UNIVERSIDADE FEDERAL DE GOIÁS / INSTITUTO DE INFORMÁTICA

Pós-treino de LLMs via Aprendizado por Reforço

Aprimoramento de Raciocínio com Ferramentas de Busca Externa

Daniel Machado Pedrozo

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

UNIVERSIDADE FEDERAL DE GOIÁS (UFG)
INSTITUTO DE INFORMÁTICA (INF)

DANIEL MACHADO PEDROZO

Pós-treino de LLMs via Aprendizado por Reforço

Aprimoramento de Raciocínio com Ferramentas de Busca Externa

Goiânia
2025

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 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 dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) 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 Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): DANIEL MACHADO PEDROZO

Título do trabalho: Pós-treino de LLMs via Aprendizado por Reforço

Aprimoramento de Raciocínio com Ferramentas de Busca Externa

2. 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)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. 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.

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

Documento assinado eletronicamente por **Daniel Machado Pedrozo**, **Discente**, em 04/02/2026, às 16:20, 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 **Fernando Marques Federson**, **Professor do Magistério Superior**, em 13/03/2026, às 11:25, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

DANIEL MACHADO PEDROZO

Pós-treino de LLMs via Aprendizado por Reforço

Aprimoramento de Raciocínio com Ferramentas de Busca Externa

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Orientador: Prof. Dr. Fernando Marques Federson

Goiânia

2025

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

PEDROZO, DANIEL MACHADO
Pós-treino de LLMs via Aprendizado por Reforço [manuscrito]: Aprimoramento de Raciocínio com Ferramentas de Busca Externa / DANIEL MACHADO PEDROZO. - 2025.
57 f.: 2025

Orientador: Prof. Dr. Fernando Marques Federson
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Informática (INF), Inteligência Artificial, Goiânia, 2025.

1. Inteligência Artificial. 2. Large Language Models. 3. Aprendizado por Reforço.

I. Federson, Fernando Marques , orient. II. Título.

CDU 004

DANIEL MACHADO PEDROZO

Pós-treino de LLMs via Aprendizado por Reforço
Aprimoramento de Raciocínio com Ferramentas de Busca Externa

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Data da Aprovação: 09 de dezembro de 2025.

Prof. Dr. Fernando Marques Federson
Orientador (INF-UFG)

Prof. Dr. Aldo André Díaz Salazar
Coordenador de TCC do BIA (INF-UFG)

Prof. Dr. Anderson da Silva Soares
Coordenador do BIA (INF-UFG)

Prof. Dr. Leonardo Antônio Alves
(INF-UFG)

DANIEL MACHADO PEDROZO

Pós-treino de LLMs via Aprendizado por Reforço

Aprimoramento de Raciocínio com Ferramentas de Busca Externa

RESUMO

Este Relatório de Conclusão de Curso tem como objetivo reunir os resultados da minha jornada para me tornar um especialista em **Aprendizado por Reforço em LLMs**. Uma ilustração e sua narrativa descrevem os períodos de trabalho. Os Apêndices contêm os Termos de Aceite de Entrega e os resultados obtidos durante cada período de trabalho.

Palavras-chave: Inteligência artificial; Large language models; Aprendizado por reforço.

ABSTRACT

This Course Completion Report aims to bring together the results of my journey to become an expert in **Reinforcement Learning in LLMs**. An illustration and its narrative describe the work periods. The Appendices contain the Delivery Acceptance Terms and the results obtained during each work period.

Keywords: Artificial intelligence; Large language models; Reinforcement learning.

Goiânia

2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

Minha Jornada

Daniel Machado Pedrozo

Especialista em:
Aprendizado por Reforço em LLMs

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

MINHA JORNADA

Nome: **Daniel Machado Pedrozo**

Especialidade: **Aprendizado por Reforço em LLMs**

Objetivo deste documento

Durante o processo da disciplina Residência em IA[1], foram gerados diversos resultados na construção da minha especialização. A cada semana, um conjunto de resultados foi formalizado por um Termo de Aceite de Entrega e avaliado por uma banca, considerando o planejado e o realizado para o período. Este documento tem como objetivo descrever esses resultados obtidos, fazendo referência aos Termos de Aceite de Entrega e seus documentos associados.

Minha Jornada

Minha Jornada começou nas **Semana 1 e 2** com atividades voltadas para definir, de forma clara, a área de conhecimento da minha especialização e me fundamentar nela. Com base nas leituras recomendadas a partir dos meus tópicos prévios de interesse por plataformas de busca acadêmica (como Consensus, ResearchRabbit e ChatGPT), pude compreender a abrangência da aplicação do Aprendizado por Reforço (*Reinforcement Learning* - RL) aos Modelos de Linguagem de Grande Porte *(Large Language Models* - LLMs). É evidente que essa área vem crescendo com grande velocidade, especialmente devido aos avanços recentes em técnicas de otimização de política como PPO, DPO, GRPO, sendo utilizadas no pós-treinamento de LLMs para aprimorar o raciocínio, como no caso do DeepSeek-R1[2]. Observa-se, ainda, que os modelos de Inteligência Artificial, especialmente os LLMs, têm se aproximado de forma significativa da capacidade humana em diversas tarefas, chegando até mesmo a superá-la em alguns casos. Entretanto, torna-se cada vez mais evidente que há um limite para o desempenho que pode ser

[1]Dez Semanas, entre setembro de 2025 e dezembro de 2025.
[2]DeepSeek-AI et al. DeepSeek-R1: *Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv, 2025. Disponível em: https://arxiv.org/abs/2501.12948.

alcançado utilizando-se exclusivamente dados produzidos por humanos. Assim, para avançar em direção a um nível mais próximo de AGI (Artificial General Intelligence), considero essencial que os modelos ultrapassem a dependência de datasets puramente humanos (encontrei respaldo dessa visão nos autores do survey[3] estudado em uma das **Semanas** seguintes). Nesse sentido, o Aprendizado por Reforço surge como um paradigma promissor, pois permite que modelos de IA aprendam por meio de sua própria interação com o ambiente, buscando aprimorar continuamente seu desempenho em tarefas específicas. Além disso, muitos LLMs são implantados em cenários reais e utilizados como assistentes ou agentes que interagem diretamente com seus usuários, realizando as funções para as quais foram desenvolvidos. Em tais contextos, é de grande interesse que esses modelos não apenas mantenham seu desempenho, mas continuem a melhorar ao longo do tempo. Portanto, por essas razões, ficou claro para mim que gostaria de me tornar um especialista em **Aprendizado de Máquina por Reforço aplicado à Large Language Models.** No **Apêndice 1**, separei uma série de fontes de conteúdo que seriam necessárias para que eu pudesse me aprofundar nessa área. Notei que devido à grande abrangência da área, seria necessário separar em 2 grupos as fontes obtidas. O primeiro, "**Fundamentos do Aprendizado por Reforço**", para garantir domínio nas principais técnicas dessa área e como elas se organizam. O segundo, "**Fundamentos em Large Language Models**", para compreender como o Processamento de Linguagem Natural evoluiu até os modelos dotados de grande capacidade de raciocínio e como o RL é normalmente aplicado nesse contexto. Além disso, estudei através de vídeos, como *The FASTEST Introduction to RL on the Internet*[4], *Reinforcement Learning: A 60 Year History*[5], e o documentário *AlphaGo*[6], que ajudaram a situar historicamente a evolução do RL e sua relevância atual. Ao mesmo tempo em que construía essa base teórica, iniciei a criação dos meus próprios instrumentos de organização: uma tabela detalhada de algoritmos de RL com suas principais características e um glossário com conceitos fundamentais. Este material, no **Apêndice 1**, foi continuamente revisitado e expandido ao longo das **Semanas** seguintes. Também realizei a

[3] Zhang, K. et al. *A Survey of Reinforcement Learning for Large Reasoning Models*. arXiv, 2025. Disponível em: https://arxiv.org/abs/2509.08827.

