PROGRAMA DE BOLSAS ITAÚ

Edital 2º Semestre de 2024

1. OBJETIVOS

Este edital tem como objetivo estabelecer as normas que regem o processo seletivo de bolsistas de **Iniciação Científica (IC)** submetidos ao **Programa de Bolsas Itaú (PBI)**, destinado ao Departamento de Engenharia de Computação e Sistemas Digitais (PCS), ao Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) da Escola Politécnica da USP, ao Programa de Pós-graduação em Ciência da Computação do IME-USP (PPGCC) e ao Programa de Pós-graduação em Sistemas de Informação (PPgSI) da EACH-USP. O PBI é vinculado ao Centro de Ciência de Dados (C^2D) e tem como intuito alavancar o desenvolvimento de pesquisas de ponta em Ciência de Dados e Inteligência Artificial nos níveis de Graduação e de Pós-Graduação, contribuindo com a capacitação de recursos humanos nestas áreas e fomentando o ecossistema em ciência de dados no país.

2. A QUEM SE DESTINA

Poderão participar do processo seletivo de Bolsas de IC do PBI **discentes de graduação da USP**, em fase de Iniciação Científica (IC) e orientadas/os por docentes do PCS, PPGEE, PPGCC e PPgSI.

3. MODALIDADES E VALORES DE BOLSAS

O PBI oferece diferentes modalidades de bolsas para apoio às pesquisas em Ciência de Dados e Inteligência Artificial vinculadas ao C^2D. As bolsas têm duração máxima de 12 meses e poderão ser renovadas caso seja solicitada a extensão do projeto no processo seletivo seguinte, e este seja novamente aprovado para receber apoio do PBI. As bolsas do PBI são distribuídas nas seguintes modalidades:

3.1. **Bolsa de Iniciação Científica (IC)**, no valor mensal líquido de **R$ 815,00 (oitocentos e quinze reais)**, por um período de até 12 meses ou até a conclusão do curso, o que for menor.

3.2. As bolsas do PBI são destinadas a bolsistas em regime de **dedicação exclusiva e integral** aos estudos e à pesquisa, não sendo permitido o acúmulo do benefício com outra atividade remunerada ou bolsa de pesquisa, com exceção das bolsas de caráter de auxílio social.

Para usufruir da bolsa a/o beneficiária/o deve, obrigatoriamente, exercer suas atividades nas dependências da USP e, preferencialmente, no C^2D.

3.3. Bolsistas de IC deverão dedicar ao menos 12h semanais ao seu projeto de pesquisa.

4. ÁREAS DE INTERESSE E QUANTIDADE DE BOLSAS

4.1. São oferecidas neste edital **até 3 (três) bolsas** de IC.

4.2. As bolsas serão determinadas para projetos de pesquisa com os seguintes temas de pesquisa:

a) Avaliação Psicométrica: Abordagem de psicologia por meio de conceitos de avaliação psicométrica para identificação de habilidades cognitivas.
b) Cybersecurity: Estudo de técnicas de defesa contra prompt injection e jailbreaking prompts no contexto de Red-Teaming.
c) Criação de Modelo: Fine-tuning de LLMs *open source* para abordagem de prompt ensemble em conteúdo ético.

O detalhamento de cada projeto é apresentado no ANEXO IV deste edital.

5. CRITÉRIOS PARA ANÁLISE

Somente serão analisadas candidaturas de alunos para os projetos de pesquisa que atendam aos seguintes requisitos:

5.1. Toda a pesquisa realizada no âmbito do projeto proposto deve utilizar dados abertos, dessa forma intensificando a troca de experiências com a comunidade científica e promovendo a divulgação dos resultados em artigos científicos publicados em anais de congressos e em periódicos especializados, com particular ênfase nas publicações com visibilidade internacional.

5.2. Cada aluna/o será selecionado para apenas 1 (um) projeto de pesquisa, porém podem se inscrever para mais de um projeto.

5.3. As candidaturas devem seguir as especificações do ANEXO I.

6. CRITÉRIOS PARA SELEÇÃO

As candidaturas serão avaliadas por um processo de avaliação conduzido e aprovado de forma colegiada pelos Comitês de Acompanhamento e Executivo do C²D, compostos por professores da Escola Politécnica da USP e pesquisadores do patrocinador do PBI, de acordo com o fluxo apresentado no ANEXO III deste edital.

