1. IDENTIFICAÇÃO

Título: Survey para comparar a atuação de dois modelos de LLM na correção de artigos científicos

Área Técnica: Engenharia de Software (ES).

Autor: Bruno Dantas e Larissa Galeno.

Afiliação: COPPE/UFRJ.

Local: Rio de Janeiro.

Data: 11 de Setembro de 2023.

2. CARACTERIZAÇÃO

Estudo aplicado para identificar qual LLM apresenta um resultado melhor perante a opinião qualitativa de usuários.

Tipo: Será um quasi-experimento acompanhado de um questionário de *follow up*.

Domínio: Engenharia de Dados e Inteligência Artificial

Língua: Português (Brasileiro).

Parceiros: não se aplica

Expectativa de Execução: Setembro/2023.
Número Estimado de Repetições: 6.
Glossário de Termos:

- Large Language Model (LLM)
- Application Programming Interface (API)
- RevAIsor: aplicação construída usando APIs de LLMs diferentes que tem como foco revisar o texto acadêmico

1. INTRODUÇÃO

A utilização de Modelos de Linguagem com Aprendizado Profundo (LLMs, do inglês "Large Language Models") tem se tornado uma prática crescentemente comum em diversas aplicações de processamento de linguagem natural (NLP) e inteligência artificial (IA). Esses modelos, como o GPT-3 (Generative Pre-trained Transformer 3) da OpenAI, têm demonstrado um impressionante poder em tarefas como geração de texto, tradução automática e resumo de documentos, entre outras.

No contexto específico de revisão de artigos científicos, onde a qualidade e clareza da linguagem são essenciais, a aplicação de LLMs para aprimorar a redação e identificar erros gramaticais ganha relevância. A busca por modelos capazes de aperfeiçoar a comunicação científica tem implicações significativas para a eficácia da divulgação de pesquisas e para a qualidade do conhecimento científico em geral.

Este estudo tem como objetivo realizar uma avaliação comparativa de dois modelos de LLMs amplamente reconhecidos, o GPT-3 e o BERT (Bidirectional Encoder Representations from Transformers), em termos de sua capacidade de aprimorar a revisão de artigos científicos. O GPT-3 é conhecido por sua capacidade de gerar texto coerente e fluido, enquanto o BERT é elogiado por seu desempenho em tarefas de entendimento de linguagem natural. Ambos os modelos representam abordagens de vanguarda no campo da IA e têm potencial para melhorar substancialmente o processo de revisão de artigos científicos.

A pesquisa se baseia na hipótese de que um desses modelos pode se destacar na tarefa de revisão de artigos, oferecendo sugestões mais precisas e relevantes para melhorar a qualidade do texto acadêmico. A compreensão do desempenho relativo desses modelos é de grande importância para pesquisadores, escritores acadêmicos e profissionais da área de publicação científica.

Além disso, este estudo busca também entender as percepções dos usuários em relação às sugestões de revisão oferecidas por esses modelos. A perspectiva do usuário desempenha um papel fundamental na avaliação da utilidade prática dessas ferramentas em contextos reais de escrita científica.

Assim, a pesquisa visa não apenas contribuir para o avanço da IA e da NLP, mas também fornecer insights valiosos para a melhoria contínua da qualidade da comunicação científica. Compreender como esses modelos de LLMs podem ser aplicados de forma eficaz na revisão de artigos científicos pode ter um impacto duradouro na forma como a pesquisa é apresentada e compreendida pela comunidade científica e pelo público em geral.

4. DEFINIÇÃO DO ESTUDO EXPERIMENTAL

- **Objeto de Estudo:** RevAIsor
- **Objetivo Global:** Este trabalho tem o propósito de realizar um questionário para coletar evidências de qual modelo de LLM conseguiu entregar melhores resultados para o usuário no contexto de revisão de artigo científico.
- **Objetivos Específicos**
  - **Analisar**: a ferramenta *RevAIsor* com LLAMA-2 e GPT-3.5
  - **Com o propósito de**: comparar
  - **Em relação**: à qualidade
  - **Do ponto de vista**: do usuário
  - **No Contexto**: de escrita científica
- **Contexto**
  - O contexto desta pesquisa se encontra no ambiente de escrita científica. Este estudo consiste em comparar dois modelos de LLM a fim de compreender, na perspectiva da opinião do usuário, qual modelo oferece melhores sugestões de revisões para os artigos científicos entrados na ferramenta.

- **Questões e Métricas**
  - ***Qual modelo de LLM (*GPT-3.5 *ou LLAMA2) possui uma melhor resposta no ponto de vista do usuário, com relação a revisão de artigo científico?***
    - Esta questão tem como objetivo identificar qual modelo de LLM apresentou uma melhor resposta na perspectiva do usuário, tendo em vista o contexto da ferramenta RevAIsor supracitado.
      - Métricas Qualitativas: opinião do usuário
- **Questões que não podem ser respondidas neste estudo**
  - Qual modelo de LLM é mais eficiente?
  - Qual modelo de LLM é menos custoso?
  - Como escrever melhores artigos científicos?