[4] The FASTEST introduction to Reinforcement Learning on the internet

[5] How Machines Learn to Act

[6] AlphaGo - The Movie | Full award-winning documentary

leitura de um importante survey[7] sobre algoritmos de RL, que serviu para compreensão dos diferentes tipos de ambientes que cada um pode ser usado, e a evolução deles até os mais atuais, como PPO, um dos principais algoritmos para otimização de política de LLMs utilizados atualmente utilizado como inspiração para outros.

A partir dessas bases, a **Semana 3** foi marcada pelo início de uma revisão bibliográfica mais profunda. Li e sistematizei artigo através da tabela de algoritmos, e elaborei um mapa mental que classifica, de forma visual, todos os algoritmos estudados. Os resultados desta **Semana** estão reunidos no **Apêndice 2.**

As **Semanas 4, 5 e 6** representaram um avanço importante na minha compreensão sobre os métodos de raciocínio usados por modelos contemporâneos. Estudei com profundidade o DeepSeek-R1, seu *pipeline* completo e suas variações (como o R1-Zero), bem como algoritmos que sustentam seu treinamento, como o GRPO. Também iniciei o estudo do Search-R1[8], que mais à frente se tornaria o centro da minha especialização. Além disso, mergulhei em conteúdos relevantes como a palestra de Dale Schuurmans[9] que relaciona LLMs com Máquinas de Turing, e o Chain-of-Thought (CoT), incluindo o custo computacional. Essa palestra me fez compreender o quão importante é o *reasoning* para a performance dos LLMs. O RL é o grande responsável pela melhoria dessa capacidade de raciocínio, logo, percebi que gostaria de trabalhar com isso de alguma forma. Realizei também a leitura inicial do *A Survey of Reinforcement Learning for Large Reasoning Models* [10], um material extenso que sintetiza muito bem a minha área de estudo e me ajudou a compreender as subáreas atuais, as técnicas e os frameworks que ainda eram desconhecidos para mim. Adicionalmente, aprofundei-me em áreas emergentes que surgiam no survey: tool reasoning, multi-agent systems (MAS-LLMs), visão e linguagem

[7] AlMahamid, F.; Grolinger, K. *Reinforcement Learning Algorithms: An Overview and Classification*. In: 2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE, 2021, p. 1–7. DOI: 10.1109/CCECE53047.2021.9569056. Disponível em: http://dx.doi.org/10.1109/CCECE53047.2021.9569056.

[8] Jin, B. et al. *Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning*. arXiv, 2025. Disponível em: https://arxiv.org/abs/2503.09516.

[9] Schuurmans, Dale. *Language Models and Computation*. YouTube, uploaded por Dale Schuurmans, 2023, https://www.youtube.com/watch?v=hMEViRcF7o0.

[10] Zhang, K. et al. *A Survey of Reinforcement Learning for Large Reasoning Models*. arXiv, 2025. Disponível em: https://arxiv.org/abs/2509.08827.

(VLMs) e aplicações biomédicas. Também estudei algoritmos de alinhamento de LLMs, como DPO e PPO, esclarecendo dúvidas que ainda me acompanhavam desde o início. Esta fase foi essencial para que eu pudesse definir, de forma madura, o foco da minha especialização. Além disso, esta foi a etapa em que avaliei diferentes frameworks de RL para LLMs. Os frameworks estudados podem ser encontrados no **Apêndice 3**

A **Semana 7** foi especialmente significativa. Estudei três artigos que influenciaram diretamente a continuidade da minha jornada — MedGemma[11], Search-R1 (de forma mais "profunda") e ChestX-Reasoner[12]. Durante esse período, ficou evidente que eu gostaria de trabalhar com o tema de "*raciocínio com busca"*, uma linha que combina RL, *reasoning* e ferramentas externas. Assim, tomei a decisão de tentar reproduzir o Search-R1 e, posteriormente, propor adaptações para alinhar um modelo de domínio específico, como o MedGemma. Os trabalhos que mais me interessaram e seus respectivos resumos podem ser encontrados no **Apêndice 4.**

Foi, então, nas **Semanas 8 e 9**, que iniciei a fase experimental da minha jornada. Preparei o ambiente de desenvolvimento, clonei repositórios, realizei vetorização da base textual e configurei os servidores necessários. No entanto, logo enfrentei obstáculos técnicos, em especial, que o repositório original do Search-R1 estava depreciado e incompatível com as bibliotecas mais recentes. As dificuldades envolveram falhas de dependências, versões conflitantes de CUDA, problemas no vLLM e ajustes nas interfaces das ferramentas. Por outro lado, essa fase foi determinante para o meu crescimento. A partir desses desafios, migrei para a versão atual do framework VeRL**.** Utilizei uma imagem Docker oficial, adaptei datasets e iniciei efetivamente meus primeiros treinamentos com os modelos Qwen2.5-3B e Qwen3-0.6B. No **Apêndice 5** pode-se visualizar os modelos base utilizados para o treinamento e os frameworks utilizados.

A **Semana 10** marcou o ponto mais alto da jornada: consegui reproduzir o Search-R1 com sucesso, obtendo um modelo capaz de raciocinar, decidir quando realizar buscas e

[11] Sellergren, A. et al. *MedGemma Technical Report.* arXiv, 2025. Disponível em: https://arxiv.org/abs/2507.05201.

[12] Fan, Z. et al. *ChestX-Reasoner: Advancing Radiology Foundation Models with Reasoning through Step-by-Step Verification.* arXiv, 2025. Disponível em: https://arxiv.org/abs/2504.20930.

recuperar informações com base em evidência. Testei também o uso do *MedGemma* no mesmo *pipeline*, mas a execução foi impossibilitada por cauda da multimodalidade do modelo original. A partir disso, formulei uma pergunta de pesquisa: será que modelos pequenos conseguem aprender tool calling? Para respondê-la, realizei experimentos com os modelos Qwen3 de 0.6B e 1.7B, utilizei um reward model especializado (Skywork Reward Model) e combinei recompensas verificáveis com avaliação do LLM Judge. Os resultados foram promissores, indicando que, sim, modelos pequenos podem apresentar desempenho competitivo em tarefas de raciocínio com apoio de ferramentas. Os resultados finais podem ser visualizados no **Apêndice 6.**

Em função de tudo que vivi nesta jornada, gostaria de deixar registrado que estas dez **Semanas** representaram um processo de profunda transformação. Pude compreender, na prática, como conceitos abstratos de RL se conectam ao comportamento real de um modelo de linguagem; enfrentei desafios técnicos que exigiram resiliência; adquiri mais familiaridade com a linguagem científica e ferramentas atuais; consegui reproduzir um *pipeline* de ponta; e explanei uma linha de investigação que me acompanha até hoje. Mais do que entender as “letrinhas do RL”, compreendi a lógica e a essência de como LLMs podem aprender a “pensar”, “agir” e buscar informações de maneira eficaz.

Essa jornada reforçou não apenas minha formação técnica, mas também minha confiança enquanto pesquisador. Assim, encerro **este percurso** e com o fim dele, abro portas para muitos outros.

Por fim, deixo meu sincero agradecimento aos professores que me acompanharam ao longo do Processo da Residência em IA — Fernando Federson, Cedric Carvalho, Leonardo Alves e Sávio Teles — cujos ensinamentos e apoio foram fundamentais durante toda a jornada. Agradeço também aos demais professores e colegas, com quem não apenas aprendi técnicas de Inteligência Artificial, mas também compartilhei momentos de grande alegria e companheirismo, os quais levarei comigo para toda a vida.

Agradeço, de modo especial, à minha namorada e colega de turma, Julia Dollis; à minha mãe, Ana Paula Machado; ao meu pai, Vladimir Pedrozo; ao meu irmão, Victor Augusto Pedrozo; e aos demais familiares, por todo o suporte ao longo da minha trajetória.

