasileiro para pessoas jurídicas. A pesquisa utiliza um conjunto de dados original composto por perguntas reais e respostas fornecidas por especialistas, com o objetivo de verificar se os LLMs, tanto de código aberto quanto proprietários, conseguem gerar respostas precisas, juridicamente válidas e adequadas ao contexto tributário corporativo do ponto de vista dos contribuintes e profissionais do direto. A metodologia emprega métricas quantitativas, como BLEU [10], ROUGE [11] e *Bert Score* [12], além de uma avaliação qualitativa [13–19] que visa garantir a acurácia factual e a relevância

semântica das respostas geradas. Os resultados demonstram uma forte correlação entre a métrica de avaliação qualitativa e o *Bert Score F1*, sugerindo que essas métricas capturam de forma eficaz aspectos semânticos relevantes para o contexto jurídico.

A relevância desta pesquisa está na crescente demanda por ferramentas tecnológicas capazes de auxiliar os profissionais do direito a lidarem com a enorme quantidade de dados legais e na interpretação precisa das normas tributárias [20]. A adoção de LLMs tem o potencial de transformar a prática jurídica, automatizando processos repetitivos e liberando os operadores do direito para se concentrarem em questões mais complexas e estratégicas [21]. Além disso, o estudo investiga a aplicação de técnicas de geração aumentada, que visam avaliar o raciocínio para geração das respostas fornecidas pelos modelos, oferecendo uma abordagem para a automação de processos complexos de compreensão textual [22].

Ao explorar o potencial desses modelos no contexto do direito tributário brasileiro, esta pesquisa contribui para o avanço da inteligência artificial aplicada ao campo jurídico [5]. Ela oferece soluções práticas para os desafios enfrentados pelos profissionais na interpretação e aplicação das leis tributárias, promovendo maior eficiência e segurança jurídica nas relações entre o Estado e as entidades jurídicas [23]. A fundamentação teórica deste estudo se apoia em uma revisão abrangente da literatura sobre direito tributário, raciocínio jurídico e PLN, investigando a interseção dessas áreas [24].

Ao avaliar a eficácia dos LLMs em tarefas de raciocínio jurídico no domínio do direito tributário, este estudo visa expandir o conhecimento acadêmico sobre a aplicação da inteligência artificial no direito, promovendo a integração dessas tecnologias como ferramentas para enfrentar os desafios cada vez mais complexos da prática jurídica [25, 26]. Espera-se que os resultados desta pesquisa possam servir como base para iniciativas futuras que busquem aprimorar a justiça a inteligência artificial aplicada ao direito tributário.

1.1 Problema de Pesquisa

A crescente complexidade e volume da legislação tributária brasileira representam um desafio significativo para profissionais do direito e contribuintes. A interpretação e aplicação corretas das leis tributárias exigem não apenas um conhecimento aprofundado das normas vigentes, mas também a capacidade de relacioná-las a situações específicas, que muitas vezes envolvem nuances e detalhes complexos.

Esse contexto exige uma investigação mais profunda sobre a capacidade desses modelos de interpretar e responder perguntas de maneira juridicamente válida e precisa, respeitando os princípios de segurança jurídica e equidade fiscal, enquanto enfrentam a complexidade inerente ao sistema tributário nacional. O problema desta dissertação reside

na lacuna existente na aplicação de LLMs para o domínio específico do direito tributário brasileiro, especialmente no atendimento às necessidades jurídicas de pessoas jurídicas.

Este problema é motivado pela necessidade de ferramentas que possam auxiliar na interpretação da legislação tributária, facilitando o acesso à informação jurídica e apoiando profissionais do direito em suas atividades. No entanto, os desafios incluem a complexidade da linguagem legal, a necessidade de contextualização das normas e a capacidade dos modelos de lidar com atualizações constantes nas leis.

1.1.1 Perguntas de Pesquisa

Esta pesquisa procura responder às seguintes perguntas de pesquisa:

1. Os LLMs conseguem interpretar e aplicar corretamente as normas tributárias brasileiras em perguntas e respostas relacionadas a entidades jurídicas?
2. Quais são as limitações dos LLMs ao lidar com a complexidade das normas tributárias e as nuances da língua portuguesa em tarefas de perguntas e respostas no direito tributário?

1.1.2 Hipóteses

Com base nas perguntas de pesquisa, foram formuladas as seguintes hipóteses:

1. **H1**: Os LLMs são capazes de gerar respostas juridicamente válidas e adequadas ao contexto tributário brasileiro, principalmente em LLMs com maior número de parâmetros.
2. **H2**: A complexidade das normas tributárias e as nuances da língua portuguesa impõem limitações significativas à acurácia dos LLMs em tarefas de raciocínio no direito tributário.

1.2 Objetivos

O objetivo geral desta pesquisa é avaliar a eficácia dos LLMs no raciocínio jurídico aplicado ao direito tributário para pessoas jurídicas.

Os objetivos específicos são:

- Desenvolver um conjunto de dados composto por perguntas reais de direito tributário, respostas de especialistas e os textos legais correspondentes.
- Aplicar diferentes LLMs, tanto de código aberto quanto proprietário na tarefa do conjunto de dados.

- Propor e analisar uma avaliação quantitativa e qualitativa das respostas geradas pelos LLMs.
- Analisar as capacidades e limitações atuais dos LLMs no tratamento de tarefas de raciocínio jurídico, identificando áreas para melhorias futuras.

1.3 Contribuições

No trabalho, foram apresentadas contribuições relevantes no campo do direito tributário e do PLN. Primeiramente, foi introduzido um conjunto de dados específico para o domínio do direito tributário, composto por perguntas reais de contribuintes, respostas elaboradas por especialistas e os textos legais de suporte. Este conjunto de dados representa um recurso para pesquisas futuras em PLN aplicada ao direito.

Além disso, foi realizada uma avaliação detalhada do desempenho de diferentes LLMs na tarefa de raciocínio jurídico. Essa avaliação fornece informações sobre a capacidade desses modelos de compreender questões complexas, aplicar normas legais e gerar respostas coerentes e precisas. Foi desenvolvida uma abordagem de avaliação de LLMs em raciocínio jurídico que integra métricas quantitativas e qualitativas, que também foi avaliada e comparada.

Foram identificadas oportunidades de melhoria no problema e foram propostas direções para pesquisas futuras, visando melhorar a capacidade dos modelos de utilizar corretamente as provisões legais relevantes. O que leva a considerar que houve contribuição para a literatura sobre a aplicação de LLMs no direito, com foco específico no direito tributário, uma área complexa e em constante evolução.

1.4 Publicações

O artigo intitulado "*Evaluating Large Language Models for Tax Law Reasoning*" [27], cujo os autores são João Paulo Presa, Celso Camilo e Sávio Teles de Oliveira, foi aceito para publicação nos anais da 34th Brazilian Conference on Intelligent Systems (*BRACIS*) com Qualis A4, que ocorrerá de 17 a 24 de novembro de 2024, em Belém, PA, Brasil.

Enquanto a dissertação oferece uma análise mais extensa e detalhada dos conceitos e métodos, o artigo apresenta de forma resumida os aspectos mais relevantes da pesquisa. A publicação em uma conferência renomada reconhece o esforço de pesquisa, o artigo é uma parte do estudo mais amplo realizado na dissertação. Este artigo e o *feedback* dos avaliadores contribuíram para a evolução da dissertação, sintetizando as principais seções do trabalho, como metodologia, experimentos, resultados e conclusão.

1.5 Organização do Trabalho

Este trabalho está estruturado em cinco capítulos, cada um abordando diferentes aspectos da pesquisa. No capítulo 1 é apresentado o tema central do estudo, juntamente com o contexto geral que o motiva. Este capítulo também discute a relevância do problema investigado, o estado da arte, os objetivos da pesquisa, e prepara o leitor para os tópicos abordados nos capítulos subsequentes.

No capítulo 2 é apresentada a fundamentação teórica que permite compreender os conceitos relacionados ao uso de LLMs em Q&A (Perguntas e Respostas). Em seguida, no capítulo 3 analisamos os trabalhos que exploram aplicações similares na aplicação de LLMs, nos aprofundando no contexto jurídico, permitindo identificar as lacunas existentes e destacar as contribuições originais desta pesquisa.

O capítulo 4 detalha os métodos empregados para a realização do estudo. Ele cobre desde a coleta e preparação dos dados até a configuração experimental necessária para a análise, assim como as métricas utilizadas para avaliar o desempenho dos modelos de linguagem. Este capítulo é essencial para que o leitor compreenda os procedimentos adotados para alcançar os resultados da pesquisa.

No capítulo 5 são apresentados os resultados obtidos a partir da avaliação dos LLMs. Neste capítulo, são discutidas as capacidades e limitações dos modelos, assim como uma análise detalhada do desempenho em diferentes métricas e a correlação entre elas, no contexto de textos legais.

Por fim, o capítulo 6 encerra o trabalho com uma análise das principais contribuições da pesquisa. Este capítulo também discute as implicações dos resultados, aponta as limitações do estudo e sugere direções para futuras pesquisas na área.

CAPÍTULO 2

Fundamentação Teórica

O Processamento de Linguagem Natural (PLN), um campo que integra Inteligência Artificial (IA) e aprendizado de máquina, tem objetivo de habilitar computadores a interpretar e manipular a linguagem humana de maneira semelhante aos seres humanos [28]. Através de técnicas que evoluíram de regras rígidas para métodos estatísticos e, mais recentemente, para redes neurais profundas como o modelo *Transformer*, o PLN permitiu o desenvolvimento de Grande Modelos de Linguagem de LLMs, como o GPT-3, que, com bilhões de parâmetros, conseguem capturar nuances semânticas e sintáticas em corpora massivos de dados [29, 30]. Esses modelos viabilizam tarefas avançadas de tradução automática, análise de sentimentos, geração de resumos e interações via *chatbots*, impactando positivamente diversos setores ao economizar tempo e recursos [31].

Este capítulo apresenta os principais conceitos e tecnologias necessários para compreender a aplicação de LLMs no contexto do trabalho, em direito tributário. Discutimos uma breve introdução ao direito tributário e ao raciocínio jurídico, passando pelos conceitos de LLMs, técnicas de criação de *prompt* e geração aumentada por recuperação (RAG), até chegar aos métodos de avaliação desses modelos.

2.1 Direito Tributário

O direito tributário, como ramo específico do direito público, é responsável por regulamentar as relações entre o Estado e os contribuintes, especialmente no que tange à arrecadação de tributos. Sua importância se estende à manutenção das funções públicas e ao controle das atividades econômicas. Neste contexto, o raciocínio jurídico assume papel central, pois é a capacidade de interpretar e aplicar normas jurídicas de maneira adequada que permite aos operadores do direito garantir que as obrigações tributárias sejam cumpridas de forma justa e eficaz. Essa habilidade de raciocínio se revela ainda mais crítica no campo do direito tributário, onde as normas frequentemente apresentam complexidade e ambiguidade [4].

O raciocínio jurídico é uma competência essencial para todos os profissionais do direito, uma vez que o processo de interpretação das normas jurídicas demanda

uma análise profunda das leis, precedentes e princípios que orientam o ordenamento jurídico. No contexto do direito tributário, essa habilidade é especialmente relevante, dado o impacto econômico das decisões jurídicas tanto sobre o poder público quanto sobre as empresas e os cidadãos. Através de um raciocínio jurídico bem fundamentado, o profissional do direito é capaz de interpretar adequadamente a legislação tributária, identificar lacunas normativas e propor soluções que sejam justas e de acordo com a Constituição e as demais normas infraconstitucionais [32].

Além disso, o raciocínio jurídico é indispensável na resolução de controvérsias judiciais e administrativas. Por meio dele, advogados, juízes e outros operadores do direito podem assegurar que as leis sejam aplicadas corretamente, evitando injustiças e garantindo a segurança jurídica nas relações entre o Estado e os contribuintes [32].

O direito tributário é notoriamente conhecido por sua complexidade. Vários fatores contribuem para essa característica, sendo fundamental que os profissionais que atuam nessa área dominem não apenas a técnica jurídica, mas também compreendam os contextos econômicos e sociais nos quais as normas estão inseridas [4].

2.2 Grandes Modelos de Linguagem

Nos últimos anos, os LLMs emergiram como uma das principais inovações no campo da inteligência artificial, revolucionando diversas áreas do conhecimento [30]. Estes modelos são baseados na arquitetura *Transformer*, que continua evoluindo e dominando o campo do PLN, oferecendo soluções para lidar com tarefas complexas que exigem compreensão contextual [29].

O aumento do tamanho dos modelos e seu grande número de parâmetros trouxeram habilidades emergentes, que são capacidades que só aparecem em escalas maiores [33]. O GPT-3, por exemplo, com seus 175 bilhões de parâmetros, pode gerar texto, traduzir idiomas e codificar, aproveitando padrões aprendidos durante o pré-treinamento [30]. Da mesma forma, o desempenho do *PaLM* em diversos *benchmarks* demonstra sua capacidade de raciocínio passo a passo [34].

Estas capacidades emergentes permitem que os LLMs realizem uma ampla gama de tarefas com pouco ou nenhum ajuste adicional, exemplificado pelo conceito de *few-shot learners* [30]. O aprendizado no contexto e o seguimento de instruções são habilidades emergentes, permitindo que o modelo adapte suas respostas com base no contexto fornecido e execute instruções complexas transmitidas em linguagem natural.

Explorações em aprendizado por reforço a partir de *feedback* humano (RLHF) têm mostrado resultados para refinar os comportamentos dos LLMs [34]. O desenvolvimento desses modelos segue diretrizes teóricas baseadas nas leis de *Kaplan* [31] e *Chin-*

chilla [35], que estabelecem relações entre o número de parâmetros, quantidade de dados de treinamento e recursos computacionais necessários.

Avanços recentes na arquitetura incluem o *Grouped Query Attention* (GQA) [36], que reduz o custo computacional do mecanismo de atenção, e o *SwigLU* [37], uma função de ativação que melhora o fluxo de informações entre os neurônios do modelo.

Os LLMs são aplicados em diversas áreas, desde geração de texto e tradução automática até análise de documentos e tomada de decisão assistida [38]. Entretanto, os LLMs enfrentam desafios, como a alta demanda por recursos computacionais [39], a presença de vieses nos dados [40], questões éticas relacionadas à privacidade e consentimento [41], e problemas de factualidade que podem resultar informações incorretas, frequentemente se usa o termo alucinações para esse comportamento. [42].

2.3 Os *Prompts*

A criação e utilização de *prompt* é o processo de otimização das instruções fornecidas aos LLMs para obter respostas com qualidade [28]. Esta área é importante pois os modelos são altamente sensíveis às instruções recebidas, onde pequenas variações na formulação podem resultar em respostas significativamente diferentes [33].

