Avaliando Ferramentas de IA Generativa no Conjunto de Perguntas e Respostas da Receita Federal

Erick de Brito[1], **Matheus Teotonio**[1], **Roberto Lotufo**[2], **Jayr Pereira**[1,2]

[1]Universidade Federal do Cariri (UFCA) – Juazeiro do Norte, CE – Brasil

[2]Universidade Estadual de Campinas (UNICAMP) – Campinas, SP – Brasil

{erick.brito , matheus.rennan} @aluno.ufca.edu.br
roberto@neuralmind.ia
jayr.pereira@ufca.edu.br

Abstract. *Artificial Intelligence language models revolutionized the pursuit of digital information, therefore being broadly utilized due to their ease of use and range of topics. However, it is impratical to guarantee that these models maintain an adequate knowledge on unsatisfactorily documented topics, even when such topics are relevant or useful, such as the Brazilian individual income tax return (DIRPF). This study evaluates the coherence of responses provided by these generative tools by submitting to them a set of frequently asked questions on the subject. The results indicate that commercial LLMs are a convenient means of obtaining support related to the DIRPF.*

Resumo. *Os modelos de linguagem de grande escala (LLMs) revolucionaram a aquisição virtual de informações. Os LLMs são amplamente utilizados em diferentes tarefas, dada sua facilidade de manuseio e alcance de tópicos. Todavia, não é possível garantir precisão na aprendizagem dos modelos acerca de questões escassamente documentadas, mesmo em se tratando de problemáticas relevantes ou úteis ao seu conhecimento, tal como a declaração de imposto de renda no Brasil. Este trabalho então avalia a coerência de respostas trazidas por essas ferramentas generativas e para isso lhes foram submetidas as perguntas mais frequentes no tema. Assim, os resultados revelam que os LLMs comerciais são convenientes para combater dúvidas sobre a DIRPF.*

1. Introdução

A Declaração do Imposto sobre a Renda da Pessoa Física (DIRPF) é um procedimento anual obrigatório para milhões de contribuintes no Brasil. Esse processo é essencial para que a Receita Federal do Brasil (RFB) possa calcular e recolher os tributos devidos, além de garantir a conformidade tributária e evitar a sonegação fiscal. A complexidade do sistema tributário brasileiro, com suas frequentes atualizações legais e regras específicas para diferentes tipos de rendimentos, torna a elaboração da DIRPF uma tarefa desafiadora para muitos contribuintes [Cabello and Nakao 2021]. Esse cenário cria uma demanda crescente por ferramentas automatizadas que possam auxiliar na interpretação das normas fiscais, no preenchimento correto das declarações e com soluções para dúvidas frequentes, promovendo assim maior precisão e eficiência no cumprimento das obrigações fiscais.

Como observado por [Maia 2017], o acesso à informação acerca do sistema tributário não se mostra suficiente para alcançar os cidadãos comuns que tem obrigação

de contribuir com seus relatórios fiscais. O descumprimento da declaração ou incoerência nos dados preenchidos, podem prejudicar a regularidade do contribuinte, que na grande parte dos casos não possuía devidas instruções de como fazê-lo. Essas pendências podem ser causadas por atrasos no recebimento, erros de digitação, confusão entre modelos de previdência, omissões de rendimentos sobre atividades secundárias, entre outras. Todos esses detalhes precisam ser considerados para que a Receita Federal não retenha a declaração inadequada para uma investigação mais profunda [Trench 2024], que acarretará em uma multa sujeita a juros até que a situação se regularize.

Na busca pela informação, a inteligência artificial (IA) e sua extensa flexibilidade se apresenta como uma grande aliada. A modelagem de uma IA generativa revela-se capaz de rapidamente recuperar dados acumulados, referentes à sua interpretação da solicitação do usuário, e os viabilizando através do seu processamento de linguagem natural [Coneglian et al. 2024]. As IAs são amplamente utilizadas para ajudar, esclarecer e debater sobre uma grande diversidade de assuntos, pois as grandes empresas que as propiciam investem intensamente no seu treinamento e acervo. Entretanto, ainda podem existir dificuldades na formulação de um retorno quando se tratam de fatos muito recentes, específicos ou pouco documentados. Isso pode ser um problema, pois a IA ainda tentará prever as palavras subsequentes mais prováveis para produzir sua resposta, e portanto quando não houver informação suficiente, praticará o que se conhece como alucinação de uma IA generativa [Maleki et al. 2024]. Em temas tais quais administrativos e jurídicos sabe-se que a veracidade das informações é fundamental, e sendo assim as falhas trazidas nas respostas impactam negativamente na solução do problema [Magalhães and Matos 2025].