APÊNDICE 1

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 2 de set. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

DANIEL MACHADO PEDROZO

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Durante esta Semana:

1. Utilizei das seguintes ferramentas para busca de artigos e fontes de estudo:
   - ChatGPT
   - Consensus
   - ResearchRabbit
2. Documentei as principais fontes fundamentais para (Fundamentação):
   - Aprendizado por Reforço
     - Reinforcement Learning: An Introduction
     - Proximal Policy Optimization Algorithms
   - LLMs
     - Attention Is All You Need
     - Fine-Tuning Language Models from Human Preferences
     - DeepSeek R1
     - Search-R1
   - Ferramentas
     - HuggingFace TRL
     - TRLx
   - Técnicas
     - RLHF
     - PPO
     - DPO
     - GRPO
3. Através de NotebookLM, utilizando essas fontes, comecei um estudo sobre o conteúdo. E à partir deste, adicionei algumas perguntas que se encontram no documento acima.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Aprofundar o estudo nas fontes de conhecimento encontradas e responder às dúvidas adicionadas

aos documentos.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 11 de set. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

Daniel Machado Pedrozo

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: Aprendizado por Reforço(RL) aplicado à LLMs

Nesta Semana, buscando aprofundar meus conhecimentos técnicos em RL:

- Assisti ao vídeo The FASTEST introduction to Reinforcement Learning on the internet a fim de compreender as principais técnicas de Aprendizado por Reforço e como elas se relacionam e classificam e fiz anotações.
- Para compreender melhor cada algoritmo, suas semelhanças e diferenças, e a linha do tempo do RL até o estado da arte atual em LLMs, criei uma tabela para documentação dos algoritmos vistos e suas características.
    - RL - Algorythms
- Criação e início do preenchimento do Glossário Dicionário - RL , para relembrar, sempre que necessários os termos fundamentais do Aprendizado por Reforço.
- Estudei o funcionamento de um Markov Decision Process (MDP), através do Reinforcement Learning: An Introduction. R. Sutton, and A. Barto.
- Leitura (incompleta) do Reinforcement Learning Algorithms: An Overview and Classification
- Assisti ao vídeo Reinforcement Learning: A 60 Year History para compreender um pouco mais da história do Aprendizado por Reforço.
- Assisti o documentário AlphaGO, o qual narra a história do primeiro modelo de IA à se tornar o melhor jogador de Go (um tradicional e famoso jogo de tabuleiro do leste asiático) do mundo.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Finalizar leitura do Survey
Mapa mental com os algoritmos e suas classificações.
Aplicação dos principais algoritmos na prática para compreender suas semelhanças e diferenças.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Fundamentos em LLM

1. Attention Is All You Need - https://arxiv.org/pdf/1706.03762
2. BERT: Pre-training of Deep Bidirectional Transformers for https://aclanthology.org/N19-1423.pdf
3. **Deep Reinforcement Learning from Human Preferences -** https://arxiv.org/pdf/1706.03741
4. Introduction to Reinforcement Learning and its Role in LLMs **-** https://huggingface.co/learn/llm-course/en/chapter12/2
5. Fine-Tuning Language Models from Human Preferences - https://arxiv.org/pdf/1909.08593
6. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs - https://dl.acm.org/doi/pdf/10.1145/3743127
7. Reinforcement Learning Enhanced LLMs: A Survey - https://arxiv.org/pdf/2412.10400
8. DeepSeek R1: https://arxiv.org/pdf/2501.12948
9. https://cameronrwolfe.substack.com/p/basics-of-reinforcement-learning
10. **O que é RLHF (aprendizado por reforço a partir do feedback humano)? -** https://www.ibm.com/br-pt/think/topics/rlhf

Fundamentos do Aprendizado por Reforço

1. Reinforcement Learning: An Introduction - https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf / https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf
2. **Simple statistical gradient-following algorithms for connectionist reinforcement learning** https://link.springer.com/article/10.1007/BF00992696?#:~:text=Williams%2C R.J. Simple statistical gradient,1007%2FBF00992696
3. Human-level control through deep reinforcementlearning https://web.stanford.edu/class/psych209/Readings/MnihEtAlHassibis15NatureControlDeepRL.pdf
4. Proximal Policy Optimization Algorithms - https://arxiv.org/pdf/1707.06347
5. AlphaGO
    1. https://deepmind.google/research/projects/alphago/
    2. https://www.youtube.com/watch?v=WXuK6gekU1Y&ab_channel=GoogleDeepMind

O RL tem duas principais vertentes que se desenvolveram independentemente antes de se entrelaçam no aprendizado por reforço moderno:

- **Aprendizagem por tentativa e erro:** Originou-se na psicologia da aprendizagem animal e se manifestou em trabalhos iniciais em inteligência artificial. Ronald J. Williams, por exemplo, é creditado por trabalhos seminais sobre algoritmos REINFORCE, uma classe geral de algoritmos de aprendizado por reforço associativo para redes conexionistas.
- **Controle ótimo:** Foca em resolver o problema de projetar um controlador para minimizar ou maximizar o comportamento de um sistema dinâmico ao longo do tempo, utilizando funções de valor e programação dinâmica. Richard Bellman foi uma figura central, introduzindo a equação de Bellman e a versão estocástica discreta do problema de controle ótimo, conhecida como processos de decisão de Markov (MDPs).

Perguntas:

1. O que é um Agente para um LLM?
2. O que é um Ambiente para um LLM?
3. O que é uma Função de Valor para um LLM?
4. O que é uma Política para um LLM?

Ferramentas para Reforço em LLMs:

1. Hugging Face TRL - https://huggingface.co/docs/trl/en/index?utm_source=chatgpt.com

Técnicas de Reforço aplicado à LLMs:

1. PPO
2. RLHF
3. DPO
4. GRPO

Perguntas:

1. Qual a diferença entre elas?

2. Quando utilizar cada uma?
3. Qual a diferença delas para o SFT?
4. Quando elas são melhores que o SFT e quando não?

Algorithm	Policy	Model	State Space	Action Space	Horizon	**Method**	**Backup Method**	Year
Actor-Critic	On/Off-policy	Model-Free	Discrete/ Continuous	Discrete/Continuous	Episodic/Continuous	Actor-Critic	Temporal Difference	1983
REINFORCE	On-policy	Model-Free	Discrete/Continuous	Discrete/Continuous	Episodic	Policy-Based	Monte Carlo	1992
Q-Learning	Off-policy	Model-Free	Discrete	Discrete	Episodic/Continuous	Value-Based	Temporal Difference	1992
SARSA	On-policy	Model-Free	Discrete	Discrete	Episodic/Continuous	Value-Based	Temporal Difference	1994
Monte Carlo Control	On-policy	Model-Free	Discrete	Discrete	Episodic	Value-Based	Monte Carlo	1998
Deep Q-Learning	Off-policy	Model-Free	Continuous	Discrete	Episodic/Continuous	Value-Based	Temporal Difference	2013
DPG	Off-policy	Model-Free	Continuous	Continuous	Episodic/Continuous	Policy-Based	Temporal Difference	2014
TRPO	On-policy	Model-Free	Discrete/Continuous	Discrete/Continuous	Episodic/Continuous	Policy-Based	Temporal Difference	2015
DQN	Off-policy	Model-Free	Continuous	Discrete	Episodic/Continuous	Value-Based	Temporal Difference	2015
A3C	On-policy	Model-Free	Discrete/Continuous	Discrete/Continuous	Episodic/Continuous	Actor-Critic	Temporal Difference	2016
AlphaGO	On-policy	Model-Based	Discrete	Discrete	Episodic	Policy-Based	Monte Carlo	2016
SVPG	On-policy	Model-Free	Discrete/Continuous	Discrete/Continuous	Episodic/Continuous	Policy-Based	Temporal Difference	2016
PPO	On-policy	Model-Free	Discrete/Continuous	Discrete/Continuous	Episodic/Continuous	Policy-Based	Temporal Difference	2017
SAC	Off-policy	Model-Free	Continuous	Continuous	Episodic/Continuous	Actor-Critic	Temporal Difference	2018
IMPALA	Off-policy	Model-Free	Discrete/Continuous	Discrete/Continuous	Episodic/Continuous	Actor-Critic	Temporal Difference	2018
Expected SARSA	Off-policy	Model-Free	Discrete	Discrete	Episodic/Continuous	Value-Based	Temporal Difference	2018
DPO	Off-policy	N/A	N/A	N/A	N/A	Policy-Based	N/A	2023
GRPO	On-policy	N/A	N/A	N/A	N/A	Policy-Based	N/A	2024