Diversas técnicas foram desenvolvidas para otimizar o desempenho dos LLMs através dos *prompts*. O *zero-shot prompting* apresenta instruções sem exemplos prévios [30], enquanto o *few-shot prompting* fornece alguns exemplos para orientar o modelo [30]. A técnica *Chain of Thought* (CoT) incentiva o modelo a desenvolver um raciocínio sequencial [33], e o *ReAct* (*Reasoning and Acting*) combina raciocínio com ações [43].

O *prompt-based fine-tuning* incorpora *prompts* durante o ajuste do modelo [44], enquanto o *instruction prompting* utiliza comandos diretos e explícitos [45]. A técnica de *self-consistency* busca aumentar a confiabilidade gerando e avaliando múltiplas respostas para o mesmo *prompt* [46].

Em domínios específicos e altamente especializados, a criação de *prompt* exige abordagens particulares para lidar com a complexidade [47]. Técnicas como contextualização nos *prompts* são empregadas para aumentar a acurácia das respostas [47, 48].

2.4 Geração Aumentada por Recuperação (RAG)

A Geração Aumentada por Recuperação (*Retrieval-Augmented Generation*, RAG) é uma tecnologia inovadora que surgiu para melhorar a acurácia e confiabilidade dos Modelos de Linguagem Grandes (LLMs). Esta tecnologia funciona combinando a capacidade dos modelos de gerar texto com informações que são buscadas em tempo real em bases de dados externas [49]. Imagine como se fosse um assistente virtual que, além

de usar seu conhecimento interno, também consulta uma biblioteca digital para garantir que suas respostas sejam mais precisas e atualizadas.

O funcionamento do RAG é baseado em dois componentes principais que trabalham juntos: o *retriever* (recuperador) e o *generator* (gerador) [49,50]. O recuperador é como um bibliotecário eficiente que busca documentos relevantes em uma base de conhecimento quando recebe uma pergunta. Ele usa técnicas avançadas chamadas *dense retrieval* e modelos de *embedding* para encontrar documentos que sejam semanticamente similares à pergunta feita [51]. Já o gerador, que é construído usando tecnologias como T5 ou GPT, funciona como um escritor especialista que lê tanto a pergunta original quanto os documentos encontrados pelo recuperador, criando uma resposta que incorpora todas essas informações [52].

Uma das grandes vantagens do RAG é que ele permite que os modelos tenham acesso a informações sempre atualizadas sem precisar passar por um processo completo de retreinamento [49]. Além disso, como sabemos exatamente quais documentos foram usados para gerar cada resposta, é possível verificar as fontes das informações, tornando o sistema mais transparente e confiável [53]. Outra vantagem importante é que podemos adaptar o sistema para áreas específicas simplesmente mudando a base de conhecimento que ele consulta [54].

Para avaliar o desempenho de sistemas RAG, os pesquisadores precisam considerar diversos aspectos. Eles medem a acurácia da recuperação usando métricas como *Precision@k*, *Recall@k* e *Mean Reciprocal Rank (MRR)* [55]. A qualidade do texto gerado é avaliada através de métricas como ROUGE, BLEU, *BertScore*, além de usar LLMs e avaliações feitas por humanos [56]. Também é importante avaliar a eficiência computacional, observando aspectos como velocidade de resposta, uso de memória e necessidade de processamento [57].

Pesquisadores têm desenvolvido diferentes versões aprimoradas do RAG original. Por exemplo, existe o RAG Adaptativo [58], que ajusta automaticamente a quantidade de informação recuperada dependendo da complexidade da pergunta. O RAG *Multi-Vector* [59] usa diferentes formas de representar cada documento para encontrar informações mais precisas e relevantes. Já o RAG Hierárquico [60] organiza a busca em camadas, começando com informações mais gerais e depois refinando para detalhes mais específicos.

O futuro da pesquisa em RAG está focado em desenvolver formas mais eficientes de organizar e recuperar informações, melhorar como as informações são selecionadas e combinadas, e reduzir o uso de recursos computacionais sem prejudicar a qualidade das respostas. Todo esse desenvolvimento contínuo faz do RAG uma tecnologia muito importante para criar sistemas que precisam trabalhar com conhecimento explícito e verificável [61].

2.5 Desenvolvimento de *Datasets*

Os conjuntos de dados (*datasets*) para PLN são coleções estruturadas de informações linguísticas essenciais para o treinamento e avaliação de modelos [62]. Sendo fundamentais para desenvolver modelos capazes de realizar tarefas como extração de informações, resposta a perguntas e classificação de textos [63].

A definição clara da finalidade é fundamental no desenvolvimento de *datasets*, se é relacionado ao RAG, por exemplo, podem incluir tanto pares de perguntas-respostas quanto corpus textuais para recuperação. Os *datasets* podem ser projetados para diferentes objetivos, como avaliação de raciocínio, compreensão de texto ou geração de linguagem natural [64].

Diferentes tipos de *datasets* são projetados para atender a objetivos específicos. Conjuntos tradicionais como SQuAD [65] e TriviaQA [66] focam em perguntas factuais, enquanto outros *datasets* podem ser desenvolvidos para tarefas mais complexas como raciocínio *multi-hop* ou inferência textual [67]. O *HotpotQA* [68], por exemplo, foi criado especificamente para avaliar o raciocínio *multi-hop*, onde a resposta requer a combinação de informações de múltiplas fontes.

A qualidade dos *datasets* é assegurada através de um rigoroso processo de anotação por especialistas, que identificam e validam as respostas corretas e as passagens relevantes [69]. Este processo pode envolver múltiplas etapas de revisão e validação para garantir a consistência e precisão das anotações.

A estrutura dos *datasets* de Perguntas e Respostas (Q&A) é tipicamente organizada em pares de pergunta-resposta, acompanhados por passagens de referência (também chamadas de passagens de ouro) que fundamentam as respostas [70]. Além disso, podem incluir metadados adicionais como tipo de pergunta, dificuldade, ou categorias temáticas.

Um aspecto do desenvolvimento de *datasets* é a consideração de diferentes aspectos linguísticos e cognitivos. Isso inclui variação de complexidade sintática, diversidade lexical, diferentes tipos de raciocínio necessários e níveis de abstração [64]. Também é importante considerar a distribuição balanceada de diferentes tipos de perguntas e temas.

A avaliação da qualidade do *dataset* pode envolver aspectos como concordância entre anotadores (*inter-annotator agreement*), cobertura temática, diversidade linguística e ausência de vieses [62]. Considerações éticas e de viés são incorporadas ao desenvolvimento dos *datasets*, visando criar recursos mais representativos e equitativos.

Aspectos práticos como tamanho do *dataset*, formato de armazenamento, documentação clara e licenciamento também são fundamentais. *Datasets* modernos frequentemente seguem princípios *FAIR* (*Findable*, *Accessible*, *Interoperable*, *and Reusable*) para facilitar seu uso pela comunidade científica [71].

2.6 Avaliação de LLMs

A avaliação de LLMs abrange tanto métricas quantitativas quanto qualitativas para mensurar seu desempenho em diversas tarefas [72]. Estas abordagens complementares permitem uma análise abrangente da eficácia dos modelos em tarefas como tradução, resumo e perguntas-respostas [34]. A seguir apresentamos os conceitos e limitações das métricas utilizadas nesse trabalho.

Hipótese:

O fato gerador do imposto ocorre mediante operação financeira realizada

Referência:

O fato gerador do tributo ocorre quando operação financeira realizada

1-grama:

O | fato | gerador | do | ocorre | operação | financeira | realizada

2-grama:

O fato | fato gerador | gerador do | ocorre | operação financeira | financeira realizada

3-grama:

O fato gerador | fato gerador do | operação financeira realizada

4-grama:

O fato gerador do

Figura 2.1: *Análise da correspondência de n-gramas e o número de ocorrências no BLEU.*

O BLEU (*Bilingual Evaluation Understudy*) é uma métrica que avalia a qualidade de um texto traduzido (ou gerado) comparando-o com uma ou mais referências. A métrica é expressa através da seguinte equação:

$$BLEU_w(\hat{S};S) := BP(\hat{S};S) \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n(\hat{S};S)\right) \tag{2-1}$$

Na equação (2-1), $\hat{S}$ representa a sentença hipótese (candidata), S é a sentença de referência, N é o n-grama máximo considerado e w_n são os pesos para cada n-grama, que tradicionalmente são uniformes com valor $w_n = \frac{1}{N}$.

O primeiro componente da equação, o BP (*Brevity Penalty*), é um fator de penalização que evita que o sistema gere traduções muito curtas. Ele é calculado da seguinte forma:

$$BP(\hat{S};S) = \begin{cases} 1 & \text{se } c > r \\ e^{1-\frac{r}{c}} & \text{se } c \leq r \end{cases} \tag{2-2}$$

Nesta equação (2-2), c representa o comprimento da hipótese e r o comprimento da referência. Quando a hipótese é mais longa que a referência ($c > r$), não há penalização (BP = 1). No entanto, quando a hipótese é mais curta ($c < r$), o sistema é penalizado proporcionalmente à diferença de comprimento.

O segundo componente principal é a precisão modificada (p_n) (2-3), calculada para cada ordem de n-grama:

$$p_n(\hat{S};S) = \frac{\sum_{ngram \in \hat{S}} \min(h_{ngram}, r_{ngram})}{\sum_{ngram \in \hat{S}} h_{ngram}} \tag{2-3}$$

Nesta equação, h_{ngram} representa o número de ocorrências do n-grama na hipótese e r_{ngram} o número de ocorrências na referência. Esta precisão modificada é calculada para diferentes ordens de n-gramas, cada uma capturando diferentes aspectos da qualidade da tradução.

Os diferentes níveis de n-gramas contribuem de maneiras distintas para a avaliação final. A precisão de 1-grama (p_1) mede principalmente a adequação lexical, verificando se as palavras corretas foram escolhidas. A precisão de bigrama (p_2) captura aspectos de ordem local e colocações de palavras. A precisão de trigrama (p_3) começa a avaliar estruturas maiores da língua, enquanto a precisão de 4-grama (p_4) captura aspectos de fluência em nível de frase.

Os pesos (w_n) na equação permitem ajustar a importância relativa de cada nível de n-grama. Na prática, é comum usar pesos uniformes ($w_1 = w_2 = w_3 = w_4 = \frac{1}{4}$), mas eles podem ser ajustados. Pesos maiores para n-gramas mais longos tendem a favorecer a fluência do texto, enquanto pesos maiores para n-gramas mais curtos favorecem a adequação lexical.

Para um cálculo típico do BLEU até 4-gramas, a equação (2-4) é a seguinte:

$$BLEU = BP \cdot \exp(\frac{1}{4}(\ln p_1 + \ln p_2 + \ln p_3 + \ln p_4)) \tag{2-4}$$

O score BLEU final é um valor entre 0 e 1, onde 1 indica uma correspondência perfeita e 0 indica nenhuma correspondência de n-gramas. Na prática, valores acima de 0.5 são considerados muito bons, embora isso possa variar dependendo do contexto e da aplicação específica.

O BLEU apresenta limitações significativas quando aplicado à avaliação de respostas em sistemas de Q&A e modelos generativos, principalmente devido à sua

rigidez na correspondência exata de n-gramas, não capturando adequadamente variações sinônimas ou respostas semanticamente equivalentes [10].

Outra métrica usada na pesquisa, o ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [11] mostra-se mais adequada para tarefas de Q&A por focar no *recall* além da precisão, utilizar *Longest Common Subsequence* em variantes como ROUGE-L, e ter sido projetado especificamente para avaliar sumários e respostas onde a ordem exata das palavras é menos crítica [11].

O ROUGE é uma métrica originalmente orientada à avaliação de tarefas de sumarização [11], mas tem sido empregada para avaliar geração de texto de LLMs [73]. Semelhante ao BLEU, o ROUGE compara as saídas do modelo com uma referência, mas com maior ênfase em termos de *recall*, ou seja, em quantos dos elementos importantes da resposta de referência foram capturados pelo modelo.

A variante ROUGE-L é utilizada para avaliar a qualidade de textos gerados ao medir a sobreposição baseada na Subsequência Comum Mais Longa (*Longest Common Subsequence* - LCS) entre um texto candidato e o texto de referência. A LCS captura a informação de sequências em ordem, sem exigir que elas sejam contíguas, o que avalia a similaridade em textos gerados.

Primeiramente, calcula-se o comprimento da LCS entre as sequências X (referência) e Y (candidato), denotado por $\text{LCS}(X,Y)$. Em seguida, determinam-se o *Recall* e a Precisão:

$$Recall = \frac{\text{LCS}(X,Y)}{\text{len}(X)} \tag{2-5}$$

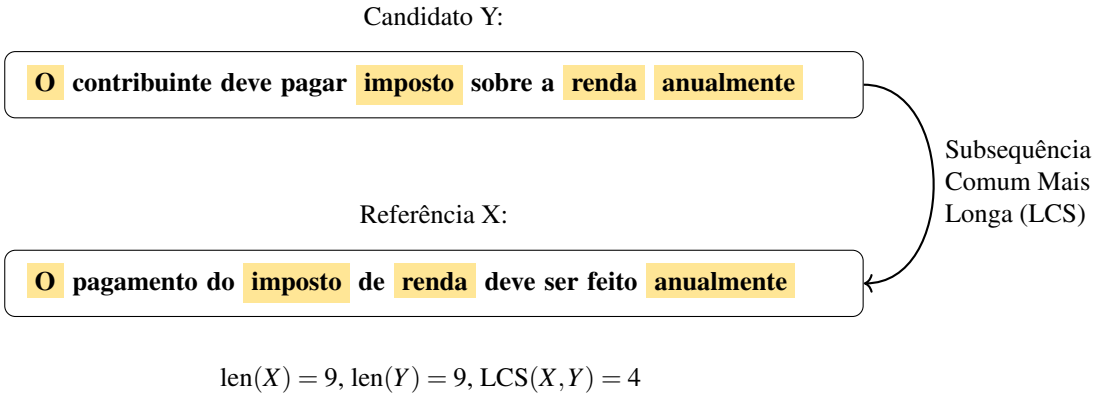
$$\text{Precisão} = \frac{\text{LCS}(X,Y)}{\text{len}(Y)} \tag{2-6}$$

onde $\text{len}(X)$ e $\text{len}(Y)$ representam os comprimentos das sequências X e Y, respectivamente. Por fim, o *F-Score* do ROUGE-L é calculado combinando o *Recall* e a Precisão:

$$F\text{-score} = \frac{(1+\beta^2) \times Recall \times \text{Precisão}}{\beta^2 \times \text{Precisão} + Recall} \tag{2-7}$$

onde β é um parâmetro que determina a importância relativa do *Recall* em relação à Precisão (comumente $\beta = 1$). Essa métrica permite uma avaliação mais flexível e informativa da qualidade dos textos gerados, ao considerar a ordem e a sequência das palavras, a Figura 2.2, ilustra como obter os parâmetros para o cálculo do ROUGE-L.