6.1. A avaliação das candidaturas será feita pelo curriculum vitae, histórico escolar e acadêmico da/o candidata/o e eventual entrevista, conforme especificado no ANEXO III.

6.2. Membros do Comitê de Acompanhamento e do Comitê Executivo do C²D que propuserem projetos no contexto deste edital não participarão do processo de seleção. Nestes casos, os respectivos comitês do C²D poderão indicar substitutas/os para participar do processo de seleção das propostas.

7. INSCRIÇÕES E RESULTADOS

7.1. O edital de bolsas para ingresso no 2º Semestre de 2024 deve obedecer aos prazos descritos no ANEXO III.

7.2. Cada proposta de projeto de pesquisa deverá ser submetida no formato especificado no item a seguir e no Anexo I, em arquivo **PDF único** para o e-mail **c2d@usp.br** com o assunto **PBI-2024S2 - NOME DA/O CANDIDATA/O**. O e-mail deve ser enviado com cópia para o/a orientador/a.

7.3. Devem acompanhar a submissão os seguintes documentos em um **único arquivo** em formato PDF, na **seguinte ordem**:;

a) Formulário de inscrição preenchido (ver ANEXO I);

b) Histórico escolar de graduação;

c) Curriculum Vitae da/o aluna/o, ressaltando os méritos escolares e acadêmicos (CV LATTES).

7.4. Apenas os projetos aprovados serão comunicados no site http://c2d.poli.usp.br/ até as 23h00 do dia previsto para a divulgação, descrito no ANEXO III.

8. OBRIGAÇÕES

Bolsistas, pesquisadores e professores dos projetos de pesquisa apoiados pelo PBI deverão atender às seguintes obrigações:

8.1. Desenvolver integralmente o projeto, dentro do período de duração da bolsa de pesquisa, executando todas as etapas estabelecidas no cronograma proposto no projeto de pesquisa.

8.2. Desenvolver seus trabalhos nas dependências do C²D, durante todo o período de duração da bolsa de pesquisa e com dedicação conforme especificada em suas bolsas (item 3).

8.3. Entregar, **mensalmente**, um relatório compacto (conforme modelo do ANEXO II) descrevendo as atividades executadas no período, de acordo com o cronograma de atividades proposto, incluindo avaliações do/a orientador/a sobre o andamento do projeto.

8.4. Bolsista e orientador/a deverão realizar reuniões de acompanhamento (não necessariamente presencial) com o grupo de especialistas do patrocinador do PBI que acompanham os projetos de pesquisa desenvolvidos no C²D.

8.5. Bolsista e orientador/a deverão participar de seminários e workshops de pesquisa promovidos pelo C²D no âmbito do PBI e destinados a compartilhar os desenvolvimentos com a comunidade científica.

8.6. Em todas as publicações vinculadas às atividades de pesquisa apoiadas no âmbito do PBI deverá constar uma referência ao patrocinador com a seguinte expressão: "*Este trabalho foi realizado com o apoio do Itaú Unibanco S.A., por meio do Programa de Bolsas Itaú (PBI), vinculado ao Centro de Ciência de Dados da Escola Politécnica da Universidade de São Paulo*" (ou expressão equivalente em língua inglesa ou na língua de redação do texto).

8.7. Em casos de desistência da bolsa por iniciativa da/o aluna/o, a/o mesma/o estará obrigado, por contrato, a devolver o valor integral já recebido, sem correções. Ficam isentos desta obrigação os casos submetidos e aprovados pelo Comitê de Acompanhamento do C²D.

9. DISPOSIÇÕES FINAIS

9.1. Os casos omissos serão resolvidos pelos Comitês de Acompanhamento e Executivo do C²D.

ANEXO I - Submissão da Proposta para Bolsa PBI

FORMULÁRIO DE SUBMISSÃO DE CANDIDATURA - A inscrição deve conter a tabela abaixo preenchida **para as bolsas de IC oferecidas:**

Modalidade de bolsa:	Iniciação Científica (IC)
Candidatura ao projeto (COLOCAR 1, 2 OU 3 DE ACORDO COM A PRIORIDADE DE INTERESSE, SENDO 1 A MAIOR E 3 A MENOR PRIORIDADE; EM BRANCO INDICA SEM INTERESSE) :	( ) Avaliação Psicométrica ( ) Cybersecurity ( ) Criação de Modelo
Nome da/o Aluna/o:	
Depto e Unidade da/o Aluna/o:	
Ano de Ingresso na USP:	
E-mail da/o Aluna/o:	
Link para o CV Lattes da/o Aluna/o:	