Instituto de Informática (INF) - UFG
Alameda Palmeiras, Quadra D, Câmpus Samambaia
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

APÊNDICE 2

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 17 de set. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

DANIEL MACHADO PEDROZO

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: Aprendizado por Reforço aplicado à LLMs:
Durante esta Semana, realizei as seguintes atividades:

- Finalizei a leitura do Survey
- Preenchi a tabela de algoritmos Residência(RL) - Algorítmos com vários dos principais algoritmos de Aprendizado por Reforço classificando-os de acordo com:
  - Policy
  - Model
  - State Space
  - Action Space
  - Horizon
  - Method
  - Backup Method
- Adicionei o ano e o url de referência de cada algoritmo, trazendo uma linha temporal.
- Fiz um mapa mental, classificando os algoritmos de acordo com o tipo de ambiente que eles atuam. E relacionando com as informações obtidas durante o preenchimento da tabela. Também relacionei seu Method(Policy-based, Actor-critic, Value-based) e se é On ou Off Policy.
  Algoritmos de Aprendizado por Reforço.pdf
- Preenchi mais termos do Residência(RL) - Glossário durante as realizações das atividades descritas acima.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Aprofundar no estudo dos algoritmos/métodos focados em LLMs:

- **RLHF**
- **DPO**
- **PPO**
- **GRPO**

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Glossário - Aprendizado por Reforço

Este documento serve como um glossário conciso dos principais termos e conceitos utilizados na área de Aprendizado por Reforço (RL).

Termos e Definições

Agente: O agente é o tomador de decisões e o aprendiz. Ele interage com o ambiente, recebendo representações do estado do ambiente e, com base nisso, selecionando uma ação. O agente tem como objetivo maximizar a quantidade total de recompensa que recebe ao longo do tempo.

Ambiente: Representa tudo o que está fora do agente. Ele recebe as ações do agente e, em resposta, emite uma recompensa numérica e um novo estado para o agente.

Policy [π]: Modo de comportamento do agente em um determinado momento. É um mapeamento de estados para ações a serem tomadas ou para probabilidades de seleção de cada ação possível quando o agente está nesses estados.

Função de Valor: Valor médio esperado de retorno (G(t)) seguindo uma policy (π) em um determinado state(s) ou par state-action (s, a).

- **State-Value [$V\pi(s)$]:** Retorno esperado quando está em um determinado state(s).
  - $V * (s)$**:** Cenário ótimo.
- **Action-Value [$Q\pi(s, a)$]:** Retorno esperado quando está em um determinado estado (s) e toma uma ação (a). Precisa de uma policy.
  - $Q * (s, a)$**:** Cenário ótimo.

Recompensa: Um sinal numérico especial (R(t)) que o ambiente envia ao agente a cada passo de tempo. O objetivo do agente é maximizar a quantidade total de recompensa que recebe, não apenas a recompensa imediata, mas a cumulativa a longo prazo.

Retorno [Gt]: Soma descontada da sequência de recompensas, onde recompensas futuras são progressivamente menos significativas.

Gamma [γ]: Um parâmetro de taxa de desconto. Ele determina o valor presente das recompensas futuras, onde uma recompensa recebida *k* passos no futuro vale gamma vezes o

que valeria se fosse recebida imediatamente. Se gamma=0, o agente é "míope" e se preocupa apenas com recompensas imediatas; se gamma=1, o agente é "previdente" e considera fortemente recompensas futuras.

Actor-Critic: Métodos que aprendem simultaneamente uma política (o "ator") e uma função de valor (o "crítico"). O crítico usa um algoritmo de aprendizado por diferença temporal (TD) para aprender a função de valor da política atual do ator e "critica" as escolhas de ação do ator, enviando erros TD (δ). O ator, por sua vez, atualiza sua política com base nessas críticas

Estocástico:

Determinístico:

Exploração:

Explotação:

Epsilon [ε]: Proporção de vezes que o agente seleciona um a ação aleatória.

Policy Gradient Methods: A política é alterada aumentando a probabilidade das ações que tiveram uma recompensa maior e diminuindo a probabilidade das que tiveram uma recompensa menor.

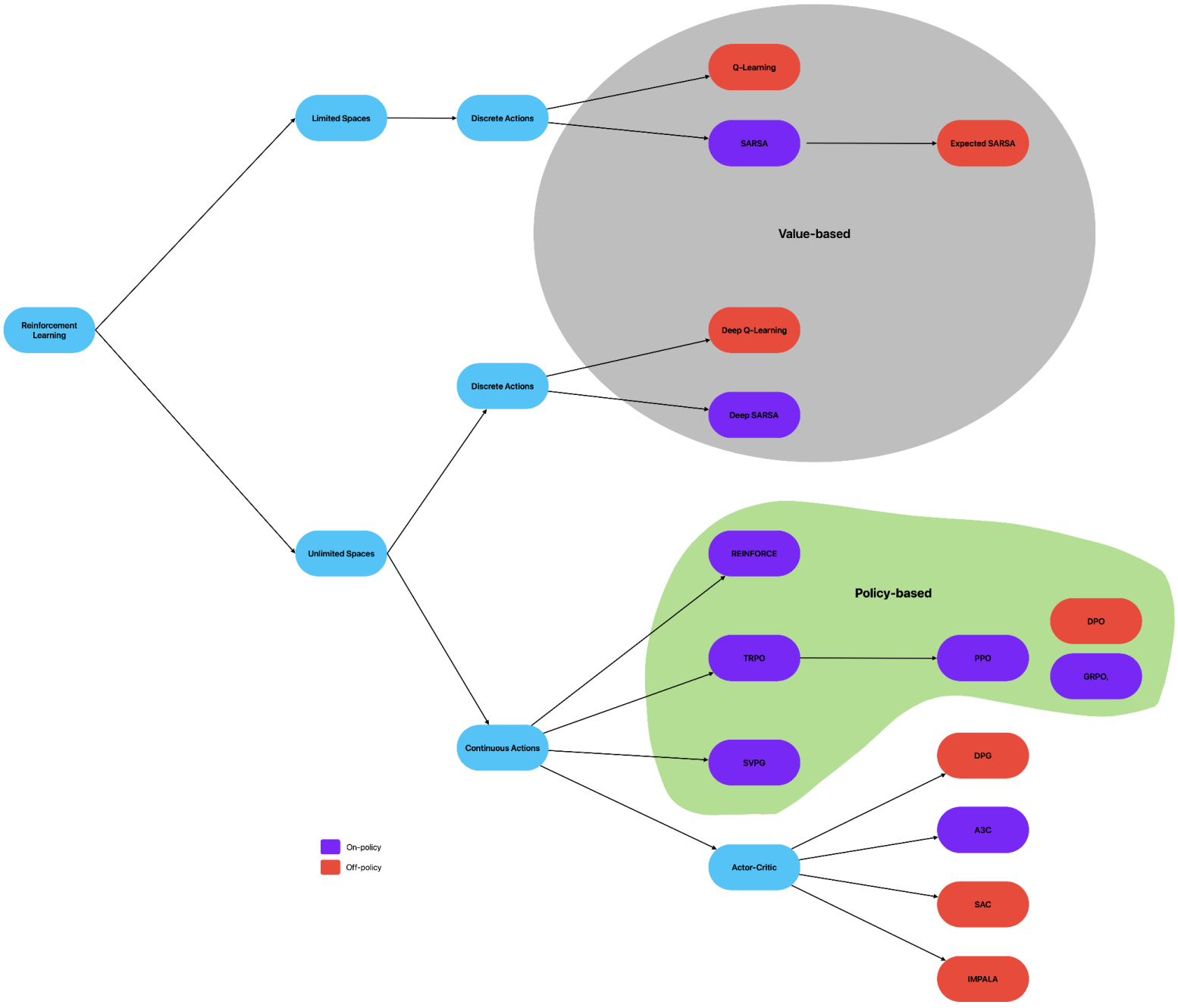
On-policy:

Off-policy:

Sample Efficiency: O quão eficiente um algoritmo é (em quantidade de amostras) para chegar a uma política ótima.