Outra métrica utilizada no trabalho é o *BertScore* é uma métrica mais recente e sofisticada, devido à sua capacidade de avaliação semântica entre textos. Segundo os autores [12] ela supera o *ROUGE-L* e o *BLEU* e tem um alinhamento com a avaliação



$$\text{len}(X) = 9, \text{len}(Y) = 9, \text{LCS}(X,Y) = 4$$

Figura 2.2: *Exemplo de parâmetros do ROUGE-L.*

humana. Esta métrica, utiliza *embeddings* de modelos como o BERT para calcular a similaridade semântica entre a resposta gerada e a de referência [12]. Em vez de comparar apenas palavras ou n-gramas, o *BertScore* avalia a proximidade das representações semânticas das frases, tornando-o uma métrica melhor para textos complexos,onde a correspondência exata de palavras não é tão importante quanto a coerência e a precisão semântica [74].

Embora as métricas quantitativas sejam úteis, muitas vezes elas não conseguem capturar completamente a qualidade e a adequação das respostas em tarefas complexas, como o raciocínio jurídico [75]. Para complementar essas métricas, a avaliação qualitativa tornou-se cada vez mais comum, especialmente com o uso de LLMs avançados [16–19]. Modelos avançados, como o GPT-4, têm sido usados para revisar e avaliar a qualidade das respostas geradas por outros modelos ou até mesmo por eles mesmos. Esses LLMs podem julgar a coerência, acurácia e relevância de uma resposta com base em seu próprio entendimento profundo da linguagem e do contexto, fornecendo informações sobre aspectos que as métricas quantitativas não capturam [76].

Ao utilizar LLMs para avaliação, é essencial considerar fatores como a acurácia factual e a relevância da resposta [77]. A acurácia factual é especialmente importante em domínios como o direito, onde informações incorretas podem ter consequências graves. A relevância, por outro lado, refere-se à capacidade do modelo de gerar uma resposta que esteja de acordo com o contexto e a pergunta original, evitando desvios ou respostas superficiais [74].

Limitações das métricas existentes

Embora as métricas quantitativas e as avaliações qualitativas ofereçam um panorama útil do desempenho dos LLMs, ambas apresentam limitações, especialmente em contextos que exigem raciocínio mais profundo, como o direito [78].

As métricas quantitativas, como BLEU e ROUGE, frequentemente falham em capturar as nuances do raciocínio jurídico, que envolve a interpretação de múltiplas fontes, a aplicação de princípios legais e a resolução de ambiguidades [74]. Como essas métricas se concentram principalmente na similaridade superficial entre textos, elas não conseguem avaliar adequadamente o processo de raciocínio por trás da resposta [72].

Tarefas complexas, como o raciocínio jurídico, exigem uma abordagem mais contextual e humanizada na avaliação [75]. Isso ocorre porque os LLMs podem gerar respostas linguisticamente corretas, mas que carecem de precisão ou de um entendimento profundo do contexto. Avaliações por humanos, especialmente especialistas de domínio, continuam sendo necessárias para garantir que as respostas sejam não apenas linguística e semanticamente corretas, mas também factualmente precisas e adequadas [79].

CAPÍTULO 3

Trabalhos correlatos

O PLN tem revolucionado o campo jurídico através de soluções inovadoras para análise, interpretação e classificação de dados legais [20]. Sendo possível realizar diversas tarefas no direito, processando textos jurídicos complexos com eficiência superior à capacidade humana [5]. Suas principais aplicações incluem análise automatizada de documentos [80], sistemas de recuperação de informação [25], análise preditiva [81], sistemas de perguntas e respostas (Q&A) e automação na redação de documentos legais [82], onde os algoritmos podem extrair informações-chave, identificar padrões e classificar documentos com base em seu conteúdo semântico [6]. No entanto, o PLN ainda enfrenta desafios significativos no contexto jurídico, como a complexidade linguística, ambiguidade semântica e a necessidade de conhecimento especializado [21, 24], sendo que a relação entre diferentes textos e seus contextos legais é crucial para a interpretação correta dos documentos [83]. Estas tecnologias servem como ferramentas de apoio, não substituindo o julgamento jurídico humano [23].

Este capítulo apresenta uma revisão abrangente dos trabalhos relacionados à aplicação de LLMs em sistemas de Q&A, com ênfase especial no domínio jurídico e no contexto do direito tributário. A análise está estruturada em três vertentes principais: primeiramente, discutimos os avanços gerais em sistemas de Q&A com LLMs, abordando aspectos metodológicos e métricas de avaliação; em seguida, exploramos aplicações específicas no campo jurídico, considerando diferentes abordagens e estratégias. Por fim, examinamos estudos focados no direito tributário em contextos internacionais e apresentamos as considerações finais. Esta organização permite uma compreensão progressiva do estado da arte, partindo de conceitos gerais até as aplicações mais específicas relacionadas ao escopo deste trabalho.

3.1 Perguntas e Respostas com LLMs

Os sistemas de Perguntas e Respostas (Q&A) com LLMs têm demonstrado avanços significativos na capacidade de processar e responder consultas complexas. Hicke et al. [84] investigou a capacidade dos LLMs em responder perguntas que exigem raciocínio

estratégico sobre múltiplos documentos, utilizando conjuntos de dados como StrategyQA e HotpotQA-Complex [68]. Este tipo de tarefa requer que os modelos combinem informações de várias fontes, demonstrando habilidades avançadas de raciocínio. Um aspecto crucial no desenvolvimento destes sistemas é a metodologia de avaliação.

Monteiro et al. [85] propôs uma abordagem similar a aplicada em nosso trabalho que fornece o documento de referência junto com a pergunta, permitindo avaliar a real capacidade do modelo em extrair e compreender informações do contexto, em vez de simplesmente recuperar conhecimento memorizado durante o treinamento. Esta metodologia revelou que modelos maiores nem sempre apresentam melhor desempenho quando precisam extrair informações de documentos fornecidos. Na avaliação da qualidade das respostas, diversas métricas são empregadas. As mais amplamente utilizadas são BLEU, ROUGE e BERTScore [86–88]. No entanto, Alinejad et al. [89] faz uso de um LLM como avaliador em *GPT4-Eval*, o estudo demonstrou alta correlação com julgamentos humanos ao retornar respostas binárias ("*Yes*"ou "*No*") para indicar equivalência entre respostas. Esta abordagem superou as limitações das métricas automáticas tradicionais, especialmente em casos onde as respostas são semanticamente equivalentes mas lexicalmente diferentes. O cálculo da acurácia é feito de maneira semelhante em nosso trabalho

O uso de LLMs como avaliadores tem se mostrado promissor. Srivastava et al. [90] demonstrou que modelos robustos como o GPT-4 podem alcançar mais de 80% de concordância com avaliações humanas em aspectos como relevância, coerência e fluência das respostas. No entanto, os autores reconhecem limitações como vieses posicionais e de verbosidade. Pesquisas específicas em domínios especializados também têm sido conduzidas. Subramanian et al. [91] desenvolveu um *benchmark* para avaliar a compreensão e recuperação de conhecimento médico, enquanto [92] focou em questões sobre artigos científicos da área de computação.

Estes trabalhos demonstram a versatilidade dos LLMs em diferentes contextos, mas também destacam os desafios específicos de cada domínio. Avanços também foram observados em abordagens que combinam diferentes técnicas. Kim et al. [93] propôs um método que utiliza LLMs tanto para aumentar as questões quanto para gerar passagens complementares, demonstrando a importância da combinação entre conhecimento paramétrico e recuperação de informações externas.

Em relação às métricas de avaliação utilizadas nos trabalhos correlatos no contexto de Q&A em diversos domínios, observa-se uma predominância de métricas tradicionais baseadas em sobreposição de texto. A métrica ROUGE foi a mais amplamente adotada, sendo utilizada em 11 dos 15 trabalhos analisados, seguida pelo BERTScore, presente em 9 trabalhos, e BLEU, empregada em 7 estudos. Pode-se notar a emergência do uso de LLMs como avaliadores, uma abordagem mais recente que já foi adotada em 6 trabalhos, demonstrando uma tendência crescente na área. Esta distribuição sugere que,

Tabela 3.1: *Métricas de avaliação utilizadas nos trabalhos correlatos no contexto de Q&A em diversos domínios*

Trabalho	BLEU	ROUGE	BERTScore	LLM Avaliador
[84]				✓
[86]	✓	✓	✓	✓
[87]	✓	✓	✓	
[88]		✓		✓
[89]	✓	✓	✓	✓
[94]				✓
[91]		✓	✓	
[95]	✓	✓	✓	✓
[93]		✓		
[96]	✓	✓	✓	
[85]		✓	✓	
[92]	✓	✓	✓	
[97]		✓	✓	
[98]	✓	✓	✓	
[90]				✓
Total	7	11	9	6

embora as métricas tradicionais ainda sejam predominantes, há um movimento em direção a métodos de avaliação mais sofisticados, que podem capturar aspectos semânticos e nuances linguísticas que as métricas baseadas em sobreposição de texto não conseguem identificar adequadamente.

Além disso, pesquisas têm explorado diferentes aspectos e aplicações de Q&A. [94] focou na avaliação de respostas longas (LFQA) em diversos domínios, enquanto [96] investigou o uso de *reinforcement learning* para melhorar o desempenho em questões sobre programação, [97] demonstrou a aplicabilidade desses sistemas em diferentes idiomas, desenvolvendo um sistema de Q&A em tailandês e [98] inovou ao utilizar Q&A com LLMs como método de verificação de qualidade para avaliação de sumários, demonstrando a versatilidade dessas técnicas.

3.2 LLMs para Q&A no contexto jurídico

A aplicação de LLMs no direito tem sido objeto de diversos estudos recentes, que abrangem desde o desenvolvimento de conjuntos de dados (*datasets*) e métricas de avaliação (*benchmarks*) especializados até a criação de sistemas inteligentes para tarefas jurídicas. Estes trabalhos visam aprimorar a capacidade dos LLMs em compreender textos legais, gerar respostas fundamentadas e incorporar conhecimento jurídico específico.

O raciocínio jurídico demanda compreensão profunda dos princípios legais, envolvendo interpretação de conceitos abstratos e consideração de princípios como razo-

abilidade e proporcionalidade [5]. Para maximizar o potencial dos LLMs no ambiente legal, é necessário desenvolver modelos que combinem eficientemente habilidades de recuperação de informações com raciocínio contextual avançado, adaptando-se às nuances e ambiguidades inerentes ao direito [74].

Embora existam numerosos trabalhos utilizando LLMs no domínio jurídico [99–103], nosso interesse está naqueles que aplicam LLMs para tarefas de perguntas e respostas (*Q&A*). Esses trabalhos podem ser classificados em três categorias principais: aqueles que usam geração aumentada por recuperação, aqueles que avaliam LLMs com base em conhecimento prévio, e aqueles que realizam *fine-tuning* e testam os modelos em tarefas de Q&A no domínio jurídico. Abaixo, discutimos os principais trabalhos encontrados em cada uma dessas categorias.

3.2.1 Avaliando Q&A LLMs com Geração Aumentada por Recuperação

No contexto jurídico, o RAG permite que o modelo busque e utilize trechos específicos de leis, regulamentos ou precedentes judiciais para formular respostas fundamentadas [7]. A eficácia dessa abordagem é avaliada considerando tanto a qualidade da recuperação quanto a acurácia da geração, frequentemente utilizando passagens de referência (*gold passages*) como *benchmarks* [104].

Aproveitando as técnicas de RAG e *prompt*, os LLMs têm demonstrado potencial significativo em tarefas de Perguntas e Respostas (Q&A) no campo jurídico [6]. Estas tarefas podem ser categorizadas em dois tipos principais: extrativas e de raciocínio jurídico.

Nas tarefas extrativas, as respostas podem ser encontradas diretamente nos textos recuperados, como em consultas sobre prazos processuais, onde o modelo atua principalmente como um sistema de busca avançado [5]. Por outro lado, tarefas que envolvem raciocínio jurídico exigem que o modelo interprete e relacione múltiplas fontes de informação, aplicando conceitos em contextos específicos [105].

Um desafio particular no domínio jurídico é que o conteúdo recuperado, como artigos de lei ou regulamentos, frequentemente requer interpretação contextual para gerar respostas adequadas [5, 74]. Isso exige que o modelo não apenas recupere e reproduza informações, mas também demonstre compreensão profunda dos princípios legais para gerar respostas juridicamente válidas [25, 105].

No estudo *Long-form Legal Question Answering* (LLeQA) [26], a técnica de RAG é aplicada ao contexto jurídico. Utilizando a abordagem denominada pelo autor como "*retrieve-then-read*", o modelo recupera artigos de um extenso corpus de legislação belga e gera respostas detalhadas com base nas informações recuperadas. O *dataset*

LLeQA contém 1.868 perguntas legais anotadas por especialistas, juntamente com suas respostas e referências legais, demonstrando a eficácia da integração de recuperação e geração de respostas para questões jurídicas complexas. A construção do *dataset* envolveu a coleta de perguntas reais e a anotação por juristas experientes, o que assegura a precisão das informações fornecidas.

Cui et al. [99] apresentam o modelo *ChatLaw*, desenvolvido para o domínio jurídico chinês. O modelo combina recuperação de informações baseada em vetores e em palavras-chave, para diminuir os problemas de alucinação (respostas incoerentes) típicos em LLMs. O *ChatLaw* utiliza uma estrutura de quatro módulos: "consulta", "referência", "auto-sugestão"e "resposta", que integra conhecimento jurídico específico no modelo durante a inferência. Além disso, o artigo descreve a criação de um *dataset* extenso de dados legais, regulamentações e consultas jurídicas reais, que foi rigorosamente processado para garantir a qualidade das respostas geradas.

3.2.2 Avaliando o Raciocínio Jurídico dos LLMs em Q&A

Fei et al. [73] propõem o *LawBench*, um *benchmark* projetado para avaliar as capacidades dos LLMs em tarefas jurídicas no sistema legal chinês. O *LawBench* divide a avaliação dos LLMs em três dimensões cognitivas: memorização, compreensão e aplicação do conhecimento jurídico. No estudo, 51 LLMs, incluindo multilíngues e especializados no direito, foram avaliados, e o GPT-4 obteve o melhor desempenho. O trabalho destaca que, embora o *fine-tuning* em textos jurídicos melhore o desempenho, os LLMs ainda enfrentam desafios para serem completamente confiáveis em tarefas jurídicas complexas.

Dai et al. [106] introduzem o *LAiW*, um *benchmark* para a avaliação de LLMs no contexto jurídico chinês. O *LAiW* avalia as capacidades dos modelos em três níveis: PLN básico, aplicação legal simples e aplicação legal complexa. O artigo também descreve o processo de construção do *benchmark* e sua categorização em tarefas específicas, com uma avaliação inicial que mostra que modelos ajustados para o direito superam versões gerais, embora ainda fiquem atrás de modelos como o GPT-4.