ANEXO II - Modelo para Relatórios Mensais

Dados da Bolsa
Tipo de Bolsa: ( ) IC ( ) PqEP ( ) ME
Nome do/a Orientador/a:
Nome do Projeto:
Período da Bolsa: **/ /** a **/ /**
Relatório: ( ) Final ( ) Parcial
Período do Relatório: **/ /** a **/ /**
Descrição das Atividades de Pesquisa do Projeto
1. Descrição das atividades acadêmicas:
2. Descrição das atividades planejadas para o relatório (repetir do relatório anterior):
3. Descrição das atividades de pesquisa realizadas:
4. Descrição das próximas atividades:
Houve alteração significativa no tema ou prazo: ( ) Sim ( ) Não **Justifique em caso positivo:**
Apreciação Circunstanciada do/a Orientador/a sobre as Atividades da/o Bolsista
Etapa cumprida no relatório: ( ) Ótimo ( ) Bom ( ) Regular ( ) Fraco
Programação para a próxima etapa: ( ) Ótimo ( ) Bom ( ) Regular ( ) Fraco
Resultados em relação às expectativas iniciais: ( ) Acima ( ) Dentro ( ) Abaixo ( ) Muito abaixo
Previsão de conclusão no prazo: ( ) Sim ( ) Não **Justifique em caso negativo:**

Apreciação do/a orientador/a:
Protocolo
Data:
Nome Completo da/o Bolsista:

ANEXO III

A. FLUXO DO PROCESSO DE SELEÇÃO DO PBI

O processo de seleção é conduzido e aprovado de forma colegiada pelos Comitês de Acompanhamento e Executivo do C²D, compostos por professores da Escola Politécnica da USP e pesquisadores do patrocinador do PBI. O processo de seleção seguirá o seguinte fluxo:

a) Inscrição das propostas: **13/05/2024 a 23/05/2024**

b) Divulgação das inscrições aceitas: **24/05/2024**

c) Entrevistas com os pré-selecionados para seleção final: **31/05/2024**

d) Divulgação no site http://c2d.poli.usp.br/ das/os aprovadas/os: **03/06/2024**

B. ITENS DO FORMULÁRIO DE AVALIAÇÃO

Os itens de avaliação são:

a) Histórico Escolar da/o candidata/o;

b) Histórico Acadêmico da/o candidata/o (participação em projetos de pesquisa, bolsas anteriores, publicações científicas, premiações, habilidades e competências);

c) Outros itens que componham a descrição das atividades acadêmicas, científicas e profissionais desenvolvidas pela/o candidata/o;

d) Resultado de entrevista, caso seja selecionada/o para tal.

ANEXO IV

Projetos de Pesquisa

Bolsista de avaliação psicométrica

Título: Abordagem de psicologia por meio de conceitos de avaliação psicométrica para identificação de habilidades cognitivas

Orientadora: **Profa. Anarosa Alves Franco Brandão (PCS - EPUSP)**

Resumo

As LLMs têm influenciado abruptamente diversas áreas do conhecimento. E para entendê-la de uma maneira mais macro é necessário que tenhamos uma abordagem multidisciplinar.

Neste contexto se encaixa o bolsista de psicologia, a ideia é usar uma abordagem diferenciada do que se encontra na literatura até então e tratar a LLM como se fosse um paciente de um psicólogo. Por meio de avaliações psicométricas podemos avaliar os modelos por meio de métricas que até então só foram usadas em humanos.

São diversos fatores que o bolsista vai avaliar de um modelo de LLM. São eles: habilidades cognitivas, inteligência geral [1], resolução de problemas; personalidade, traços de personalidade, predisposições padrões de pensamento e comportamento, por meio de situações que podemos submetê-la; avaliação educacional, identificar se há necessidades educacionais especiais, estilos de aprendizagem que melhor se encaixem para tirar o melhor resultado possível, pontos forte e fracos de aprendizado; e por último avaliar role-playing e como isso afeta diretamente todos os fatores supramencionados.

Metodologia

Para colocar em prática essa análise, deverão ser estudados conceitos de avaliação psicométrica, análise de personalidade, análise de inteligência e medidas cognitivas entre outros disponíveis na literatura especializada de psicologia [1, 2].