Nesse contexto, este trabalho tem o objetivo de avaliar a qualidade de respostas obtidas ao fazer buscas em algumas ferramentas comerciais de IA generativa de uso genérico, tais como ChatGPT, Perplexity e outras. Essas ferramentas são amplamente conhecidas e utilizados pela população em geral. As entradas que lhes foram enviadas tratavam-se de algumas perguntas frequentes no tema, as quais foram selecionadas pela própria Receita Federal. Sendo assim, este trabalho pretende responder à seguinte pergunta de pesquisa: Qual o desempenho das ferramentas comerciais de IA generativa de uso genérico em responder as Perguntas do conjunto de Perguntas e Respostas da RF?

2. Metodologia

Nesta seção, descreve-se como o projeto foi realizado, desde a seleção dos modelos até a comparação entre seus resultados. Este trabalho é uma das etapas iniciais de um trabalho maior focado em elucidar as vantagens do uso e estruturação de uma IA generativa no escopo de questões legais e socioeconômicas imprecisas ao conhecimento popular. O processo passou por diversas etapas: (i) Coleta de dados referentes às perguntas e respostas disponibilizadas pela RFB; (ii) coleta de novas respostas produzidas ao inserir as perguntas mencionadas em ferramentas de IA generativa comerciais; e (iii) avaliar essas novas respostas com base nas referências originais de respostas.

2.1. O Dataset

Como base de dados, foi usado o recurso construído por [Júnior et al. 2025]. A base de dados de perguntas e respostas foi adquirida diretamente por ofícios publicados pelo Sistema Digital[1] da RFB e facultados pelo Conselho Administrativo de Recursos Fiscais

[1]https://www.gov.br/receitafederal/pt-br/centrais-de-conteudo/publicacoes/perguntas-e-respostas/dirpf

(CARF). No total, o *dataset* conta com 715 perguntas e respostas, que foram arranjadas e complementadas com as leis e normativas às quais faziam citação, para criar um *dataset* preenchido com as perguntas e quaisquer informações adicionais referentes a cada uma. Esse resultado foi possível com o amparo de ferramentas automáticas capazes de apanhar textos em PDFs e *scripts* responsáveis por extrair os documentos que sustentavam as respostas destacadas, além da verificação manual para garantir a corretude dos processos.

2.2. Coleta de Respostas

Com a base de dados definida, seguidamente foi realizada a coleta de novas respostas, armazenando de forma ordenada em planilhas no Google Planilhas[2]. Para isso, os pesquisadores que conduziram este trabalho se dividiram para o preenchimento e coleta das respostas. Inicialmente, foram designadas quatro ferramentas de IA generativa de uso genérico para comparar suas respostas geradas com as originais, porém uma foi descartada. As restantes, onde foi possível contornar o limite de prompts são:

- **ChatGPT:** Utilizando o GPT-4o e GPT-4o-mini como modelos base e também a funcionalidade de busca na internet, foram geradas respostas às perguntas 1-150.
- **Perplexity.ai:** Utilizando o modelo *Sonar*[3] com a função *DeepResearch*.
- **Grok:** Utilizando sua versão 3 e com apoio do *DeepResearch*.

2.3. Avaliação

Ainda se tratando do recolhimento das respostas, é necessário comparar a corretude de cada uma dessas implicações com o texto fundamentado na resolução oficial da respectiva pergunta. As respostas coletadas das ferramentas de IA generativa foram avaliadas por meio de um conjunto de métricas da ferramenta RAGAS detalhadas em [Es et al. 2024], amplamente utilizadas para medir a qualidade das gerações. A relevância da resposta (*Response Relevancy*) foi usada para verificar se os modelos abordaram diretamente o tema proposto pelas perguntas, sem desviar ou omitir aspectos centrais. Já a correção factual (*Factual Correctness*) avaliou minuciosamente com base na legislação e nas diretrizes da RFB, observando-se a frequência com que os modelos forneceram informações juridicamente corretas e atualizadas.

Para reforçar a análise, também foram aplicadas métricas de similaridade e fidelidade textual. A Similaridade Semântica (*Semantic Similarity*) foi utilizada para quantificar o grau de alinhamento entre as respostas dos modelos e as respostas oficiais da RFB, considerando o significado geral dos textos. As métricas *BLEU Score*[Papineni et al. 2002] e *ROUGE Score*[Lin 2004], comumente usadas em tarefas de tradução e resumo automático, foram aplicadas para avaliar a sobreposição e cobertura de fragmentos de texto entre as respostas geradas e as referências oficiais. Cabe destacar que a ferramenta RAGAS foi originalmente desenvolvida para a língua inglesa, utilizando modelos de *embeddings* e avaliadores, por exemplo o `gpt-4o-mini`, ajustados para esse idioma. No presente estudo, as métricas foram aplicadas diretamente ao português, sem adaptação específica, o que pode impactar a acurácia das avaliações. Esses indicadores permitiam observar não apenas a precisão literal das respostas, mas também a proximidade, vocabulário e estrutura informacional.