3.2.3 *Fine-tuning* e Avaliação de LLMs para Q&A

Yue et al. [107] apresentam o *FedJudge*, uma estrutura baseada em Aprendizado Federado (FL), que preserva a privacidade dos dados ao treinar LLMs em clientes locais, com os parâmetros sendo agregados em um servidor central. O *FedJudge* é avaliado em tarefas de perguntas e respostas jurídicas, utilizando métricas como ROUGE, BLEU e *BertScore*. A abordagem demonstrou ser mais eficaz que métodos tradicionais, garantindo

respostas mais precisas e relevantes, além de preservar a privacidade dos dados em diferentes contextos legais.

Por fim, Yue et al. [108] propõem o *DISC-LawLLM*, um sistema inteligente que usa LLMs para fornecer uma variedade de serviços jurídicos. Utilizando estratégias de *prompting* de silogismo jurídico e treinado em dados supervisionados do domínio jurídico chinês, o *DISC-LawLLM* também incorpora um módulo de recuperação de conhecimento jurídico externo. Avaliado por meio do *DISC-Law-Eval*, o sistema demonstrou eficácia em diversas tarefas jurídicas, como extração de elementos jurídicos e predição de julgamentos, sendo avaliado tanto por métricas objetivas quanto subjetivas, incluindo o uso de GPT-3.5 como avaliador.

3.3 LLMs para tarefas jurídicas em contexto de direito tributário no mundo

A aplicação de LLMs no domínio do direito tributário tem se mostrado um campo emergente e promissor de pesquisa. Diversos estudos têm explorado as capacidades destes modelos em tarefas como interpretação legal, Q&A e raciocínio jurídico aplicado à legislação tributária.

Xu et al. [109] apresenta uma avaliação similar a do nosso trabalho de LLMs aplicados a exames de qualificação tributária na China, testando especificamente a capacidade dos modelos de raciocinar sobre provisões legais fornecidas diretamente no *prompt*. O estudo demonstra que os modelos mantêm consistência e acurácia quando o contexto legal necessário é incluído na entrada, permitindo avaliar sua capacidade de interpretação e resposta baseada apenas nas informações fornecidas, sem depender de recuperação externa de dados.

Na mesma linha, [110] explora o uso de LLMs como "advogados tributários virtuais", avaliando sua eficácia em responder questões sobre legislação tributária americana através de diferentes técnicas de *prompting* e métodos de recuperação de contexto legal. Os resultados indicam que os modelos apresentam capacidades emergentes significativas em Q&A jurídico.

Zhang [111] realizou um dos primeiros estudos avaliando o desempenho de LLMs em questões específicas de direito tributário, revelando que os modelos apresentam dificuldades em questões técnicas que exigem interpretação precisa da legislação, embora demonstrem melhor desempenho em questões conceituais sobre pratica tributária.

Contribuições significativas também vêm de [112], que apresenta um LLM especializado em contabilidade e direito tributário, treinado com mais de 15 mil diálogos

reais, sugerindo a viabilidade de especializar modelos para domínios jurídicos específicos, como o tributário.

Weller et al. [113] introduz uma métrica específica (*QUIP-Score*) para avaliar a fidelidade das respostas dos LLMs a fontes confiáveis, incluindo o código tributário, enquanto [114] explora o uso de modelos para previsão de desfechos em disputas fiscais, alcançando alta acurácia em suas análises.

Amrullah et al. [115] fornece uma perspectiva sobre a implementação prática de IA em serviços tributários, destacando aspectos importantes sobre eficiência e acurácia no contexto de sistemas tributários complexos.

3.4 Considerações Finais

A análise dos trabalhos correlatos revela uma evolução significativa na aplicação de LLMs para tarefas jurídicas, especialmente no contexto de perguntas e respostas. Ao comparar estes trabalhos com a presente proposta, surgem pontos de convergência e distinções que são discutidos nesta seção.

No contexto das métricas de avaliação, a presente proposta se alinha com a tendência observada em trabalhos recentes de combinar métricas tradicionais (BLEU, ROUGE, *BERTScore*) com avaliação por LLMs. Esta abordagem híbrida, similar à adotada por [86, 89, 95] e [90] que usou apenas LLMs, demonstra-se particularmente eficaz no contexto jurídico, onde a avaliação puramente automatizada pode não capturar nuances importantes do raciocínio legal. A metodologia inova ao aplicar estas métricas especificamente no contexto do direito tributário brasileiro, um domínio ainda pouco explorado na literatura.

O uso de LLMs como avaliadores, uma tendência crescente observada em [89] e [90], foi incorporado na metodologia através do GPT-4 como substituto do julgamento humano. Esta escolha metodológica encontra respaldo nos resultados de [94], que demonstrou alta correlação entre avaliações realizadas por LLMs e julgamentos humanos em tarefas complexas de análise textual.

Em relação aos trabalhos focados em Q&A jurídico, a abordagem compartilha similaridades metodológicas com [26] no desenvolvimento do LLeQA, principalmente na construção de um *dataset* especializado. No entanto, enquanto o LLeQA focou na legislação belga com 1.868 questões, o presente trabalho se diferencia ao criar um corpus específico do direito tributário brasileiro com questões cuidadosamente selecionadas da Coordenação-Geral de Tributação (Cosit) [1], oferecendo uma contribuição única para este domínio específico.

A metodologia de avaliação dos modelos adotada na presente pesquisa aproxima-se da abordagem de [85, 109], que também testou a capacidade dos LLMs

em raciocinar sobre provisões legais fornecidas diretamente no *prompt*. A contribuição se destaca ao expandir esta análise para um conjunto mais amplo de modelos (mais de 20 LLMs), incluindo tanto proprietário quanto *open-source*, especificamente no contexto tributário brasileiro.

Uma característica distintiva da presente proposta em relação aos trabalhos anteriores é o foco na avaliação em português brasileiro. Enquanto a maioria dos estudos, como [99] com o *ChatLaw* e [73] com o *LawBench*, concentrou-se em sistemas jurídicos em inglês ou chinês, esta pesquisa contribui para preencher uma lacuna importante na literatura sobre LLMs aplicados ao direito em português.

No âmbito específico do direito tributário, a presente proposta se diferencia dos trabalhos de [111] e [110] pela abordagem sistemática na construção do *dataset* e pela avaliação abrangente de múltiplos modelos. Enquanto estes trabalhos focaram principalmente em aspectos conceituais ou em questões técnicas isoladas, esta pesquisa oferece uma análise mais abrangente do desempenho dos LLMs no raciocínio jurídico tributário.

Diferentemente de trabalhos como [107] e [108], não foi realizado *fine-tuning* dos modelos, focando na avaliação de sua capacidade de geração aumentada. Esta escolha metodológica permite uma avaliação mais direta das capacidades dos modelos em seu estado atual.

A análise de correlação entre métricas, utilizando correlações de *Pearson* e *Kendall*, encontra paralelo metodológico em [107], que também empregou análises estatísticas robustas para validar seus resultados. A presente contribuição adiciona a análise de *Bland-Altman*, oferecendo uma perspectiva complementar sobre a concordância entre diferentes métodos de avaliação.

O desempenho superior do *Qwen2-72B-Instruct* e do *Mixtral-8x22B-Instruct-v0.1* nos experimentos dialoga com os resultados obtidos por [108] e [106], que também observaram variações significativas no desempenho entre diferentes arquiteturas de modelos. No entanto, a descoberta sobre o desempenho competitivo de modelos menores bem ajustados adiciona uma perspectiva importante para a otimização de recursos em aplicações práticas.

As limitações identificadas no presente trabalho, como a ausência de avaliação direta por especialistas e a não exploração completa do RAG, dialogam com desafios similares relatados por [115] e [114]. No entanto, estas limitações também apontam direções promissoras para pesquisas futuras, especialmente considerando o potencial de integração com técnicas avançadas de recuperação de informação.

Em relação à dimensão temporal do *dataset*, uma limitação também observada em [5] e [74], a abordagem se diferencia ao focar em um corpus legislativo relativamente atualizado e cuidadosamente curado. Esta escolha metodológica, embora limitando a

escala do conjunto de dados, garante maior acurácia e relevância das análises no contexto atual do direito tributário brasileiro.

A decisão de não explorar cenários mais complexos ou casos hipotéticos, embora represente uma limitação, alinha-se com a abordagem metodológica de [84] e [92], que privilegiaram a profundidade da análise em um escopo bem definido. Esta escolha permitiu uma avaliação mais rigorosa e controlada do desempenho dos modelos nas tarefas específicas propostas.

A revisão da literatura evidenciou uma lacuna significativa nos estudos sobre a aplicação de LLMs no contexto do direito tributário brasileiro, tanto em termos de avaliação sistemática de desempenho quanto no uso do português jurídico. Enquanto diversos trabalhos correlatos, focam na aplicação de LLMs em áreas jurídicas gerais ou em outros sistemas legais, identificamos a necessidade de uma investigação específica no domínio tributário nacional. Esta lacuna é especialmente relevante considerando as particularidades do sistema tributário brasileiro e seus desafios linguísticos únicos, aspectos ainda não abordados adequadamente na literatura existente. Os trabalhos correlatos também apontam para oportunidades promissoras na integração de técnicas de RAG, na análise jurisprudencial e no desenvolvimento de metodologias de avaliação específicas para o contexto jurídico em português, direções estas que fundamentam os objetivos desta pesquisa.

CAPÍTULO 4

Metodologia

Neste capítulo, detalhamos a metodologia empregada para atingir os objetivos desta pesquisa, que busca avaliar a eficácia de Modelos de Linguagem de Grande Escala (LLMs) no raciocínio jurídico aplicado ao direito tributário para pessoas jurídicas em português brasileiro. A metodologia abrange todo o processo de seleção dos modelos, coleta e preparação dos dados, criação de um *dataset* relevante e configuração dos experimentos, incluindo os parâmetros e *prompts* específicos utilizados nos LLMs escolhidos.

Além disso, descrevemos a configuração experimental adotada, as ferramentas utilizadas para viabilizar o estudo e as métricas implementadas na avaliação quantitativa e qualitativa do desempenho dos modelos. Nosso enfoque também inclui a estratégia de avaliação qualitativa, que complementa as métricas tradicionais, garantindo uma análise mais profunda da acurácia e relevância das respostas geradas. O objetivo é fornecer uma descrição detalhada o suficiente para que outros pesquisadores possam replicar e validar este estudo.

4.1 Coleta e Preparação de Dados

A base de dados utilizada nesta pesquisa consiste em uma série de perguntas reais relacionadas à tributação de pessoas jurídicas, com ênfase em questões envolvendo impostos federais. O conjunto de dados foi construído a partir de soluções de consulta selecionadas da Coordenação-Geral de Tributação (Cosit) [1] da Receita Federal do Brasil, abrangendo principalmente dúvidas sobre tributos específicos como Imposto de Importação (II), Imposto sobre Produtos Industrializados (IPI), Imposto sobre Operações Financeiras (IOF), entre outros tributos federais. Este conjunto de dados é especialmente relevante por refletir dúvidas genuínas de contribuintes e fornecer respostas elaboradas por especialistas no campo tributário brasileiro, focando em casos que envolvem a interpretação e aplicação da legislação tributária federal.

4.1.1 Seleção das Perguntas

Inicialmente, extraímos uma subamostra do conjunto abrangente de perguntas e respostas fornecido pela Cosit [1]. O conjunto é composto por mais de 1000 (mil) perguntas e respostas feitas por pessoas jurídicas à Cosit. A cada ano o documento é atualizado e ajustado às mudanças da lei. O documento usado nesse trabalho está atualizado para 2023.

Durante o processo de seleção, focamos em perguntas cujas respostas incluíam apenas as referências legais específicas, como a citação do Normativo completa. A Figura 4.1 ilustra um exemplo de pergunta que se encaixa nos critérios de seleção. O exemplo está fiel a como é representado no documento da Cosit. Na Figura 4.1 é possível ver a pergunta, uma resposta objetiva e o Normativo com o art. correspondente. Antes de proceder com a seleção das perguntas filtramos apenas que possuíam o Normativos, o conjunto disponível caiu para 700 pares de perguntas e respostas.

005

Os condôminos na propriedade de imóveis estão sujeitos à equiparação como pessoa jurídica?

Os condôminos na propriedade de imóveis não são considerados sociedades em comum (antiga sociedade de fato), ainda que deles também façam parte pessoas jurídicas. Assim, a cada condômino pessoa física serão aplicados os critérios de caracterização da empresa individual e demais dispositivo legais, como se ele fosse o único titular da operação imobiliária, nos limites da sua participação.

Normativo: RIR/2018, art. 167.

Figura 4.1: *Exemplo de pergunta e resposta com o respectivo normativo da coleção da Cosit [1]*

É importante notar que na coleção da Cosit há apenas a referência, sem o texto completo do dispositivo legal. A coleta do texto integral dos artigos ou dispositivos legais foi realizada em uma etapa posterior.

Embora a maioria das respostas contivesse essas referências, muitas foram elaboradas de forma extensa pelos especialistas, indo além do escopo da pergunta original ou incluindo detalhes excessivos, como tabelas e múltiplos exemplos. Essas características tornavam tais respostas inadequadas para uso em contextos como RAG.

Para garantir uma avaliação justa dos LLMs, excluímos respostas excessivamente detalhadas ou que não se alinhavam diretamente com as perguntas. Além disso, asseguramos que cada resposta selecionada estivesse vinculada a artigos de lei ou regu-

lamentos específicos, permitindo uma avaliação precisa das capacidades dos modelos em utilizar referências legais para gerar suas respostas. Com isso, mais de uma centena de perguntas e respostas com referência legal foram aptas a compor o *dataset*. Os resultados iniciais desse processo de seleção estão apresentados nas três primeiras colunas da Tabela 4.1.

4.1.2 Criação do Corpus de Normativos

Após a seleção das perguntas e suas respectivas referências legais, procedemos à coleta de cada documento normativo referenciado pelos especialistas nas respostas.

Nesta etapa, foram coletados mais de 30 documentos, incluindo leis, instruções, decretos e pareceres. Cada documento podendo conter até milhares de artigos, compostos por múltiplas disposições. Esses documentos representam uma fração da legislação tributária brasileira e incluem as normas que fundamentam as respostas dos especialistas presentes no conjunto de dados.

Para construir as passagens de ouro, ou seja, o texto na íntegra dos artigos e dispositivos legais referenciados na etapa anterior com o respectivo normativo, como leis, decretos, entre outros. Iniciamos pela coleta de todas os normativos, no total foram coletados 30 normativos. Várias perguntas podem se referir a diferentes artigos ou dispositivos de um único normativo, o que explica o número de 30 documentos, apesar de a quantidade de perguntas chegar às centenas. Este número reflete a complexidade do direito tributário.

Para garantir a qualidade do corpus e possibilitar a resposta às perguntas por RAG (Retorno de Informação), realizamos uma rigorosa limpeza nos normativos, removendo todas as disposições revogadas até a data de criação do conjunto de dados. Além disso, qualquer pergunta cuja base normativa já estivesse sido revogada foi eliminada durante a fase de seleção de perguntas.