A abordagem deste trabalho será nestas duas direções a primeira parte será uma abordagem e na segunda a outra, a ordem não importa e quem decidirá isso será o bolsista junto com seu orientador.

Referências

- [1] Intelligence, cognition, and major neurocognitive disorders: From constructs to measures --- Métricas de inteligência e habilidades cognitivas
- [2] Big Five Factor Model, Theory and Structure --- Estudo de personalidade

Bolsista de Cybersecurity

Título: Estudo de técnicas de defesa contra prompt injection e jailbreaking prompts no contexto de Red-Teaming

Orientadora: **Profa. Sarajane Marques Peres (EACH-USP)**

Resumo

Modelos de LLM vêm revolucionando a ciência em diversos campos devido sua capacidade generativa. Cada vez mais estamos integrando esses modelos em aplicações reais no mercado. Mas o que se tem mostrado é que esses modelos apresentam muitas vulnerabilidades e riscos quando se trata de segurança da informação.

Para um uso consciente da IA é necessário que haja salvaguardas capazes de impedi-la de condutas inadequadas que podem comprometer a imagem de um banco. Com isso é importante sempre estar atento às novas tendências de ataque para poder se defender de novas formas que surgem todos os dias.

O acesso a informações privadas, a criação de conteúdo nocivo, conteúdo falso, entre outros são exemplos de como os modelos de IA generativa são vulneráveis.

Esses tipos de ataque estão inclusos em termos como prompt injection e jailbreaking prompts. Prompt Injection refere-se à técnica maliciosa de inserir ou modificar inputs (prompts) para modelos de linguagem com o objetivo de enganar o modelo e fazer com que execute ações não intencionadas ou revele informações que não deveria. Jailbreaking Prompts, por outro lado, é um conceito específico de tentativa de "quebrar" ou contornar as limitações e restrições impostas a um modelo de linguagem. Esses prompts são projetados de maneira inteligente para explorar vulnerabilidades no modelo ou em sua implementação, permitindo ao usuário acessar funcionalidades ou informações que normalmente estariam restritas ou bloqueadas.

Existem bastantes técnicas dentro destes assuntos, e neste contexto entra esse bolsista. Ele estará incumbido de estudar o que foi usado até então na área de prompt injection e jailbreaking prompts, tendo esses ataques em mente ele deve elaborar guardrails mais robustas para cada um dos ataques estudados.

Metodologia

Para alcançar esse objetivo é necessário estudar os métodos existentes. Um trabalho que deve ser usado é [1], essa *survey* passa por métodos muito atuais (visto que ela foi lançada dia 3 de março de 2024) e pode ser um bom norte de referências e conteúdos a se estudar. Todos os artigos que esse trabalho cita deverão ser analisados pelo bolsista para aprofundar em cada tópico, assim como em [2].

Outro trabalho que será usado como base é o [3], este trabalho utiliza um modelo de linguagem para gerar automaticamente casos de teste ("red teaming") com o objetivo de encontrar situações em que uma LM alvo se comporta de maneira prejudicial. E avaliam as respostas da LM a esses testes com a ajuda de um classificador treinado para detectar conteúdo ofensivo, eles exploram várias técnicas, desde a geração em zero-shot até aprendizado por reforço, para produzir casos de teste com diferentes graus de diversidade e dificuldade. E esse método deste paper deve ser usado para avaliação, e saber se a guardrail criada para aquele mecanismo está funcionando adequadamente.

Foi dito um pouco sobre métodos de avaliação e deve seguir nesta linha que fora exposta, criando guardrails específicas para aquela forma específica de ataque que o bolsista estudou das referências e aplicação de ataque em cima dela para avaliar sua robustez e efetividade. Para realizar a tarefa de ataque é importante além do próprio teste do bolsista e do orientador utilizar agente automáticos de ataque. Muitos papers já foram feitos criando abordagens diferenciadas para agentes de ataques e esses são alguns que devem ser utilizados como base [4, 5].