Fontes

- Reinforcement Learning, An Introduction. Barto, Sutton
- Reinforcement Learning Algorithms: An Overview and Classification


Reinforcement Learning
Limited Spaces
Discrete Actions
Q-Learning
SARSA
Expected SARSA
Value-based
Deep Q-Learning
Discrete Actions
Deep SARSA
Unlimited Spaces
REINFORCE
Policy-based
DPO
TRPO
PPO
GRPO,
Continuous Actions
SVPG
DPG
A3C
Actor-Critic
SAC
IMPALA
On-policy
Off-policy

APÊNDICE 3

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 25 de set. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

DANIEL MACHADO PEDROZO

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: Aprendizado por Reforço aplicado a LLMs

Durante esta Semana, desenvolvi as seguintes atividades:

- Senti que ainda me faltava maior familiaridade com o passo a passo dos algoritmos de Aprendizado por Reforço. Para suprir essa necessidade, estudei novamente o **Q-Learning** e **DQN** por meio do *Deep RL Course*(Hugging Face).
  - **Anotações:** Q-Learning.pdf
- Após consolidar essa base, realizei a leitura do artigo **DeepSeek-R1**, explorando os seguintes pontos:
  - Diferença entre **R1** e **R1-Zero**
  - Pipeline de treinamento
    - *Cold Start*
    - *Reasoning RL*
    - *Rejecting Sample*
    - *Diverse RL*
- Aprendizado do algoritmo **GRPO** através **DeepSeek-R1** e complementando com o **DeepSeekMath**.
    - Estrutura geral do algoritmo
    - Função objetivo
    - *Advantage*
    - Termo de divergência de Kullback-Leibler (**DKL**)
- **Anotações:** RL em LLMs.pdf

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Leitura do A Survey of Reinforcement Learning for Large Reasoning Models para complementar o conhecimento dos demais algoritmos.

Explorar tópicos de interesse:

- **Treinamento de Agentes**
  - **https://github.com/OpenPipe/ART**
- **Tool calling reasoning**
  - **Search R1**
- **VLM Reasoning**

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 2 de out. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

DANIEL MACHADO PEDROZO

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: Aprendizado por Reforço aplicado à LLMs

Nos 3 primeiros Stages, estudei **Aprendizado por Reforço** em geral, buscando aumentar minha maturidade na área. No 4o. Stage, afunilar meu estudo em **LLMs**, com foco em GRPO.

Durante esta Semana, desenvolvi as seguintes atividades:

- **Leitura do Survey (págs 20/100):**
  - Na introdução e contextualização, o artigo reforça uma ideia que eu já tinha: que o **uso de RL em LLMs pode reduzir a dependência de dados humanos** e se mostra uma **tecnologia promissora para a ASI** (Artificial Super Intelligence) - **Como poderíamos superar os humanos em tarefas utilizando dados humanos?**
  - Algumas dúvidas que eu tinha, por estar estudando de forma fragmentada, foram esclarecidas — por exemplo, a **viabilidade de granularidade na modelagem das ações do LLM**.
  - **Conheci novas políticas de otimização** que eu ainda não havia visto, traçando também uma história mais completa do campo.
  - Classificação interessante de tipos de recompensas e políticas.
  - Com essa leitura, ficou evidente o quanto a escolha de um bom **survey impacta o conhecimento do estado da arte em uma área de estudo** e que eu teria **economizado um bom tempo se tivesse começado ele antes**.
  - Anotações: Survey.pdf

- **Assisti à Talk Dale Schuurmans: Language Models and Computation**, a qual trouxe reflexões sobre a importância do reasonings para os LLMs:
  - A arquitetura do LLM confere **universalidade computacional** (capacidade de simular Máquinas de Turing) a modelos **até mesmo não treinados**.
  - O **Chain of Thought (CoT) é a computação** (execução) do algoritmo. O problema é que **não se pode compilar algoritmos longos (lineares) em tempo constante**. E se você tentar fazer isso o

modelo vai ‘‘chutar” a resposta.
    - **A Praticidade é o Desafio:** O **Policy Gradient** (RL) está funcionando para treinar o LLM, mas é **muito demorado**. O modelo já viu todas as rotinas (*routines*), mas o desafio está na **exploração eficiente** e em como **programar** essa capacidade máxima.
    - Lacuna de conhecimento: Como aumentar a eficiência da exploração (Atualmente lenta com os Policy Optimization atuais) - Acredito que não vou abordar
    - Me mostrou a **importância de utilizar reasoning** quando buscamos resolver um problema que envolve uma **sequência de passos lógicos**.
    - Anotações: Linguagem Modelos Até Computer.pdf

- **Comecei a leitura do Search-R1**(5/30):
    - Gera autonomamente e de forma estratégica search queries intercalando busca e raciocínio multi-turn com tokens explícitos (<search>, <think>)
    - Resolve a ineficácia de prompts simples e a dependência de dados anotados em larga escala em métodos RAG tradicionais.

- **Exploração do Agent Reinforcement Trainer(ART)**: Explorei a documentação e fiz algumas anotações de pontos de interesse:
    - Framework de treinamento de código aberto para o treinamento de LLM Agents através da experiência.
    - Aceita diferentes arquiteturas(CPU, GPU)
    - Integração com Weights & Bias para observabilidade
    - **Integração com LangGraph - Otimização para uso de tools**

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

Iniciar experimentos utilizando o ART
Analizar demais frameworks e fazer comparação
Continuar leitura do Survey o e Search-R1

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 9 de out. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

DANIEL MACHADO PEDROZO

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: Aprendizado por Reforço aplicado à LLMs

Nos 3 primeiros Stages, estudei **Aprendizado por Reforço** em geral, buscando aumentar minha maturidade na área. À partir do 4o Stage, especifiquei meu estudo em **LLMs**, com foco em **GRPO**.

Durante esta Semana, desenvolvi as seguintes atividades:

- **Estudo do DPO e PPO**:
  - Li os artigos originais do DPO e do RLHF-PPO com foco em compreender suas funções objetivo, semelhanças e diferenças.
  - Consegui esclarecer duas dúvidas que eu tinha:
    - Por que o **DPO não é considerado um método de RL**?
    - O motivo pelo qual o **GRPO é visto como mais eficiente que o PPO,** mesmo sendo bem mais simples?
- **Aplicação:**
  - Através do Survey, estudei as principais áreas de estudo e aplicação emergentes.
  - As técnicas/aplicações/tarefas que mais me chamaram atenção foram.
    - **Tool Reasoning** - capacidade dos modelos de linguagem em interagir com ferramentas externas - já estava interessando antes.
      - Trabalhos: Schick et al., 2023; Dong et al., 2025a; Wei et al., 2025c
    - **MAS–LLMs** (*Multi-Agent Systems*), o uso de aprendizado por reforço tem se destacado por permitir **colaboração, coordenação e atribuição de crédito** entre múltiplos agentes
      - Trabalhos: Lowe et al., 2017; Foerster et al., 2018; Yu et al., 2022
    - **VLMs** (*Vision-Language Models*) têm se beneficiado do uso de RL para alinhar melhor **entendimento multimodal e geração de conteúdo**
      - Trabalhos: Wu et al., 2025f; Jiang et al., 2025b; Duan et al., 2025
    - **Tarefas médias** - Auxílio diagnóstico
      - Trabalhos: Pan et al, 2025d; Chen et al., 2024a; Fan et al., 2025e
- **Frameworks**
  - Busquei os frameworks buscando compreender suas semelhanças:

- TRL
- OpenRLHF
- ART

- Organizei suas informações em Frameworks - RL/LLMs

- **Repositório:**
  - Criei um repositório para colocar meus experimentos.
  - Tentei usar o ART, mas esbarrei com vários problemas de dependências e ainda não consegui.