Esse corpus também faz parte do *dataset* final, sendo um retrato das leis na data de criação, garantindo que nenhuma mudança altere as respostas. É importante ressaltar que, embora o conjunto original de perguntas e respostas da Cosit tenha sido elaborado e revisado por mais de doze especialistas em direito tributário, o *dataset* desta pesquisa foi construído sem auxílio direto de profissionais jurídicos, o que pode impactar sua qualidade devido a possíveis limitações na interpretação das leis. Para mitigar este risco, foram utilizadas estritamente as leis e artigos explicitamente citados nas respostas originais dos especialistas da Cosit, sem interpretações adicionais da legislação. A Figura 4.2 apresenta uma visão geral dos documentos que compõem o corpus legislativo utilizado nesta pesquisa.

ADI SRF nº 5, de 2001	**Lei nº 6.766, de 1979**
ADN Cosit nº 4, de 1996	**Lei nº 9.249, de 1995**
Decreto-Lei nº 1.381, de 1974	**Lei nº 9.316, de 1996**
Decreto-Lei nº 1.510, de 1976	**Lei nº 9.430, de 1996**
Decreto-Lei nº 1.598, de 1977	**Lei nº 9.532, de 1997**
IN DPRF 21, de 1992	**Lei nº 9.718, de 1998**
IN RFB nº 1.252, de 2012	**Lei nº 11.051, de 2004**
IN RFB nº 1.520, de 2014	**PN CST nº 1, de 1983**
IN RFB nº 1.700, de 2017	**PN CST nº 2, de 1983**
IN RFB nº 2.004, de 2021	**PN CST nº 4, de 1981**
IN RFB nº 2.055, de 2021	**PN CST nº 58, de 1977**
IN SRF nº 213, de 2002	**PN CST nº 72, de 1975**
IN SRF nº 51, de 1978	**PN CST nº 146, de 1975**
IN nº 122, de 1989	**Portaria MF nº 356, de 1988**
Lei nº 6.404, de 1976	**RIR 2018**

Figura 4.2: *Documentos do corpus legislativo*

4.1.3 Coleta de Normas (Passagens de Ouro)

Após a seleção das perguntas e de suas respectivas referências legais, ou seja, leis, artigos e outros dispositivos, assim como a coleta dos documentos normativos que compõem o corpus na etapa anterior, iniciamos a extração de cada artigo ou dispositivo legal referenciado no conjunto de dados. A Figura 4.3 é a passagem de ouro do par pergunta e resposta da Figura 4.1.

Normativo: RIR 2018

Art. 167. Os condôminos na propriedade de imóveis não são considerados sociedades em comum, ainda que pessoas jurídicas também façam parte deles (*Decreto-Lei nº 1.381, de 1974, art. 7º*).

Parágrafo único. A cada condômino, pessoa física, serão aplicados os critérios de caracterização da empresa individual e os demais dispositivos legais, como se ele fosse o único titular da operação imobiliária, nos limites de sua participação (*Decreto-Lei nº 1.381, de 1974, art. 7º, parágrafo único*).

Figura 4.3: *Exemplo de passagem de ouro extraída do normativo RIR 2018*

Esses dispositivos são a íntegra dos dispositivos legais citadas pelos especialistas em suas respostas às perguntas formuladas pelas pessoas jurídicas. Embora essa tarefa tenha exigido um tempo e esforço considerável, foi fundamental para avaliar as capacidades de raciocínio dos LLMs em relação a textos legais. Mesmo o exemplo da Figura 4.1, não ser tão complexo, a resposta dificilmente poderia ser respondida por uma abordagem extrativa. Esse exemplo, é valioso pois o especialista criou um resposta com o texto próximo ao do artigo do normativo citado na pergunta.

Ao final dessa etapa, o conjunto de dados passou a incluir a pergunta, a resposta, a referência à norma e a própria norma completa (passagens de ouro). A Tabela 4.1 apresenta o conjunto de dados final.

Tabela 4.1: *Exemplos de Perguntas, Respostas e Referências Legais*

Pergunta	**Resposta do Especialista**	**Referência**	**Passagem de Ouro**
Quais pessoas jurídicas estão dispensadas de apresentar a ECF?	Estão dispensadas de apresentar a ECF: I - as...	IN RFB nº 2004, de 2021, art. 1º, § 1º.	Art. 1º A Escrituração Contábil Fiscal (ECF) deverá...
Quais os efeitos tributários em caso de retificação da ECF?	Quando a retificação da ECF evidenciar maior imposto devido...	IN RFB nº 2055, de 2021, art. 148.	Art. 148. O crédito relativo a tributo administrado...
A fruição da isenção do IRPJ depende de reconhecimento prévio?	Não. O benefício da isenção do IRPJ não depende...	RIR/2018, art. 192.	Art. 192. As isenções de que trata esta Seção...
Em quais situações a pessoa física é equiparada à pessoa jurídica?	Para fins de imposto de renda, são equiparadas...	RIR/2018, art. 162, § 1º, incisos I a III.	Art. 162. Consideram-se firmas individuais...
Condomínios em edificações estão sujeitos ao imposto de renda?	Os condomínios em edificações não estão sujeitos...	RIR/2018, art. 167.	Art. 167. Os condomínios em edificações...

Além do *dataset* representado na Tabela 4.1, o produto final também inclui o corpus legislativo (Figura 4.2). A conclusão do processo de coleta e preparação dos dados resultou em um conjunto robusto, composto por perguntas, respostas especializadas, referências normativas e as respectivas passagens de ouro, que são os trechos integrais dos dispositivos legais, que não estavam inclusos no conjunto original da Cosit [1], no conjunto original estava apenas a referência. Este processo cuidadoso, que envolveu tanto a seleção de perguntas relevantes quanto a coleta de normativos atualizados, foi fundamental para assegurar a qualidade e precisão do *dataset*. Ao garantir que todas as disposições revogadas fossem excluídas e que apenas normas vigentes fossem incluídas, o corpus final reflete fielmente a legislação tributária brasileira no momento da criação do conjunto de dados [1]. Com este *dataset*, é possível avaliar de forma rigorosa as capacidades dos LLMs em utilizar e interpretar textos legais, tornando-o também adequado para aplicações de RAG.

4.2 Ambiente Experimental

Neste estudo, realizamos uma avaliação abrangente dos LLMs em termos de sua capacidade de raciocinar sobre leis, com foco específico na tributação corporativa para pessoas jurídicass. Avaliamos os LLMs utilizando o *dataset* criado neste trabalho, o qual foram composto por perguntas e respostas reais sobre a tributação, fornecidas por especialistas no assunto.

Selecionamos mais de 20 LLMs para a avaliação, englobando tanto modelos proprietários quanto de código aberto. Os modelos escolhidos incluem exemplos notáveis, como Mistral AI, Llama, Gemma, Qwen, várias versões ajustadas pela comunidade desses modelos, e um modelo proprietário. Cada modelo possui características e capacidades únicas, proporcionando uma variedade diversificada de perspectivas para nossa análise.

Para garantir a consistência nas avaliações, padronizamos o parâmetro de temperatura em 0,1 para todos os modelos escolhidos. Esse ajuste de baixa temperatura foi selecionado para reduzir a aleatoriedade nas saídas, incentivando respostas mais determinísticas. A decisão de não impor um limite máximo de *tokens*, permitindo que os modelos gerassem respostas sem restrições de comprimento, foi tomada após cuidadosa consideração. Embora a imposição de limites de *tokens* pudesse potencialmente melhorar as métricas de avaliação ao forçar respostas mais concisas e focadas, optou-se por não utilizar esta abordagem para preservar a completude do raciocínio jurídico. Esta escolha metodológica fundamenta-se na natureza complexa das questões tributárias, onde respostas mais extensas podem ser necessárias para abordar adequadamente todos os aspectos legais relevantes, incluindo fundamentação, contextualização e referências normativas específicas.

Um *prompt* específico (ver *Prompt* Pergunta-Resposta no Apêndice A.1) foi elaborado para guiar os modelos no raciocínio sobre as leis e na geração de respostas adequadas. O *prompt* instrui explicitamente os modelos a raciocinarem com base no contexto legal fornecido e a formularem uma resposta. Caso um modelo não consiga gerar uma resposta satisfatória, ele é orientado a declarar que não sabe a resposta. Esse *prompt* em particular, foi criado em inglês, seguindo as diretrizes da biblioteca *langchain* [116], que sugere essa abordagem para geração aumentada.

As informações legais necessárias para responder a cada pergunta, como um artigo de lei ou um documento jurídico, são incluídas no *prompt*. Essas informações são as mesmas utilizadas pelos especialistas para criar as respostas de referência, garantindo uma base justa para comparação. Ao utilizar *prompts* padronizados e incorporar as disposições legais relevantes, asseguramos que os modelos tenham acesso às mesmas informações que os especialistas humanos. Isso permite uma avaliação rigorosa de suas capacidades de raciocínio.

É importante destacar que as perguntas e as respostas de referência são apresen-

tadas em português brasileiro. Este aspecto do estudo testa tanto as habilidades de raciocínio dos modelos quanto sua proficiência em gerar respostas precisas e contextualmente apropriadas na língua portuguesa. Dado que muitos LLMs são treinados principalmente em *datasets* em inglês, a avaliação de seu desempenho em textos jurídicos em português brasileiro é essencial para entender a aplicabilidade e as limitações desses modelos em jurisdições que não falam inglês.

Embora o *dataset* utilizado neste experimento contenha um corpus adequado para RAG, nossa avaliação concentrou-se exclusivamente nas tarefas de geração e raciocínio. Essa decisão foi inspirada em outros *datasets* proeminentes, como o SQuAD 2.0 [65] e o *HotpotQA* [68], que também fornecem as passagens esperadas juntamente com as respostas de referência, permitindo uma avaliação direta das capacidades de geração dos modelos, sem a etapa de recuperação. Ao concentrar-se nesses aspectos, nosso objetivo foi isolar e avaliar a capacidade dos LLMs de gerar respostas precisas e fundamentadas, com base unicamente no contexto legal fornecido.

4.2.1 Métricas de Avaliação

Nosso estudo avaliou os LLMs utilizando uma abordagem abrangente, integrando métodos quantitativos e qualitativos. Para a avaliação quantitativa, empregamos as métricas BLEU e ROUGE [10, 11]. No campo do PLN, essas métricas avaliam a qualidade da geração de texto, comparando as respostas dos modelos com um conjunto predefinido de respostas de referência. Especificamente, no domínio de perguntas e respostas relacionadas à tributação corporativa, essas métricas fornecem uma medida quantitativa de quão próximas, em termos lexicais, as respostas geradas estão das respostas ideais.

Apesar de seu uso difundido, métricas como BLEU, ROUGE [107] e METEOR [26], amplamente utilizadas para avaliar modelos de linguagem, oferecem uma perspectiva predominantemente quantitativa e podem não captar totalmente a acurácia das respostas em cenários de perguntas e respostas [16]. Essa limitação decorre do fato de que essas métricas não avaliam adequadamente a acurácia factual ou a relevância das respostas geradas, fatores que são fundamentais para determinar se as perguntas foram respondidas corretamente.

Para lidar com essa lacuna, adotamos uma abordagem qualitativa mais detalhada, utilizando as capacidades de um poderoso modelo de linguagem como substituto do julgamento humano. Especificamente, empregamos o GPT-4 para avaliar o desempenho dos outros modelos. Essa abordagem é baseada na premissa de que um LLM robusto, como o GPT-4, pode emular de maneira eficaz o julgamento humano na avaliação de respostas [13–19] para perguntas abertas, proporcionando uma aproximação mais fiel aos critérios de avaliação humana.

Para a avaliação qualitativa, utilizamos um *prompt* cuidadosamente elaborado para verificar a acurácia factual das respostas dos modelos. Para criar esse *prompt* usamos as técnicas *few-shot prompting*, passando alguns exemplos de avaliação e *Chain of Thought* pedindo para o modelo explicar o raciocínio de avaliação antes de dar a resposta. A definição desse *prompt* foi empírica, ele foi o que melhor avaliou os modelos nos testes. A construção de um *prompt* ou ainda uma avaliação mais robusta pelo LLM podem ser objetos de estudos futuros.

A acurácia dos modelos avaliados foi calculada com base na avaliação do LLM GPT-4. O *prompt* específico utilizado para essa avaliação pode ser encontrado no Apêndice A.2, com mais detalhes.

4.2.2 Implementação

A implementação deste estudo foi realizada em *Python*, utilizando *Jupyter Notebooks* para facilitar a reprodução e compartilhamento do código, que está disponível em um repositório *Git* (Seção 6.2). Os LLMs foram acessados por meio de *APIs* na nuvem, utilizando a *Together AI* [117] para os modelos *open-source* e a *OpenAI* [118] para o modelo proprietário e o processo de avaliação, ambos integrados via a biblioteca *Langchain*, que oferece suporte para comunicação com essas *APIs*.

As seguintes bibliotecas foram empregadas na avaliação e manipulação de dados:

A biblioteca *datasets* foi utilizada para carregar e manipular o *dataset* [119], fornecendo uma interface eficiente para lidar com grandes volumes de dados. A métrica ROUGE foi calculada com a biblioteca *Rouge* [11], que oferece uma implementação robusta e otimizada para essa métrica de avaliação textual. Já a métrica BLEU foi calculada utilizando a biblioteca *sacrebleu* [120], reconhecida por sua padronização e confiabilidade. A similaridade semântica foi avaliada com a biblioteca *bert_score* [12], permitindo uma análise mais refinada da relação entre as respostas geradas e as de referência.

A manipulação e análise de dados tabulares foram feitas com a biblioteca *pandas*, que foi essencial para organizar os resultados dos experimentos [121]. Por fim, a biblioteca *Langchain* facilitou a integração das *APIs* de modelos, orquestrando o fluxo de geração e avaliação das respostas [116].

Esse conjunto de bibliotecas permitiu uma implementação simples, garantindo a reprodutibilidade dos resultados e a análise automatizada das respostas geradas pelos modelos. Todo o código pode ser encontrado no repositório *Git* (Seção 6.2).

CAPÍTULO 5

Resultados

Nesta seção, apresentamos os resultados da avaliação do método proposto, detalhando as métricas utilizadas, os resultados obtidos e uma discussão aprofundada sobre esses resultados. Nosso objetivo é analisar o desempenho dos LLMs no contexto do direito tributário para pessoas jurídicas, identificando suas capacidades e limitações.