5. PLANEJAMENTO

- **Formulação de Hipóteses**
  - **H0**: Não há diferença entre o modelo de LLM GPT-3.5 e o modelo de LLM LLAMA2 para a atividade de revisão de artigo.
  - **H1**: Há diferença entre o modelo de LLM GPT-3.5 e o modelo de LLM LLAMA2 para a atividade de revisão de artigo.

- **Seleção de Variáveis**
  - **Dependentes**: "melhor" modelo de LLM
  - **Independentes**: Modelos de LLM, ordem de uso dos modelos, experiência dos participantes

- **Seleção dos Participantes**
  - Critério de Seleção de Participantes:
    - Alunos de Mestrado ou Doutorado
  - Critério de Seleção de Grupos:
    - Ordem de uso do Modelo de LLM
      - Os participantes serão alocados nos grupos (A ou B) de forma aleatória
      - Grupo A: usa primeiro GPT-3.5 e depois LLAMA2
      - Grupo B: usa primeiro LLAMA2 e depois GPT-3.5
  - Técnicas Probabilísticas de Amostragem: não se aplica
  - Técnicas não Probabilísticas de Amostragem: por conveniência.

- **Recursos**
  - Software: aplicação revAIsor (GPT-3.5 e LLAMA2), Google Forms e E-mail (disparo do questionário)
  - Hardware: Um computador com acesso a internet e navegador instalado.
  - Questionários: somente um com as seguintes seções
    - Termo de consentimento
    - Caracterização do Participante
    - Experimento de fato com os textos dos modelos para avaliar

- Parte do artigo a ser apresentado no questionário para os participantes

- **Design do Experimento**
  - Objetos: revAIsor.
  - Medições: opinião do usuário
  - Técnicas: modelos de LLM: (i) GPT-3.5 e (ii) LLAMA2.

- **Instrumentação**
  - Descrição da Instrumentação: serão coletados dados qualitativos a partir de um questionário em que será apresentado o texto de entrada para o RevAIsor e as correções sugeridas
  - Apoio à Análise Quantitativa: não se aplica;
  - Apoio à Análise Qualitativa: Google Planilhas;
  - Critérios de Observação:
    - Dados registrados no questionário
  - Artefatos (Questionários, Procedimentos, etc):
    - Roteiro para análise dos dados

- **Mecanismos de Análise**
  - Critérios para Eliminação de *Outliers:* não se aplica;

- **Validade dos Resultados**
  - **Validade Interna:** para maximizar a validade interna do estudo será alterada a ordem em que os modelos são apresentados para os participantes, para não ter um viés em relação a ordem. Além disso, será necessário certificar de que as experiências dos participantes são comparáveis, por conta disso será necessário um formulário de caracterização.
  - **Validade Externa:** a validade externa pode ser afetada pela limitação de quantidade de participantes do estudo voluntário e, também, pelo perfil pouco diversificado (alunos de mestrado e doutorado da área de computação). Dessa forma, é possível impactar a generalização do resultado, porém é possível coletar indícios acerca de qual modelo de LLM apresenta melhores resultados para os usuários.
  - **Validade de Conclusão:** a limitação da amostra, novamente, pode afetar a validade de conclusão como, também, o fato do estudo lidar com dados qualitativos somente. Mesmo utilizando as análises recomendadas, o uso de dados qualitativos pode apresentar uma ameaça à validade de conclusão.
  - **Validade de Constructo:** novamente, a natureza do experimento de usar dados qualitativos pode apresentar uma ameaça à validade do constructo. Outro aspecto é a tendência do participante de preferir um dos modelos (por exemplo, o GPT-3.5 por ser mais famoso), com isso os modelos serão anonimizados.

6. TREINAMENTO

- Definição do Treinamento e Procedimentos:
    - No questionário, ao apresentar o texto de entrada, será perguntado para os participantes a opinião acerca do texto para confirmar que fizeram a leitura e estão prontos para avaliar as observações feitas pelo RevAIsor

7. PROCEDIMENTOS DE EXECUÇÃO

- Definição de Execução do Estudo Experimental:
    - Preparar material para execução da sessão:
        - Artigo
    - Selecionar 6 voluntários para participar do estudo
    - Enviar o questionário por e-mail
    - Após o tempo limite sumarizar as respostas dos questionários em uma planilha para realizar o processo de codificação dos comentários
    - Realizar análise dos dados levando em conta os dados do questionário
    - Concluir estudo: refutar hipótese nula e disponibilizar o pacote experimental
- Artefatos (Instruções, Documentos, etc):
    - Template do e-mail a ser enviado
    - Termo de consentimento
    - Plataforma RevAIsor
    - Questionário