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

Estudo de mais alguns frameworks:
- MARTI
- NeMo-rl
- veRL

Ler as referências citadas.
Buscar datasets que podem ser utilizados para aplicação prática:
- QA médico
- VQA médico

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

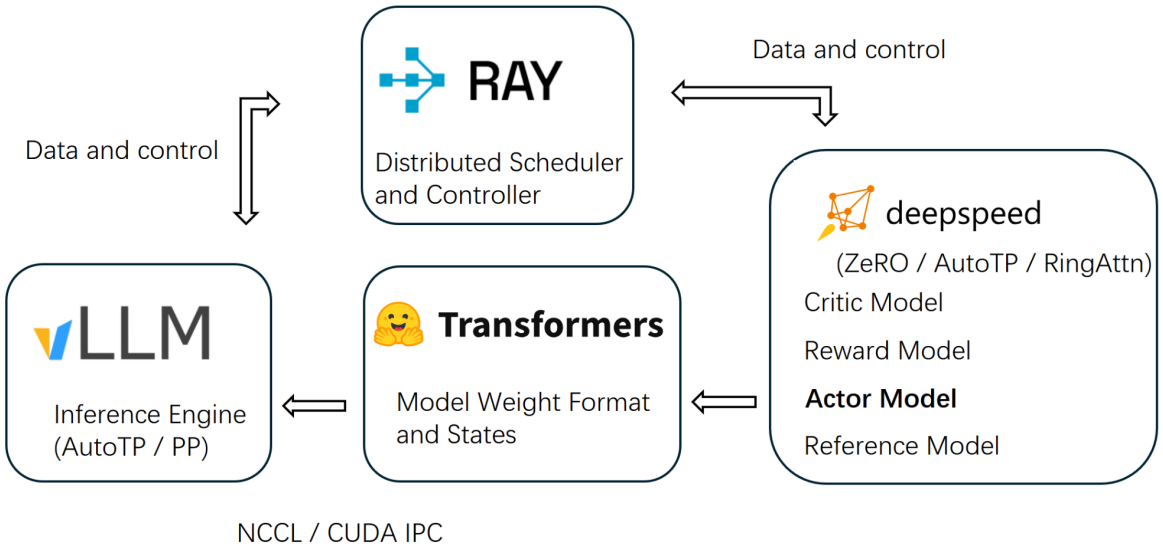
CEDRIC LUIZ DE CARVALHO: Go!

Frameworks

TRL - 15.8k stars

Conjunto de ferramentas para treinar modelos de linguagem de transformers com métodos como SFT, GRPO, DPO, Modelagem de Recompensa.
Mais focado em baixo nível.

OpenRLHF - 8.1k stars



Nível um pouco mais alto que Transformers TRL. Feito para treinamento LLMs em larga escala, já se integrando com: Ray, vLLM, ZeRO-3 and HuggingFace Transformers

ART - 7.5k stars

Agent Reinforcement Trainer é uma estrutura de treinamento de código aberto para o ensino de LLMs de agente para melhorar o desempenho e a confiabilidade através da experiência.

Anotações de pontos de interesse:

Framework de treinamento de código aberto para o treinamento de LLM Agents através da experiência.
Aceita diferentes arquiteturas(CPU, GPU)
Integração com Weights & Bias para observabilidade
Integração com LangGraph - Otimização para uso de tools

Separa o client do server:

VeRL - 14.3k stars

Treinamento RL flexível, eficiente e pronta para produção para grandes modelos de linguagem (LLMs)
Integração com Tools
Integração nativas com vários algoritmos:

- GRPO
- PPO
- DAPO

AReaL - 2.8k stars

AReaL é um sistema de treinamento de aprendizagem de reforço totalmente assíncrono de código aberto para grandes modelos de raciocínio e agentes.
Integração com Tools
Versão leve: AReal-Lite (prototipagem rápida)

APÊNDICE 4

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 15 de out. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

DANIEL MACHADO PEDROZO

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: Aprendizado por Reforço aplicado à LLMs

Nos 3 primeiros Stages, estudei **Aprendizado por Reforço** em geral, buscando aumentar minha maturidade na área. À partir do 4o Stage, especifiquei meu estudo em **LLMs**, os principais algoritmos para otimização de política e suas principais aplicações.

Durante esta última Semana, desenvolvi as seguintes atividades:

1. Analisei os trabalhos que me interessaram na Semana passada, aprofundando em alguns e fiz anotações em RL em LLMs - Trabalhos Interessantes. Dentre eles, os que mais me interessei foi:
    a. MedGemma
    b. Search-R1(e seus semelhantes)
    c. ChestX-Reasoner
2. Sobre os trabalhos acima, fiquei curioso em fazer:
    a. RL no MedGemma para aprendizado de Tool Calling
    b. Reprodução do Search-R1 adaptada
    c. Reprodução do MedGemma adaptada
    d. Reprodução do Chest-X-Reasoner adaptada
3. De todas essas opções, o que mais me chamou atenção foi **reproduzir o Search-R1** e trocar o modelo inicial para o **MedGemma**, para ensiná-lo a utilização de tools e analisar se há **melhorias nos benchmarks médicos**.
4. Estudei mais alguns frameworks encontrados:
    a. VeRL
    b. AReal
    c. Frameworks - RL/LLMs
2. Rodei o ART e testei algumas inferências.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Análise de custo e viabilidade das ideias:
- Custo do uso das Tools
- GPUs necessárias
- Custo/Tempo de treinamento

Iniciar reprodução da ideia com maior viabilidade

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

1. **Agent/Tool Reasoning** - capacidade dos modelos de linguagem em interagir com ferramentas externas - já estava interessando antes.
    1. Trabalhos:
        - Search:
            1. Agentes de pesquisa Wikipedia / Google:
                1. Search-R1: Jin et al., 2025b: foca em permitir que um LLM aprenda, via RL puro, a intercalar raciocínio e buscas em motores de busca como ferramenta durante o processo de inferência, sem depender de supervisão de processo ou distilação pesada. (293 citações)

                   R1-Searcher: Song et al., 2025a: parte do problema de que muitos LLMs raciocinam usando apenas seu conhecimento interno, o que causa erros quando perguntas exigem informação atual ou externalidade, e propõe que o modelo aprenda via RL a invocar buscas externas quando apropriado. (25 citações)

                2. R1-Searcher++ [Song et al., 2025b]: busca levar isso adiante, fazendo o modelo aprender a usar **tanto** conhecimento interno quanto externo de forma adaptativa, com uma fase inicial de SFT para “cold start” e depois RL para “Dynamic Knowledge Acquisition”, incluindo mecanismos de internalização de informação recuperada. (2 citações)
            2. Schick et al., 2023;
            3. Recompensas de diversidade:
                1. Dao e Le[2025]
                2. Mei et al. [2025]
            4. KG:
                1. WebSharper [Tao et al., 2025]
            5. Semelhantes - pós treino com chamada de tools intercaladas.
                1. ARPO - Dong et al., 2025a;
                2. AutoTIR - Wei et al., 2025c;
                3. CoRL - Li et al., 2025b
                4. ToRL - Li et al., 2025q
        - Recompensa baseado nas ferramentas:
            1. Li et. Al [2025v
            2. Paprunia et al. [2025]
            3. Xue et al. [2025b]