5.1 Análise de Desempenho dos Modelos

As versões mais recentes das famílias *Llama*, *Qwen* e *Mistral* apresentam avanços significativos em comparação as suas predecessoras. Esses modelos incorporam diversas melhorias arquiteturais, incluindo a ativação *SwiGLU* [37] e *Grouped Query Attention* (*GQA*) [36]. Tanto o modelo *Qwen2-72B-Instruct* [122] quanto o *Llama-3-70b-chat-hf* [123] se beneficiaram dessas melhorias, especialmente das modificações no *tokenizer* e da inclusão de *GQA*, resultando em ganhos de desempenho notáveis. Como resultado, o modelo *Qwen2-72B-Instruct* [122] alcançou a maior acurácia. Resultados semelhantes foram observados em outros *benchmarks* de avaliação de LLMs [122], destacando o desempenho superior de modelos que incorporam essas técnicas.

A análise de desempenho dos modelos revelou que o tamanho do modelo impacta significativamente os resultados, embora esse impacto nem sempre seja direto. Modelos maiores, como *Qwen2-72B-Instruct* [122] e *Mixtral-8x22B-Instruct-v0.1* [124], obtiveram desempenho superior, apresentando as maiores métricas de ROUGE-L, BLEU, *Bert Score F1* e a acurácia avaliada pelo *GPT-4*. No entanto, observamos que modelos menores, como *Mistral-7B-Instruct-v0.3* [125] e *OpenHermes-2p5-Mistral-7B* [126], superaram alguns modelos maiores em métricas específicas. Por exemplo, o *Mistral-7B-Instruct-v0.3* atingiu um *Bert Score F1* de 0.71, superando vários modelos maiores, e o *OpenHermes-2p5-Mistral-7B* demonstrou desempenho notável com acurácia comparável a modelos significativamente maiores. Esses achados sugerem que, embora modelos maiores geralmente ofereçam melhores resultados devido à sua capacidade de capturar informações mais complexas, modelos menores bem treinados e ajustados podem apresentar desempenho competitivo em contextos específicos. Esse padrão indica que a qualidade do

Tabela 5.1: *Métricas de Desempenho dos Modelos*

Modelo	ROUGE-L	BLEU	Bert Score F1	Acc. GPT-4
Mistral-7B-Instruct-v0.2	0.35	0.20	0.67	0.54
Mistral-7B-Instruct-v0.3	0.40	0.26	0.71	0.55
Mixtral-8x7B-Instruct-v0.1	0.38	0.24	0.70	0.53
Mixtral-8x22B-Instruct-v0.1	**0.44**	**0.30**	**0.73**	0.59
Llama-2-70b-chat-hf	0.38	0.19	0.69	0.49
Llama-2-13b-chat-hf	0.37	0.20	0.68	0.43
Llama-2-7b-chat-hf	0.32	0.14	0.65	0.34
Llama-3-70b-chat-hf	0.34	0.16	0.65	0.60
Llama-3-8b-chat-hf	0.35	0.15	0.65	0.54
Qwen1.5-110B-Chat	0.39	0.21	0.71	0.60
Qwen1.5-72B-Chat	0.41	0.24	0.71	0.62
Qwen1.5-14B-Chat	0.34	0.16	0.68	0.48
Qwen2-72B-Instruct	0.43	0.29	**0.73**	**0.64**
gemma-7b-it	0.40	0.22	0.70	0.45
Yi-34B-Chat	0.38	0.26	0.70	0.52
gpt-3.5-turbo	0.38	0.15	0.69	0.56
Platypus2-70B-instruct	0.41	0.29	0.70	0.57
vicuna-13b-v1.5	0.41	0.27	0.71	0.50
vicuna-7b-v1.5	0.37	0.23	0.69	0.39
openchat-3.5-1210	0.42	0.28	0.72	0.55
WizardLM-13B-V1.2	0.36	0.25	0.68	0.49
SOLAR-10.7B-Instruct-v1.0	0.36	0.23	0.70	0.51
OpenHermes-2p5-Mistral-7B	0.41	0.25	0.71	0.55

treinamento e o ajuste do modelo a *datasets* diversos, mesmo fora do domínio, são fatores que podem mitigar a disparidade de tamanho entre os modelos.

Embora o volume de dados em português utilizados no treinamento desses modelos ainda precise ser verificado, as melhorias arquiteturais e de treinamento sugerem um desempenho aprimorado dos LLMs em tarefas Q&A no domínio do direito tributário. Na família *Mistral*, o modelo *Mixtral-8x22B-Instruct-v0.1* [127] se destacou com as maiores pontuações em *ROUGE-L*, *BLEU* e *Bert Score F1*, indicando o potencial da arquitetura de *mixture of experts* [124] para textos legais em português.

A análise dos modelos *open-source* ajustados (*fine-tuned*) revela melhorias significativas em relação aos modelos base. Os modelos *openchat-3.5-1210* [128] e *OpenHermes-2p5-Mistral-7B* [126], ambos derivados do *Mistral-7B-v0.1* [125], apresentaram aumentos notáveis de acurácia após o *fine-tuning* em *datasets* abertos. Da mesma forma, os modelos *vicuna-13b-v1.5* e *vicuna-7b-v1.5* [19], ajustados a partir do *Llama 2* [129], também demonstraram avanços na acurácia das respostas. Além disso, modelos como *WizardLM-13B-V1.2* [130], *SOLAR-10.7B-Instruct-v1.0* [131, 132], e *Platypus2-70B-instruct* [133], derivados do *Llama 2*, melhoraram os resultados de seus modelos base. Notavelmente, esses processos de *fine-tuning* foram realizados em *datasets* diver-

sos, e não no *dataset* experimental em si, mas ainda assim levaram a métricas aprimoradas no *dataset* experimental. Essas melhorias sugerem que o *fine-tuning* pode aprimorar efetivamente as capacidades de geração de textos legais e de tarefas de Q&A quando aplicado a *datasets* específicos.

5.1.1 Origem e Características dos Modelos Avaliados

Este trabalho focou exclusivamente na avaliação do desempenho de modelos de linguagem previamente treinados, sem realizar nenhum processo de treinamento ou *fine-tuning* adicional. Entre os modelos avaliados, diversos são resultados de processos de *fine-tuning* realizados pela comunidade em modelos base conhecidos. O modelo *openchat-3.5-1210* [128] e *OpenHermes-2p5-Mistral-7B* [126] são derivados do *Mistral-7B-v0.1* [125], tendo sido ajustados em *datasets* abertos da comunidade para melhorar seu desempenho em tarefas conversacionais. Os modelos *vicuna-13b-v1.5* e *vicuna-7b-v1.5* [19] são baseados no *Llama 2* [129], assim como o *WizardLM-13B-V1.2* [130], *SOLAR-10.7B-Instruct-v1.0* [131, 132] e *Platypus2-70B-instruct* [133].

É importante ressaltar que nenhum desses modelos passou por processo de *fine-tuning* específico para o domínio do direito tributário ou para o *dataset* experimental deste trabalho. Os processos de *fine-tuning* mencionados foram realizados pela comunidade em *datasets* diversos e de domínio geral. Por exemplo, o *OpenHermes-2p5-Mistral-7B* foi ajustado utilizando um *dataset* aberto focado em instruções gerais, enquanto o *vicuna* utilizou dados conversacionais compartilhados pela comunidade. O presente trabalho se concentrou exclusivamente em avaliar as capacidades de raciocínio e geração de texto desses modelos em seu estado atual, sem qualquer ajuste adicional, permitindo assim uma análise objetiva de suas capacidades em um domínio específico mesmo quando treinados em contextos mais amplos.

5.1.2 Análise das Métricas de Avaliação

As métricas tradicionais, como BLEU e ROUGE, podem não capturar completamente as nuances necessárias para uma avaliação precisa em tarefas de Q&A. A métrica *Bert Score F1* é reconhecida por seu alinhamento com a avaliação humana, devido à sua capacidade de capturar semelhanças semânticas profundas entre os textos, superando as capacidades de correspondência lexical das métricas tradicionais, como *ROUGE-L* e *BLEU*. Esses achados foram documentados no artigo do *Bert Score* [12].

Embora este estudo não tenha como objetivo provar que LLMs podem atuar como avaliadores perfeitamente alinhados à avaliação humana, estudos recentes têm explorado essa correlação [16–19]. Nossa pesquisa avalia a qualidade das respostas geradas pelos LLMs em tarefas de Q&A no domínio jurídico. A maior correlação

entre a Acurácia do LLM (*GPT-4*) e o *Bert Score F1*, conforme demonstrado pelas correlações de *Pearson* (0.657) e *Kendall* (0.491) (ver Tabela 5.2), sugere que ambas as métricas capturam aspectos semânticos relevantes para a qualidade percebida por humanos. Os resultados estão em conformidade com estudos [12] que recomendam o uso das correlações de *Pearson* e *Kendall* para avaliar a qualidade das métricas.

As correlações foram calculadas considerando os resultados agregados de cada modelo no conjunto completo de dados, onde cada modelo gerou respostas para todas as perguntas da base de avaliação. Para cada modelo, foram computadas diferentes métricas de avaliação: *Rouge-L*, *BLEU*, *Bert Score F1* e a acurácia avaliada pelo *GPT-4*. As correlações de *Pearson* e *Kendall* foram então calculadas entre estas métricas utilizando os valores médios obtidos por cada modelo, ou seja, cada observação na matriz de correlação representa o desempenho global de um modelo específico em todas as perguntas e respostas do conjunto de dados. Este método permite analisar como as diferentes métricas automáticas se relacionam entre si e com a avaliação do *GPT-4*, que serve como aproximação da avaliação humana. Em particular, a correlação encontrada entre o *Bert Score F1* e a acurácia do *GPT-4* (0.657 para *Pearson* e 0.491 para *Kendall*) indica uma concordância moderada entre estas métricas na avaliação da qualidade das respostas geradas.

Tabela 5.2: *Coeficientes de Correlação entre as Métricas*

Métricas	**Pearson**	**Kendall**
Acurácia (GPT-4) ↔ BERTScore F1	0,658	0,491
Acurácia (GPT-4) ↔ ROUGE-L	0,539	0,393
Acurácia (GPT-4) ↔ BLEU	0,373	0,289
BERTScore F1 ↔ ROUGE-L	0,908	0,838
BERTScore F1 ↔ BLEU	0,781	0,666
ROUGE-L ↔ BLEU	0,821	0,670

Embora seja usada, alguns estudos sugerem que a correlação não é uma boa estratégia para comparar técnicas de medição [134]. Em contrapartida, a análise de *Bland-Altman* é particularmente adequada para comparar métodos de medição [134].

O cálculo de Bland-Altman é realizado primeiramente obtendo-se a média entre os dois métodos de medição (*means*) e calculando as diferenças entre eles (*differences*). O viés (*bias*) é então calculado como a média dessas diferenças, representando o erro sistemático entre os métodos. O desvio padrão das diferenças é calculado para determinar a variabilidade dessas diferenças. O valor 1,96 é utilizado por ser o valor crítico da distribuição normal para um intervalo de confiança de 95%, assumindo que as diferenças seguem uma distribuição normal. Multiplicando 1,96 pelo desvio padrão e somando/subtraindo do viés, obtêm-se os limites superior e inferior de concordância (*upper_limit* e *lower_limit*), que definem o intervalo onde espera-se encontrar 95% das diferenças entre

os métodos. Em outras palavras, o desvio padrão quantifica a dispersão das diferenças, enquanto o valor 1,96 permite estabelecer um intervalo que captura 95% dessas diferenças, assumindo normalidade dos dados.

Essa análise confirma que *Bert Score F1* é mais concordante com a Acurácia calculada pelo LLM (*GPT-4*) do que as métricas lexicais, como evidenciado pela menor variação na dispersão dos pontos e pela menor amplitude dos limites de concordância (ver Figura 5.1). Na análise de Bland-Altman comparando ROUGE-L e Bert Score F1 com a acurácia do GPT-4, observou-se que, embora inicialmente o ROUGE-L aparente ter maior concordância devido ao seu viés menor (-0.14 vs 0.23), uma análise mais profunda revela que o *Bert Score F1* possui concordância superior. Esta conclusão baseia-se em três aspectos principais: primeiro, ambas as métricas têm amplitude similar nos limites de concordância (0.23), mas o *Bert Score F1* apresenta distribuição mais sistemática dos pontos; segundo, o *Bert Score F1* mostra uma tendência previsível de diminuição das diferenças conforme a média aumenta, enquanto o ROUGE-L tem dispersão errática; terceiro, o viés maior do *Bert Score F1* não compromete sua superioridade, pois representa um erro sistemático corrigível, ao contrário da variabilidade imprevisível do ROUGE-L, que constitui uma fonte de incerteza mais problemática na medição. Portanto, em termos de diferença média, o *Bert Score F1* é considerado melhor não pela magnitude das diferenças em si, mas pela forma como essas diferenças se comportam de maneira mais sistemática e previsível ao longo das medições.

Esses achados indicam que a Acurácia do LLM (*GPT-4*), de forma semelhante ao *Bert Score F1*, pode ser uma métrica viável e representativa para avaliar o desempenho real dos LLMs. Embora ROUGE-L e BLEU apresentem correlações mais altas com o *Bert Score F1*, a correlação e concordância mais fortes da Acurácia do LLM com o *Bert Score F1* indicam seu potencial alinhamento com a avaliação humana. Isso apoia o desenvolvimento de métricas de avaliação que reflitam mais precisamente a qualidade percebida por humanos, alinhando-se à direção das pesquisas atuais que investigam o potencial dos LLMs para se alinharem à avaliação humana [16–19].