8. AVALIAÇÃO DO PLANO

- Objetivos:
    - verificar se a instrumentação é suficiente;
    - verificar se as perguntas do questionário estão claras e são suficientes;
- Participantes: 1 mestrando;
- Procedimentos de Execução:
    - Convidar o participante
    - Seguir as mesmas diretrizes mencionadas no plano
    - Validar se o planejamento do estudo é suficiente;
- Artefatos Utilizados: roteiro de tarrafas, vídeo tutorial, ficha para anotação e questionário online;
- Artefatos Gerados (Lições Aprendidas, Sugestões de Modificação do Plano):
    - Lições Aprendidas:
        - O plano de estudo apresenta uma estrutura clara e bem definida, o que é fundamental para a condução de uma pesquisa rigorosa.

        - A formulação das hipóteses é apropriada e alinhada com os objetivos do estudo.
        - A descrição dos procedimentos de recrutamento dos participantes, coleta de dados e análise está detalhada e bem organizada.
        - As ameaças à validade são consideradas, destacando a importância da transparência na pesquisa.
    - Sugestões de Modificação do Plano:
        - Para aumentar a validade externa, seria interessante explorar a possibilidade de recrutar participantes de diferentes instituições acadêmicas ou áreas de estudo, se possível.
        - Considerar a inclusão de uma seção no questionário para coletar informações demográficas dos participantes, o que pode ser útil para análises posteriores.

9. PLANEJAMENTO DE CUSTOS

- Custos do Estudo Experimental
    - Custos de Planejamento
        - Plano em si: não se aplica;
        - Instrumentação: não se aplica;
        - Material de Treinamento: não se aplica;
        - Avaliação do Plano: não se aplica;
    - Custos de Execução
        - Deslocamentos: não se aplica;
        - Treinamentos: não se aplica;
        - Recursos Humanos: não se aplica;
        - Recursos Materiais:computadores, software.
    - Custos de Análise: não se aplica;
    - Custos de Empacotamento: não se aplica;

10. REFERÊNCIAS BIBLIOGRÁFICAS

- Brown, T. B., et al. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Bidirectional encoder representations from transformers. arXiv preprint arXiv:1810.04805.

11. ANEXO

- **Modelo de e-mail de convite para o estudo**

  Assunto: Convite para Participar de um Estudo Importante sobre Revisão de Artigos Científicos

  Prezado [Nome do Participante Potencial],

Espero que esta mensagem o encontre bem. Gostaríamos de convidá-lo a participar de um estudo crucial que está sendo realizado pela equipe de pesquisa da COPPE/UFRJ na área de Engenharia de Software.

Este estudo tem como objetivo avaliar e comparar diferentes modelos de linguagem natural em um contexto muito relevante: a revisão de artigos científicos. A sua participação é fundamental para o sucesso deste projeto e para o avanço da pesquisa na área.

Por favor, leve alguns minutos do seu tempo para preencher um formulário online que contém algumas questões sobre sua experiência e opiniões em relação à revisão de artigos científicos. O formulário é de preenchimento rápido e direto, e não requer conhecimento prévio sobre os modelos que estamos avaliando. Seu feedback é extremamente valioso para nós.

Link para o Formulário: https://forms.gle/aRC4Lif8qqKBrwBY8

Data Limite para Resposta: 23/09/2023 às 23:59

O tempo estimado para preencher o formulário é de aproximadamente 10 a 20 minutos. Sabemos que seu tempo é precioso, e agradecemos antecipadamente por sua contribuição.

Os resultados deste estudo podem ter um impacto significativo no desenvolvimento de ferramentas de revisão de artigos científicos, beneficiando a comunidade acadêmica e científica como um todo. Sua participação ajudará a fornecer informações valiosas para a pesquisa e o desenvolvimento futuro nessa área.

Lembramos que todas as informações fornecidas serão tratadas com a mais alta confidencialidade e anonimato. Seu nome e informações pessoais não serão divulgados ou associados às suas respostas.

Agradecemos imensamente por considerar nosso convite e esperamos contar com a sua colaboração. Se você tiver alguma dúvida ou precisar de esclarecimentos adicionais, não hesite em entrar em contato conosco pelos endereços de e-mail bdantas@cos.ufrj.br galeno@cos.ufrj.br.

Mais uma vez, agradecemos por sua participação e contribuição para a pesquisa. Seu envolvimento é de grande importância.

Atenciosamente,
Bruno Dantas de Paiva e Larissa Monteiro da Fonseca Galeno
COPPE/UFRJ