        4. ToolRL [Qian et al., 2025]
    - frameworks:
        1. veRL [Sheng et al., 2025]
        2. AReaL [Fu et al., 2025b]
    - features necessárias:
        1. asynchronous generation and training

2. **Tarefas médias** - Auxílio diagnóstico
    1. **Chen et al., 2024a** — aprimora a capacidade de raciocínio ao sintetizar dados confiáveis de trajetórias de raciocínio com um verificador médico e treinar o modelo com SFT e RL. *(107 citações)*
    2. **Gazal-R1 [Arora et al., 2025]** — propõe um sistema de recompensas multicritério que refina a precisão, a aderência ao formato e a qualidade do raciocínio por meio do GRPO, com foco em aprimorar o raciocínio médico. *(0 citações)*
    3. **ProMed [Ding et al., 2025]** — muda o paradigma dos LLMs médicos de reativos para proativos, permitindo que os modelos façam perguntas clinicamente relevantes antes da tomada de decisão, usando recompensas de *Shapley Information Gain* durante a exploração de trajetórias guiada por MCTS e RL. *(1 citação)*
    4. **VLM-R1 [Pan et al., 2025d]** — emprega um framework de RL que incentiva o modelo a descobrir trajetórias de raciocínio interpretáveis por humanos, sem depender de referências explícitas, utilizando recompensas de formato e precisão. *(75 citações)*
    5. **ARMed [Liu and Wei, 2025]** — aborda o colapso de recompensas em tarefas abertas de VQA médica com recompensas semânticas adaptativas, que se ajustam dinamicamente com base na distribuição histórica de recompensas. *(2 citações)*
    6. **MMedAgent-RL [Xia et al., 2025b]** — apresenta um framework multiagente baseado em RL que possibilita colaboração dinâmica e otimizada entre agentes médicos. *(16 citações)*
    7. **MedGemma [Sellergren et al., 2025]** — *(47 citações)*
        - O artigo propõe **MedGemma**, uma família de modelos *foundation* multimodais médicos (baseados em Gemma 3), incluindo uma variante de **4 bilhões de parâmetros** (texto + imagem) e uma variante de **27 bilhões** focada em texto, ambas ajustadas para tarefas médicas.
        - Também introduz o **MedSigLIP**, um encoder visual especializado para imagens médicas. Os autores mostram que o MedGemma supera modelos da mesma escala em tarefas como QA multimodal, classificação de imagens médicas e raciocínio clínico, mantendo boa performance em tarefas gerais. Além disso, um *fine-tuning* adicional

traz ganhos adicionais para subdomínios específicos (ex.: relatórios de raios-X, histopatologia, EHR).

8. **ChestX-Reasoner [Fan et al., 2025e]** — incorpora recompensas de processo derivadas de relatórios clínicos para treinar o modelo a emular o raciocínio passo a passo de radiologistas. *(9 citações)*
9. **CX-Mind [Li et al., 2025k]** — combina SFT e RL com recompensas de formato, resultado e processo para treinar raciocínio intercalado em tarefas de diagnóstico de raios-X torácicos. *(1 citação)*
10. Alex J. Goodell [2025] - https://www.nature.com/articles/s41746-025-01475-8 - Large language model agents can use tools to perform clinical calculations
    - Modelos de linguagem puros cometem muitos erros em cálculos clínicos, mas que essa limitação pode ser superada ao transformá-los em agentes que integram ferramentas especializadas — como mecanismos de recuperação (RAG), interpretadores de código e APIs médicas validadas (OpenMedCalc)

APÊNDICE 5

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 22 de out. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

DANIEL MACHADO PEDROZO

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: Aprendizado por Reforço aplicado à LLMs

Nos 3 primeiros Stages, estudei **Aprendizado por Reforço** em geral, buscando aumentar minha maturidade na área. A partir do 4o Stage, especifiquei meu estudo em **LLMs**, os principais algoritmos para otimização de política e suas principais aplicações.
Compreendi como o Aprendizado por Reforço é utilizado em LLMs, em grande parte, para **aprimoramento da capacidade de ''raciocinar'' através de Chain-of-Thoughts**. Em artigos como "**ToolRL**: Reward is All Tool Learning Needs" e "**Search-R1**: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning", estudei sobre o aprimoramento dessas **Chain-of-Thoughts com chamadas de Tools** e fiquei bastante interessado sobre essa linha de pesquisa.
Na última Semana, defini que iria fazer uma análise de viabilidade da replicação do Search-R1 e o iniciar esse processo.

Durante esta última Semana, desenvolvi as seguintes atividades:

- Buscando fazer a análise da viabilidade da replicação do Search-R1, o qual treina um modelo, através de Aprendizado por Reforço, para aprimorar sua capacidade de raciocínio intercalado com chamadas **tools de Search Engines**:
  - Busquei entender como ele implementa o aprimoramento do uso de tools através do repositório que eles fornecem no artigo.
  - Compreendi que o as Search Tools podem ser:
    - **Estáticas:** Retrieval em bases de dados - **Não levaria um alto custo de API**
    - **Dinâmicas:** Engines de busca na internet
  - Modelos utilizados no artigo variam de 3B à 7B
    - Não são tão "grandes" - É viável a reprodução pois não dependem de GPUs muito grandes.
    - Seria interessante o teste de modelos ainda menores para utilização em computadores pessoais.
  - Compreendi que seria viável a replicação(adaptada) do trabalho
- Clonei o repositório e **depois de muitas tentativas e erros** consegui preparar um ambiente.
  - Baixei uma base de dados textual.

- Fiz sua vetorização para que possa ser utilizada como base de dados para busca do LLM Agent.
- Tive mais erros relacionados ao ambiente e dependências.

Planejamento: **[descrever o que pretende fazer para realizar a próxima ENTREGA]**

Continuar implementação do Search-R1:
- Finalizar implementação do retrieval
- Iniciar o treinamento de um modelo

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 2 de set. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

DANIEL MACHADO PEDROZO

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: Aprendizado por Reforço aplicado à LLMs

Nos 3 primeiros Stages, estudei **Aprendizado por Reforço** em geral, buscando aumentar minha maturidade na área. A partir do 4o Stage, especifiquei meu estudo em **LLMs**, os principais algoritmos para otimização de política e suas principais aplicações.
Compreendi como o Aprendizado por Reforço é utilizado em LLMs, em grande parte, para **aprimoramento da capacidade de "raciocinar" através de Chain-of-Thoughts**. Em artigos como "**ToolRL**: Reward is All Tool Learning Needs" e "**Search-R1**: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning", estudei sobre o aprimoramento dessas **Chain-of-Thoughts com chamadas de Tools** e fiquei bastante interessado sobre essa linha de pesquisa. Preparei o ambiente de treinamento utilizando o framework fornecido pelo Search R1, baixei o dataset e consegui rodar o servidor retriever para servir como a Search Engine.