5.1.3 Discussão dos Resultados

Os resultados deste trabalho revelam informações relevantes sobre a aplicabilidade dos LLMs no domínio do direito tributário brasileiro, particularmente para consultas de pessoas jurídicas. Os resultados demonstram que avanços arquiteturais recentes, como a incorporação de *SwiGLU* e *Grouped Query Attention* (GQA), têm impacto direto na qualidade das respostas geradas. O *Qwen2-72B-Instruct*, que implementa essas melhorias, alcançou a maior acurácia (0.64) em nossa avaliação, sugerindo que a combinação de uma arquitetura otimizada com um grande número de parâmetros (72B) oferece van-

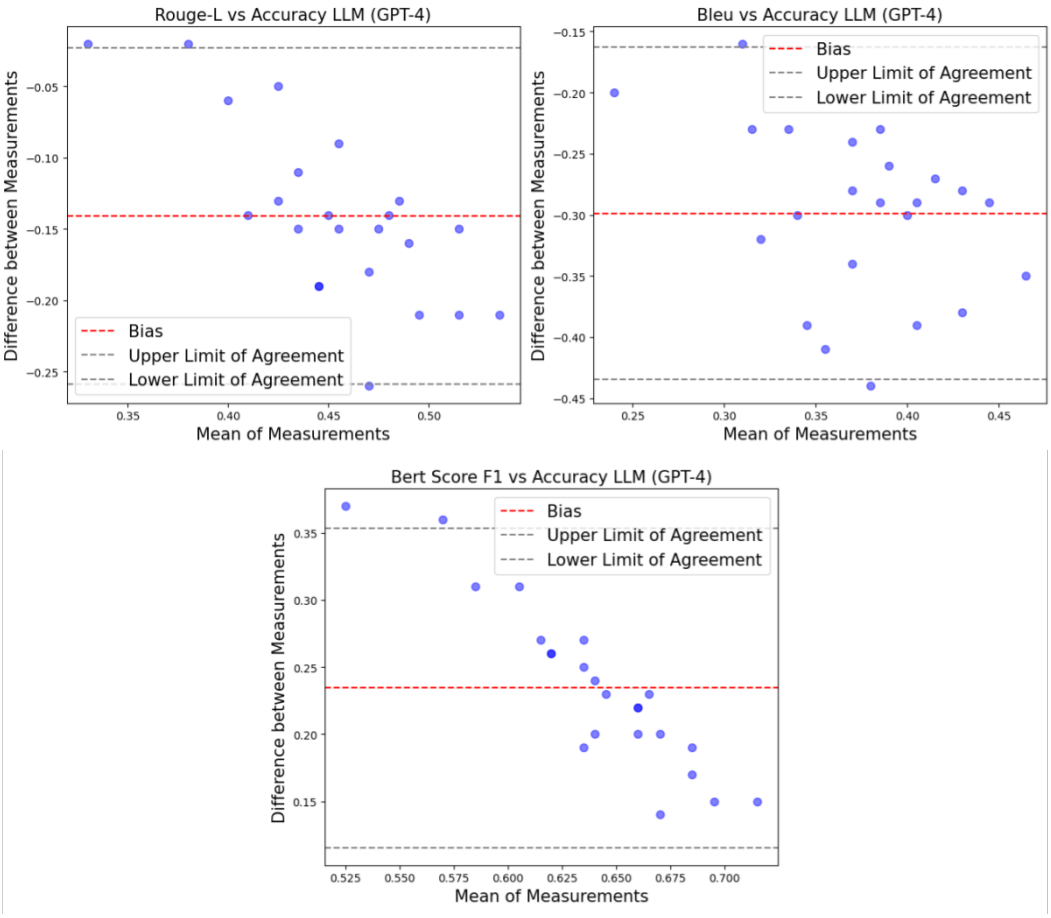


Figura 5.1: *Análise de Bland-Altman das Métricas vs Acurácia Avaliada pelo GPT-4*

tagens significativas para o processamento de textos jurídicos em português.

A análise comparativa entre diferentes famílias de modelos revela padrões importantes para a seleção de LLMs em contextos jurídicos. Os modelos da família Mixtral, utilizando a arquitetura de *mixture of experts* (MoE), demonstraram desempenho notável, com o *Mixtral-8x22B-Instruct-v0.1* atingindo as maiores pontuações em métricas baseadas em texto (ROUGE-L: 0.44, BLEU: 0.30, *Bert Score F1*: 0.73). Este resultado sugere que a arquitetura MoE pode ser particularmente eficaz para processar a complexidade e especificidade da linguagem jurídica, possivelmente devido à sua capacidade de ativar diferentes especialistas para diferentes aspectos do texto legal.

Um resultado importante é que o tamanho do modelo não é o único determinante de desempenho. Modelos menores bem ajustados, como o *Mistral-7B-Instruct-v0.3* (*Bert Score F1*: 0.71) e *OpenHermes-2p5-Mistral-7B*, demonstraram competitividade com modelos significativamente maiores. Este padrão indica que, para aplicações com restrições de recursos computacionais, modelos menores com *fine-tuning* adequado podem oferecer um equilíbrio eficaz entre desempenho e eficiência. Em cenários que priorizam acurácia máxima, como suporte a decisões jurídicas complexas, os modelos *Qwen2-72B-Instruct*

ou *Mixtral-8x22B-Instruct-v0.1* se mostram mais adequados pela sua maior acurácia geral e melhor capacidade de processamento de nuances jurídicas.

Uma descoberta relevante é o impacto positivo do *fine-tuning* mesmo em *datasets* não específicos do domínio jurídico. Modelos como *openchat-3.5-1210* e *vicuna-13b-v1.5* demonstraram melhorias significativas após ajustes em *datasets* diversos, sugerindo que o *fine-tuning* pode aprimorar a capacidade geral de processamento de linguagem natural, beneficiando também o domínio jurídico. Para sistemas que requerem baixa latência, os modelos da família *Llama-3* de menor escala apresentam bom desempenho, sendo adequados para aplicações que precisam de respostas em tempo real.

A correlação forte entre a avaliação do GPT-4 e o *Bert Score F1* sugerem que essas métricas capturam aspectos semânticos relevantes para o contexto jurídico. No entanto, reconhecemos limitações importantes, mesmo os modelos com melhor desempenho ainda apresentam lacunas significativas em relação à expertise humana, especialmente em interpretações jurídicas mais complexas. A variação nas métricas de avaliação sugere que diferentes aspectos da qualidade da resposta são capturados por diferentes métricas.

Identificamos como possíveis direções para avanços futuros o desenvolvimento de técnicas de *fine-tuning* específicas para o domínio jurídico em português, a integração mais efetiva de conhecimento jurídico estruturado nos modelos e o aprimoramento dos métodos de avaliação para capturar melhor a acurácia jurídica das respostas. Estas direções foram reveladas pelos resultados e podem ser um passo para aproximar o desempenho dos LLMs ao nível de expertise necessário para aplicações jurídicas práticas, especialmente no contexto do direito tributário brasileiro.

CAPÍTULO 6

Conclusão

Este estudo destaca a importância do direito tributário na sociedade e o potencial dos modelos de linguagem para auxiliar em sua compreensão e aplicação. Desenvolvemos um *dataset* com perguntas reais sobre direito tributário e respostas de especialistas em português brasileiro, e conduzimos uma avaliação com diversos modelos de linguagem. Embora nossos resultados sugiram que esses modelos apresentam potencial para compreender e raciocinar sobre textos legais complexos, são necessárias mais pesquisas para demonstrar plenamente sua eficácia no raciocínio jurídico em uma gama mais ampla de cenários e tarefas.

Nossa avaliação mostrou que os avanços na arquitetura dos modelos têm um impacto perceptível no desempenho, e que o *fine-tuning* de modelos *open-source*, mesmo quando feito em *datasets* diversos e não específicos do domínio jurídico, ainda podem melhorar sua capacidade de gerar respostas relevantes e precisas. Isso sugere que melhorias contínuas e adaptações são importantes para aprimorar as capacidades dos modelos de linguagem em tarefas jurídicas.

Para avaliar o desempenho dos modelos, utilizamos o *Bert Score F1*, conhecido por sua forte correlação com avaliações humanas em tarefas que envolvem entendimento descritivo e estrutural, e uma métrica mais recente, a avaliação de acurácia por LLM. Embora o *Bert Score F1* já seja uma medida consolidada e alinhada ao julgamento humano, especialmente em tarefas descritivas, nossos resultados mostraram que a avaliação de acurácia por LLM demonstrou forte correlação com o *Bert Score F1* através das correlações de *Pearson* e *Kendall*. A análise de *Bland-Altman* confirmou ainda mais que a métrica *LLM* se alinha de perto com o *Bert Score*, sugerindo seu potencial como uma alternativa confiável nas avaliações. No entanto, é importante notar que, embora esses resultados sejam encorajadores, o uso dessas métricas para tarefas baseadas em raciocínio, como as deste estudo, ainda requer validação adicional. A métrica *LLM* é uma ferramenta promissora, mas mais pesquisas são necessárias para estabelecer plenamente sua eficácia, especialmente na captura das nuances do raciocínio jurídico.

6.1 Limitações e Trabalhos Futuros

Uma limitação significativa do nosso estudo é que, embora tenhamos focado na avaliação das capacidades de geração e raciocínio dos LLMs, não exigimos que os modelos identificassem disposições legais específicas como parte de suas respostas. Apesar de nosso *dataset* incluir um corpus abrangente contendo as leis necessárias, não exploramos completamente seu potencial para técnicas de RAG. Trabalhos futuros poderiam explorar a integração do RAG para aprimorar a habilidade dos modelos de não apenas gerar respostas corretas, mas também identificar e citar as disposições legais apropriadas, alcançando um raciocínio estatutário mais robusto e abrangente.

Outra limitação importante refere-se à avaliação da qualidade das respostas, que foi realizada principalmente através de métricas automatizadas e avaliação por LLM (GPT-4). Embora tenhamos demonstrado correlações significativas entre essas métricas e encontrado resultados promissores, a ausência de avaliação direta por especialistas em direito tributário limita nossa compreensão da real aplicabilidade prática das respostas geradas. Pesquisas futuras poderiam incorporar um painel de especialistas jurídicos para avaliar não apenas a acurácia técnica das respostas, mas também sua utilidade prática e adequação ao contexto legal brasileiro.

A dimensão temporal do conjunto de dados representa outra limitação relevante, uma vez que o direito tributário está em constante evolução, com frequentes atualizações e mudanças na legislação. Nosso *dataset* reflete um momento específico no tempo, não capturando atualizações posteriores nas leis e regulamentações. Trabalhos futuros poderiam desenvolver metodologias para manter o corpus legal atualizado automaticamente, possivelmente através de técnicas de *web scraping* e processamento automático de documentos legais, garantindo que os modelos trabalhem sempre com a legislação mais recente.

Uma limitação adicional diz respeito à escala do conjunto de dados utilizado, que representa apenas uma fração do universo de questões possíveis no direito tributário brasileiro. O número limitado de exemplos pode restringir a capacidade dos modelos de generalizar para casos mais complexos ou incomuns. Pesquisas futuras poderiam expandir o *dataset* através da coleta sistemática de mais perguntas e respostas, possivelmente incluindo decisões judiciais e pareceres técnicos, além de explorar técnicas de *data augmentation* específicas para o domínio jurídico.

A ausência de análise do impacto do viés linguístico constitui outra limitação importante. Como muitos dos modelos foram primariamente treinados em língua inglesa, não foi possível determinar completamente como isso afeta seu desempenho em português, especialmente no contexto jurídico brasileiro. Trabalhos futuros poderiam investigar especificamente o impacto do idioma no desempenho dos modelos, possivelmente através de *fine-tuning* com corpus jurídicos em português e análise comparativa com modelos

treinados primariamente em português.

Por fim, nosso estudo limitou-se a avaliar as respostas dos modelos em um formato relativamente simples de pergunta e resposta, não explorando cenários mais complexos como análise de casos hipotéticos ou interpretação de situações que envolvam múltiplas disposições legais inter-relacionadas. Pesquisas futuras poderiam explorar a capacidade dos modelos em lidar com casos mais complexos, incluindo análise de precedentes, interpretação de jurisprudência e avaliação de cenários que envolvam múltiplas áreas do direito tributário simultaneamente.

6.2 Disponibilidade do Conjunto de Dados e Código

O conjunto de dados utilizado neste estudo, bem como o código para reproduzir os experimentos e análises, estão disponíveis publicamente. O conjunto de dados pode ser acessado no seguinte link: https://github.com/joaopaulopresa/dataset. O código está disponível em: https://github.com/joaopaulopresa/code.

Referências

[1] General Coordination of Taxation (Cosit). Questions and answers for legal entities 2023. https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/declaracoes-e-demonstrativos/ecf/perguntas-e-respostas-pj-2023.pdf, 2023. Accessed: 11/11/2023.

[2] Humberto Ávila. Direito tributário. *Porto Alegre: Verbo Jurídico*, 2008.

[3] André FOLLONI and Camila Beatriz SIMM. Direito tributário, complexidade e análise econômica do direito. *Revista Eletrônica do Curso de Direito da UFSM*, 11(1):49–70, 2016.

[4] Miguel Teixeira De Sousa. *Introdução ao direito*. Leya, 2023.

[5] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*, 2020.

[6] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. Neural legal judgment prediction in english. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4302–4313, 2020.

[7] Vu Tran, Minh Le Nguyen, and Ken Satoh. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. In *Proceedings of the seventeenth international conference on artificial intelligence and law*, pages 275–282, 2019.

[8] Lorenzo Marocco Diehl. A tributação dos incentivos fiscais de imposto sobre circulação de mercadorias e serviços (icms): a ilegalidade da incidência de imposto de renda de pessoa jurídica (irpj) e da contribuição social sobre o lucro líquido (csll). 2021.

[9] Alexandre Machry and Caroline Orth. Análise tributária do regime simples nacional: Comparação entre as recomendações da ocde e do cpee e as alterações propostas pela anfip. *REVISTA DE CONTABILIDADE DOM ALBERTO*, 12(23):54–77, 2023.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[11] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[12] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[13] Yu Du, Fangyun Wei, and Hongyang Zhang. Anytool: Self-reflective, hierarchical agents for large-scale api calls. *arXiv preprint arXiv:2402.04253*, 2024.

[14] Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. Is gpt-4 a reliable rater? evaluating consistency in gpt-4's text ratings. In *Frontiers in Education*, volume 8, page 1272229. Frontiers Media SA, 2023.

[15] Charles Koutcheme, Nicola Dainese, Sami Sarsa, Arto Hellas, Juho Leinonen, and Paul Denny. Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4-as-a-judge. *arXiv preprint arXiv:2405.05253*, 2024.

[16] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[17] Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. Evaluation metrics in the era of gpt-4: Reliably evaluating large language models on sequence to sequence tasks. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[18] Fangyun Wei, Xi Chen, and Lin Luo. Rethinking generative large language model evaluation for semantic comprehension. *arXiv e-prints*, pages arXiv–2403, 2024.

[19] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[20] Kevin D Ashley. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, Cambridge, 2017.

[21] L Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. Scalable and explainable legal prediction. *Artificial Intelligence and Law*, 29:213–238, 2021.

[22] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020.

[23] Harry Surden. 719Ethics of AI in Law: Basic Questions. In *The Oxford Handbook of Ethics of AI*. Oxford University Press, 07 2020.

[24] Doris Liebwald. Law's capacity for vagueness. *International Journal for the Semiotics of Law*, 26(2):391–423, 2013.

[25] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.

[26] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. Interpretable long-form legal question answering with retrieval-augmented large language models. *arXiv preprint arXiv:2309.17050*, 2023.

[27] João Presa, Sávio Teles de Oliveira, and Celso Camilo-Junior. Evaluating large language models for tax law reasoning. In *BRACIS 2024 ()*, may 2024.

[28] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[29] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[30] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[31] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[32] Felipe Oliveira de Sousa. O raciocínio jurídico entre princípios e regras. 2011.

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large

language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[34] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[35] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[36] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, 2023.

[37] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[38] Mohammad Affan Habib, Shehryar Amin, Muhammad Oqba, Sameer Jaipal, Muhammad Junaid Khan, and Abdul Samad. Taxtajweez: A large language model-based chatbot for income tax information in pakistan using retrieval augmented generation (rag). In *The International FLAIRS Conference Proceedings*, volume 37, 2024.

[39] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696, 2020.