[2]http://bit.ly/46TbkqS

[3]https://www.perplexity.ai/hub/blog/meet-new-sonar

Table 1. Representação numérica da pontuação das IAs generativas de uso genérico abordadas no estudo em cada métrica descrita

Method	Response Relevancy	Factual Correctness	Semantic Similarity	BLEU	ROUGE-L
ChatGPT + Search tool	0.738	0.389	0.793	0.158	0.251
Perplexity.ai + DeepResearch	0.665	0.469	0.757	0.075	0.106
Grok 3 + DeepSearch	0.509	0.454	0.745	0.099	0.089

3. Resultados

A Tabela 1 apresenta os resultados obtidos pelas ferramentas de IA generativa avaliadas neste estudo. As métricas de *Response Relevancy*, *Factual Correctness* e *Semantic Similarity* foram calculadas com base nas respostas geradas pelas ferramentas em comparação com as respostas oficiais da Receita Federal do Brasil. As pontuações variam de 0 a 1, onde valores mais altos indicam melhor desempenho.

Os resultados demonstram que o ChatGPT, com suporte da ferramenta de busca, obteve desempenho superior em todas as métricas avaliadas. Seu destaque em *Response Relevancy* (0.738) e *Semantic Similarity* (0.793) indica uma alta capacidade de produzir respostas alinhadas às perguntas e semanticamente próximas das resoluções oficiais, evitando redundâncias, incompletude e desvio da informação original. No entanto, a métrica de *Factual Correctness* (0.389), que descreve a capacidade de implicação e qualidade da interseção entre resposta e o *ground truth*, evidencia uma fragilidade na precisão de suas respostas, o que pode comprometer sua utilidade em contextos sensíveis como o fiscal.

O Perplexity.ai apresentou a melhor pontuação em *Factual Correctness* (0.469), sugerindo que seu recurso de pesquisa contribui positivamente para a precisão das respostas, ainda que sua coerência semântica e relevância geral sejam inferiores às do ChatGPT. Por fim, o Grok 3 apresentou desempenho inferior nas cinco métricas, indicando maior limitação para uso no contexto avaliado. Apesar das limitações das métricas *BLEU* e *ROUGE-L*, por trabalharem uma análise mais técnica de conjuntos de caracteres, ainda são valiosas para comparar correspondência e estruturação nos modelos.

4. Conclusões

Este estudo buscou responder à seguinte pergunta de pesquisa: “Qual o desempenho das ferramentas comerciais de IA generativa de uso genérico em responder as Perguntas do conjunto de Perguntas e Respostas da Receita Federal?”. A partir dos resultados obtidos, conclui-se que o ChatGPT, com funcionalidade de busca, apresentou o melhor desempenho geral, embora nenhuma das ferramentas tenha atingido níveis satisfatórios em todas as métricas.

Dado o papel central da DIRPF no cumprimento das obrigações fiscais no Brasil, o uso dessas ferramentas pode representar um avanço na inclusão digital e na acessibilidade à informação tributária. No entanto, é imprescindível que os usuários estejam cientes das limitações dessas ferramentas, evitando confiar exclusivamente em suas respostas sem validação adicional.

References

Cabello, O. G. and Nakao, S. H. (2021). Complexidade, conformidade e arrecadação tributária. *Economia e Sociedade*, 30(3):1033–1050.

Coneglian, C. S., Torino, E., Segundo, J. E. S., and Vidotti, S. A. B. G. (2024). Inteligência artificial generativa e recuperação da informação: Tendências e oportunidades de pesquisa. In *XXIII ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO*.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In Aletras, N. and De Clercq, O., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Júnior, J. D., Faria, A., de Oliveira, E. S., de Brito, E., Teotonio, M., Assumpção, A., Carmo, D., Lotufo, R., and Pereira, J. (2025). Br-taxqa-r: A dataset for question answering with references for brazilian personal income tax law, including case law. *arXiv preprint arXiv:2505.15916*.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Magalhães, B. and Matos, F. (2025). Falsas verdades. o impacto das alucinações de ia nos processos judiciais administrativos. *Revista Eletrônica de Direito Processual*, 26(2).

Maia, A. S. (2017). Declaração de imposto de renda de pessoas físicas: principais dificuldades dos contribuintes.

Maleki, N., Padmanabhan, B., and Dutta, K. (2024). Ai hallucinations: a misnomer worth clarifying. In *2024 IEEE conference on artificial intelligence (CAI)*, pages 133–138. IEEE.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). 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.

Trench, E. E. (2024). Raça, gênero e outros atributos do contribuinte e probabilidade de cair na "malha fina" da receita federal.