Cronograma

- Primeira parte: Métodos de prompt injection e jailbreaking prompts
- Segunda parte: Agente de ataque
- Terceira parte: Guardrails

Referências

- [1] *Breaking Down the Defenses: A Comparative Survey of Attacks on Large Language Models*
- [2] A survey on Large Language Model (LLM) security and privacy: The Good, The Bad, and The Ugly
- [3] *Red Teaming Language Models with Language Models*
- [4] Jailbreaking Black Box Large Language Models in Twenty Queries
- [5] Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack

Bolsista de criação de modelo

Título: Fine-tuning de LLMs *open source* para abordagem de prompt ensemble em conteúdo ético

Orientador: **Prof. Artur Jordão (PCS - EPUSP)**

Resumo

As LLMs têm cada vez mais nos levado a um modelo agnóstico a tarefa e isso tem movimentado muito a área acadêmica sobre como torná-la mais robusta, e atraído as empresas para utilizar os seus próprios modelos em aplicações do mercado. Mas um problema que encontramos é a variabilidade dentre as respostas dadas pelo modelo, para um mesmo input o modelo para dar inúmeras combinações e respostas diferentes.

É nesse contexto que se insere o bolsista de criação de modelo. A ideia é usar conceitos de *fine-tuning* de modelos open-source aplicados ao escopo de ética, junto com a abordagem de *prompt ensemble*.

O bolsista deverá criar um modelo usando fine-tuning de modelos open-source, existem vários modelos de diferentes versões e o bolsista deve ser capaz de testar e avaliar qual faz melhor a tarefa proposta a fim de torná-lo específico em ética.

Metodologia

A fim de conseguir alcançar essa meta o bolsista deve seguir um cronograma orientado.

Para começar é necessário que o bolsista estude a bibliografia destes assuntos. Começando pelos dois principais benchmarks que serão o paper sobre a criação de múltiplos prompts a fim de performar melhor, uma técnica chama *prompt ensemble* [1].

Paralelamente o outro tópico que o bolsista deve pesquisar usando como referência base o paper sobre fine-tuning de modelos [2]. Esse artigo em específico usa o conceito de *Reinforcement Learning from Human Feedback* (RLHF), mas se o bolsista preferir pode optar por utilizar uma outra abordagem como por exemplo o *Supervised Learning*.

Depois disso o bolsista irá realmente começar a pôr em prática esses conceitos e criar esse novo modelo. Primeiramente ele deve focar no dataset para realizar o treinamento. Um dataset que deve ser usado como base para a construção de modelos é o CyberSecEval 2 da Meta [3]. Esse dataset é o que há de mais recente além de ser perfeitamente encaixável neste projeto pois ele testa o modelo em si, que é diferente de testar um prompt.

O dataset robusto é uma das partes mais importantes para se construir o modelo mais robusto possível, por isso o CyberSecEval deve ser a base, o bolsista vai ficar a encargo junto com seu orientador de aumentar esse dataset levando em consideração o que for surgindo de mais novo no assunto de *prompt injection* e *safeguards*.

Tendo esse dataset o bolsista deve focar sua atenção em fine-tuning de modelos. Usando as máquinas disponíveis do C2D para fazer o treinamento, que dispõem de GPUs potentes capazes de tornar esse processo muito mais rápido. Dessa forma é possível que ele faça o fine-tuning do máximo possível de modelos disponíveis open-source com a intenção de fazer uma análise comparativa no fim para publicar em conferência do assunto.

Depois vem a parte do prompt ensemble, o bolsista deve aplicar essa abordagem dentro do modelo já treinado para avaliar a melhoria e dessa forma conseguir criar o modelo mais robusto, juntando abordagens que já se demonstraram eficazes dentro da literatura especializada.

Por último vem a avaliação do modelo que foi construído. Além das comparações feitas anteriormente entre os modelos open-source que passaram pelo processo de fien-tuning, deve ser usado para comparar resultados o modelo do Meta Llama-Guard [4]. Esse modelo é basicamente o que desejamos que o bolsista seja capaz de superar.

Depois de todo esse material feito que foi explicado anteriormente o bolsista em conjunto com o seu orientador deve escrever um artigo a fim de publicar em alguma conferência de maior impacto.

Cronograma

- Primeira parte: Dataset
- Segunda parte: Fine-tuning
- Terceira parte: Prompt ensemble

Referências

- [1] ASK ME ANYTHING: A SIMPLE STRATEGY FOR PROMPTING LANGUAGE MODELS --- Prompt Ensemble
- [2] Fine-Tuning Language Models from Human Preferences --- Fine-Tuning
- [3] CYBERSECEVAL 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models ----- CyberSecEval 2
- [4] Meta-Lla