[40] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[41] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[42] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[43] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[44] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[45] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[46] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[47] Brydon T Wang. Prompts and large language models: A new tool for drafting, reviewing and interpreting contracts? *Law, Technology and Humans*, 6(2):88–106, 2024.

[48] Fangyi Yu, Lee Quartey, and Frank Schilder. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596, 2023.

[49] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[50] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[51] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

[52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[53] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

[54] Jihyeok Kim, Seungtaek Choi, Reinald Kim Amplayo, and Seung-won Hwang. Retrieval-augmented controllable review generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2284–2295, 2020.

[55] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*, 2020.

[56] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*, 2023.

[57] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023.

[58] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*, 2024.

[59] Susav Shrestha, Narasimha Reddy, and Zongwang Li. Espn: Memory-efficient multi-vector information retrieval. In *Proceedings of the 2024 ACM SIGPLAN International Symposium on Memory Management*, pages 95–107, 2024.

[60] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[61] Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*, 2023.

[62] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11), 2021.

[63] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*, 2018.

[64] Anna Rogers, Matt Gardner, and Isabelle Augenstein. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45, 2023.

[65] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

[66] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, 2017.

[67] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

[68] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[69] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*, 2021.

[70] Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. Question answering is a format; when is it useful? *arXiv preprint arXiv:1909.11291*, 2019.

[71] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59, 2019.

[72] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

[73] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*, 2023.

[74] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, 2022.

[75] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[76] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, Li Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022.

[77] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[78] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[79] Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168, 2021.

[80] Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 19–28, 2017.

[81] Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698, 2017.

[82] Helena Haapio and Margaret Hagan. Design patterns for contracts. In *Networks. Proceedings of the 19th international legal informatics symposium IRIS*, pages 381–388, 2016.

[83] Enrico Francesconi, Sebastiano Faro, Elisabetta Marinai, and G Perugi. A methodological framework for thesaurus semantic interoperability. In *Proceeding of the Fifth European Semantic Web Conference*, pages 76–87, 2008.

[84] Yann Hicke, Anmol Agarwal, Qianou Ma, and Paul Denny. Chata: Towards an intelligent question-answer teaching assistant using open-source llms. *arXiv preprint arXiv:2311.02775*, 2023.

[85] Joao Monteiro, Pierre-Andre Noel, Etienne Marcotte, Sai Rajeswar, Valentina Zantedeschi, David Vazquez, Nicolas Chapados, Christopher Pal, and Perouz Taslakian. Repliqa: A question-answering dataset for benchmarking llms on unseen reference content. *arXiv preprint arXiv:2406.11811*, 2024.

[86] Hongyu Yang, Liyang He, Min Hou, Shuanghong Shen, Rui Li, Jiahui Hou, Jianhui Ma, and Junda Zhao. Aligning llms through multi-perspective user preference ranking-based feedback for programming question answering. *arXiv preprint arXiv:2406.00037*, 2024.

[87] Hanane Djeddal, Pierre Erbacher, Raouf Toukal, Laure Soulier, Karen Pinel-Sauvagnat, Sophia Katrenko, and Lynda Tamine. An evaluation framework for attributed information retrieval using large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5354–5359, 2024.

[88] Liang Zhang, Katherine Jijo, Spurthi Setty, Eden Chung, Fatima Javid, Natan Vidra, and Tommy Clifford. Enhancing large language model performance to answer questions and extract information more accurately. *arXiv preprint arXiv:2402.01722*, 2024.

[89] Ashkan Alinejad, Krtin Kumar, and Ali Vahdat. Evaluating the retrieval component in llm-based question answering systems. *arXiv preprint arXiv:2406.06458*, 2024.

[90] Akchay Srivastava and Atif Memon. Towards robust evaluation: A comprehensive taxonomy of datasets and metrics for open domain question answering in the era of large language models. *IEEE Access*, 2024.

[91] Anand Subramanian, Viktor Schlegel, Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Vijay Prakash Dwivedi, and Stefan Winkler. M-qalm: A benchmark to assess clinical reading comprehension and knowledge recall in large language models via question answering. *arXiv preprint arXiv:2406.03699*, 2024.

[92] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqa: A dataset for multimodal question answering on scientific papers. *arXiv preprint arXiv:2407.09413*, 2024.

[93] Minsang Kim, Cheoneum Park, and Seungjun Baek. Qpaug: Question and passage augmentation for open-domain question answering of llms, 2024.

[94] Meghana Moorthy Bhat, Rui Meng, Ye Liu, Yingbo Zhou, and Semih Yavuz. Investigating answerability of llms for long-form question answering. *arXiv preprint arXiv:2309.08210*, 2023.

[95] Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. Qgeval: Benchmarking multi-dimensional evaluation for question generation. *arXiv preprint arXiv:2406.05707*, 2024.

[96] Alexey Gorbatovski and Sergey Kovalchuk. Reinforcement learning for question answering in programming domain using public community scoring as a human feedback. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2294–2296, 2024.

[97] Kietikul Jearanaitanakij, Chananchida Srithongdee, Sirinoot Ketkham, Onwanya Ardsana, Tiwat Kullawan, and Chankit Yongpiyakul. Thai question-answering system using similarity search and llm. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 18(3):406–416, 2024.

[98] Junyuan Liu, Zhengyan Shi, and Aldo Lipani. Summequal: Summarization evaluation via question answering using large language models. In *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ ACL 2024)*, pages 46–55, 2024.

[99] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*, 2023.

[100] Shengjie Ma, Chong Chen, Qi Chu, and Jiaxin Mao. Leveraging large language models for relevance judgments in legal case retrieval. *arXiv preprint arXiv:2403.18405*, 2024.

[101] Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. Better call gpt, comparing large language models against lawyers. *arXiv preprint arXiv:2401.16212*, 2024.

[102] Joel Niklaus, Lucia Zheng, Arya D McCarthy, Christopher Hahn, Brian M Rosen, Peter Henderson, Daniel E Ho, Garrett Honke, Percy Liang, and Christopher Manning. Flawn-t5: An empirical examination of effective instruction-tuning data mixtures for legal reasoning. *arXiv preprint arXiv:2404.02127*, 2024.

[103] Ruizhe Zhang, Haitao Li, Yueyue Wu, Qingyao Ai, Yiqun Liu, Min Zhang, and Shaoping Ma. Evaluation ethics of llms in legal domain. *arXiv preprint arXiv:2403.11152*, 2024.

[104] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[105] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[106] Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. Laiw: A chinese legal large language models benchmark (a technical report). *arXiv preprint arXiv:2310.05620*, 2023.

[107] Linan Yue, Qi Liu, Yichao Du, Weibo Gao, Ye Liu, and Fangzhou Yao. Fedjudge: Federated legal large language model. *arXiv preprint arXiv:2309.08173*, 2023.

[108] Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, et al. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*, 2023.

[109] Lifeng Xu, Chuanrui Hu, Hua Zhang, Jiahui Zhai, Wei Tang, Yuchen Li, Zhao Peng, Qiuwu Chen, Shiyu Sun, Ao Ji, et al. Surpassing human counterparts: A breakthrough achievement of large language models in professional tax qualification examinations in china. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 1365–1370. IEEE, 2024.

[110] John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*, 382(2270):20230159, 2024.

[111] Libin Zhang. Four tax questions for chatgpt and other language models. 2023.

[112] Jiayuan Luo, Songhua Yang, Xiaoling Qiu, Panyu Chen, Yufei Nai, Wenxuan Zeng, Wentao Zhang, and Xinke Jiang. Kuaiji: the first chinese accounting large language model. *arXiv preprint arXiv:2402.13866*, 2024.

[113] Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. "according to...": Prompting language models improves quoting from pre-training data. *arXiv preprint arXiv:2305.13252*, 2023.

[114] Benjamin Alarie. The rise of the robotic tax analyst. *Tax Notes Federal*, page 57, 2023.

[115] Ahmad Syarief Amrullah, Amelia Cahyadini, and Tasya Safiranita. Potensi artificial intelligece (ai) dalam pelayanan dan pengawasan pajak di indonesia ditinjau dari uu ite, pp pste dan uu kup. *Equality: Journal of Law and Justice*, 1(2):79–94, 2024.

[116] Harrison Chase. LangChain, October 2022.

[117] Together AI. Together ai. https://www.together.ai/, 2023. Acessado em 1 de outubro de 2023.

[118] OpenAI. Openai. https://openai.com/, 2023. Acessado em 1 de outubro de 2023.

[119] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[120] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[121] The pandas development team. pandas-dev/pandas: Pandas, September 2024.

[122] Qwen2 blog. https://qwenlm.github.io/blog/qwen2/, 2024. Accessed: 08/06/2024.

[123] AI@Meta. Llama 3 model card, 2024.

[124] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[125] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[126] Inferless Team. Openhermes-2-5-mistral-7b, 2024. Accessed on: 15 May 2024.

[127] Mistral.ai. Introducing the mixtral-8x22b-instruct-v0.1 model, 2024. Accessed on: 15 May 2024.

[128] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. In *The Twelfth International Conference on Learning Representations*, 2023.

[129] AI@Meta Touvron, H. Llama 2: Open foundation and fine-tuned chat models, 2023.

[130] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

[131] Dahyun Kim, Yungi Kim, Wonho Song, Hyeonwoo Kim, Yunsu Kim, Sanghoon Kim, and Chanjun Park. sdpo: Don't use your data all at once, 2024.

[132] Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling, 2023.

[133] Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. Platypus: Quick, cheap, and powerful refinement of llms. 2023.

[134] Shahab Haghayegh, Hyeon-Ah Kang, Sepideh Khoshnevis, Michael H Smolensky, and Kenneth R Diller. A comprehensive guideline for bland–altman and intra class correlation calculations to properly compare two methods of measurement and interpret findings. *Physiological measurement*, 41(5):055012, 2020.

APÊNDICE A

Apêndice 1

A.1 *Prompt* Perguntas e Respostas

```
Use the following pieces of legal information from laws to
   answer the user's question.
If the answer is not clear in context, try to figure out by
   interpreting the information.
If you don't know the answer, just say that you don't know,
   don't try to make up an answer.
Context: {context}
Question: {question}
Do not quote the "contextual information" provided in the
   answer, do not say "according to the information" or
   anything like that, use the information only to answer the
    question.
Only return the helpful answer below and nothing else.
Answer the question in Portuguese.
Helpful answer:
```

A.2 Prompt de Avaliação

```
Instruções:
Avalie a resposta gerada pela IA com base nos seguintes crité
   rios:

1. Verifique se a Resposta da IA está contida na resposta
   Resposta do Especialista, ou seja, se não existe nenhuma
   contradição. Ignore termos diferentes ou pequenas informaç
```

```
   ões a mais ou a menos.
2. A Resposta do Especialista pode conter mais informações do
    que foi solicitado na pergunta, se a informação da
   Resposta do Especialista não for necessária para responder
    a pergunta não use para avaliar a Resposta da IA.
3. Se a Resposta da IA tiver mais informações do que a
   Resposta do Especialista não deve ser levada em consideraç
   ão para avaliação desde que as informações estejam
   corretas.
3. Verifique se a resposta pode responder à questão. Para
   isso veja se a resposta fornece as informações solicitadas
    na questão, se é suficiente. Por exemplo, se a questão
   pode ser respondida com um simples "Não", isso é aceitável
   .

Inclua um raciocínio que justifique a Avaliação. Se os crité
   rios forem atendidos, retorne 'CORRETO'. Se qualquer um
   dos critérios não for atendido, retorne 'ERRADO'.

A Avaliação deve ser um objeto JSON, com as chaves raciocínio
    e resultado.

Exemplos:

1.
### Questão:
Os rendimentos, decorrentes da prestação de serviços de
   consultoria técnica,  pagos por pessoa jurídica
   domiciliada no Brasil a sua matriz no exterior estão
   sujeitos à legislação de preços de transferência?

### Resposta do Especialista:
Em primeiro lugar, há que distinguir se a prestação dos servi
   ços no Brasil implicou transferência de tecnologia. Na hip
   ótese de ficar comprovada a transferência de tecnologia,
   com a anuência do Instituto Nacional de Propriedade
   Industrial (INPI), a transação não estará sujeita às
   regras de preços de transferência consoante o estabelecido
    pelo art. 55 da IN RFB nº 1.312, de 2012. Nessa hipótese,
    a dedução de tais despesas está sujeita aos limites
```

```
   estabelecidos pelos arts. 362 a 365 do RIR/2018. Caso
   inexista transferência de tecnologia, esses serviços
   passam a se submeter às regras de preços de transferência.

### Resposta da IA:
Sim, estão sujeitos, a menos que haja transferência de
   tecnologia com anuência do INPI.

### Avaliação:
{{
  "raciocinio": "A resposta da IA está alinhada com a
     resposta do especialista, abordando corretamente a quest
     ão sem contradições, embora seja mais curta.",
  "resultado": "CORRETO"

}}

2.
### Questão:
O que deve ser considerado como "contraprestações creditadas
   "?

### Resposta do Especialista:
Para efeito do disposto no art. 175 da Instrução Normativa
   RFB nº 1.700, de 2017, consideram-se contraprestações
   creditadas as contraprestações vencidas.

### Resposta da IA:
As contraprestações vencidas.

### Avaliação:
{{
  "raciocinio": "A resposta da IA cobre as principais hipó
     teses mencionadas pelo especialista, sem apresentar
     contradições, embora seja menos detalhada.",
  "resultado": "CORRETO"
}}

3.
### Questão:
```

```
Existe prazo para a compensação de prejuízos fiscais da
   atividade rural?

### Resposta do Especialista:
Não existe qualquer prazo para compensação de prejuízos
   fiscais da atividade rural.

### Resposta da IA:
O prazo é de 7 dias a partir da data do prejuízo, podendo se
   estender até 30 dias para compensação de prejuízos fiscais
    da atividade rural.

### Avaliação:
{{
  "raciocinio": "A resposta da IA está incorreta porque
     menciona '7 dias' e 'até 30 dias', que contradizem a
     resposta do especialista.",
  "resultado": "ERRADO"
}}

4.
### Questão:
A base de cálculo negativa da CSLL poderá ser compensada com
   resultados  apurados em períodos subsequentes?

### Resposta do Especialista:
Sim. A base de cálculo da CSLL, quando negativa, poderá ser
   compensada até o limite de 30% dos resultados apurados em
   períodos subsequentes, ajustados pelas adições e exclusões
    previstas na legislação.

### Resposta da IA:
Não. A base de cálculo da CSLL poderá ser compensada dos
   resultados apurados em períodos subsequentes.

### Avaliação:
{{
  "raciocinio": "A resposta da IA contradiz a resposta do
     especialista, fornecendo uma informação oposta sobre a
     necessidade possibilidade de compensação da CSLL.",
```

```
  "resultado": "ERRADO"
}}

Agora pense passo a passo e faça essa Avaliação:

### Questão:
{questao}

### Resposta do Especialista:
{resposta_especialista}

### Resposta da IA:
{resposta_ia}

### Avaliação:
```