Durante esta última Semana, desenvolvi as seguintes atividades:

- Durante o processo de treinamento do modelo, precisei lidar com diversos obstáculos. Inicialmente, utilizei o repositório disponibilizado pelo artigo original do Search-R1. Porém, ele se encontra depreciado e não oferece suporte às versões mais recentes das bibliotecas necessárias, como **veRL** e **vLLM**, além de não contemplar arquiteturas de modelos mais atuais, como a **Qwen3**. Tentei contornar essa limitação atualizando manualmente as dependências dentro do container, sem sucesso.
- Aprofundando o estudo dos frameworks envolvidos, percebi que o Search-R1 é essencialmente um *fork* do **veRL**. Diante disso, optei por clonar diretamente a versão mais recente do **veRL** e passar a trabalhar sobre ela.
- Outro desafio surgiu na imagem Docker utilizada como base, que apresentava conflitos de dependências, incluindo incompatibilidades relacionadas ao compilador CUDA. Após diversas tentativas de correção, encontrei no Docker Hub uma imagem oficial mantida pelo próprio **veRL**. Ao adotá-la, consegui estabilizar o ambiente e resolver os problemas de compatibilidade.
- Também foi necessário ajustar o *dataset* para atender às mudanças na nova versão do **veRL**,

especialmente no formato de chamada das *tools*.
    - Qwen2.5-3B (o mesmo usado pelo Search R1 original)
    - Qwen3-0.6B
- Os dois treinamento estão em andamento, o Qwen2.5 já possui alguns métricas parciais metrics_searchr1_qwen2.5

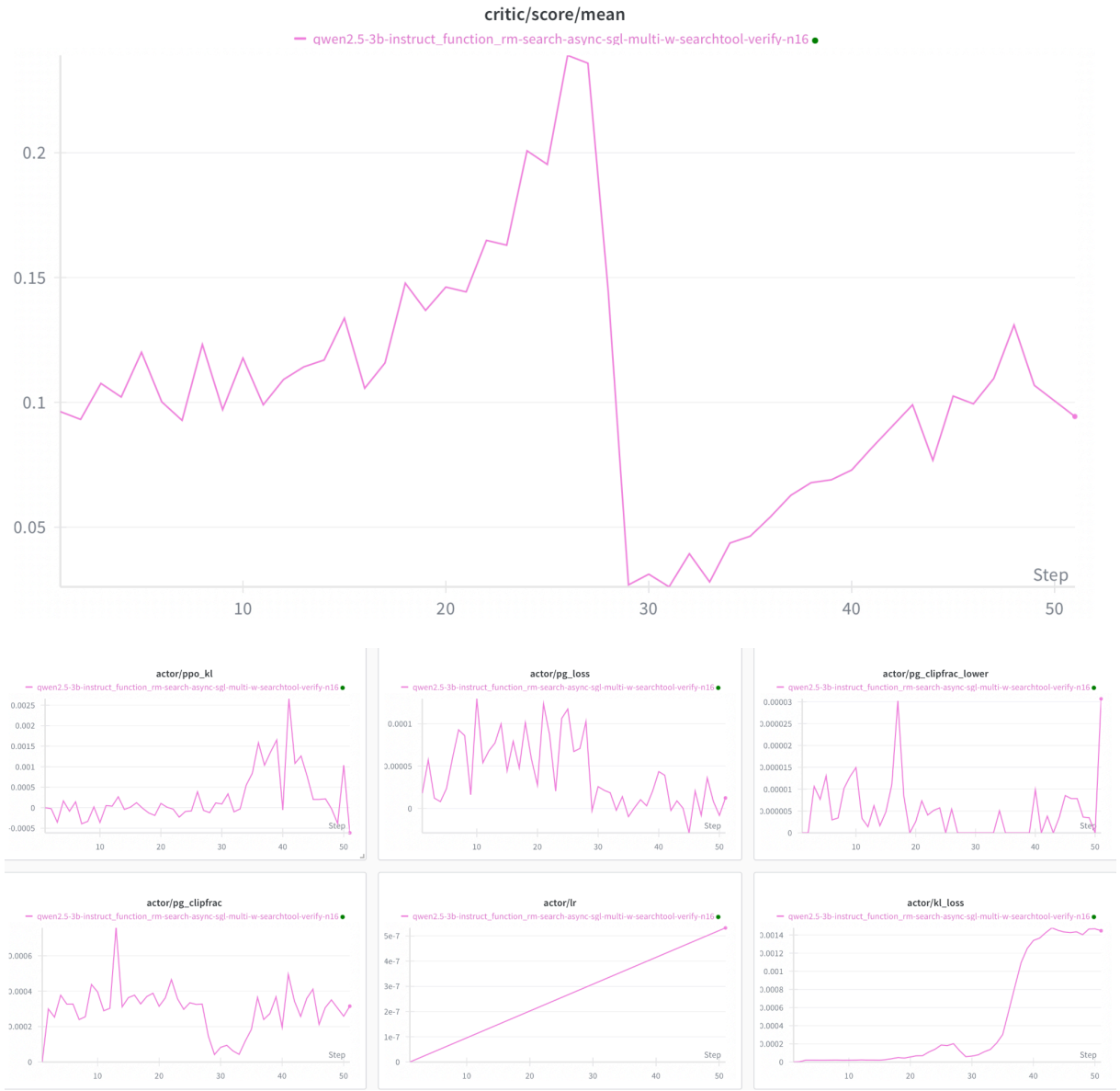
Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Continuar o treinamento e fazer diferentes experimentações com diferentes modelos.

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!


critic/score/mean
qwen2.5-3b-instruct_function_rm-search-async-sgl-multi-w-searchtool-verify-n16
0.2
0.15
0.1
0.05
Step
10
20
30
40
50
actor/ppo_kl
actor/pg_loss
actor/pg_clipfrac_lower
actor/pg_clipfrac
actor/lr
actor/kl_loss


val-core/searchR1_triviaqa/reward/mean@1
val-core/searchR1_popqa/reward/mean@1
val-core/searchR1_nq/reward/mean@1
val-core/searchR1_musique/reward/mean@1
val-core/searchR1_hotpotqa/reward/mean@1
val-core/searchR1_bamboogle/reward/mean@1
Step

APÊNDICE 6

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião ("Gate") de aprovação: 12 de nov. de 2025

Participantes da Entrega [matriculados em Residência em IA]:

DANIEL MACHADO PEDROZO

Entrega: [descrever a ENTREGA - requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Tema: Aprendizado por Reforço aplicado à LLMs

Nos 3 primeiros Stages, estudei **Aprendizado por Reforço** em geral, buscando aumentar minha maturidade na área. A partir do 4o Stage, especifiquei meu estudo em **LLMs**, os principais algoritmos para otimização de política e suas principais aplicações.
Compreendi como o Aprendizado por Reforço é utilizado em LLMs, em grande parte, para **aprimoramento da capacidade de ''raciocinar'' através de Chain-of-Thoughts**.

Em artigos como "**ToolRL**: Reward is All Tool Learning Needs" e "**Search-R1**: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning", estudei sobre o aprimoramento dessas **Chain-of-Thoughts com chamadas de Tools** e fiquei bastante interessado sobre essa linha de pesquisa.

Preparei o ambiente de treinamento utilizando o framework fornecido pelo Search R1, baixei o dataset e consegui rodar o servidor retriever para servir como a Search Engine.

Durante esta última Semana, desenvolvi as seguintes atividades:

- **Replicação do Search R1:** Finalizei o treinamento do modelo Qwen 2.5 3B, replicando com sucesso os resultados obtidos no artigo Search R1.
  - Dataset: Perguntas e Respostas variadas.
  - Modelo Raciocina sobre a pergunta.
  - Se "sentir" necessidade, faz uma busca na base de dados da Wikipedia.
  - Responde a pergunta usando o conteúdo buscado.
- **Tentativa com MedGemma:** Foi realizada uma tentativa de treinamento com o modelo MedGemma, utilizando a mesma *pipeline*. No entanto, devido à sua natureza multimodal, surgiram diversos impedimentos que inviabilizaram a continuidade do treinamento.
- **Investigação de Modelos Pequenos:** Decidi focar em uma nova questão de pesquisa: "Modelos de linguagem de pequena escala conseguem aprender a utilizar ferramentas de busca (*Search Tools*)?".
- **Objetivo Redefinido:** O objetivo passou a ser a utilização de modelos significativamente

menores que os do artigo original para a obtenção de resultados comparáveis.

- **Modelos Utilizados:** Foram empregados o Qwen3 0.6B e o Qwen3 1.7B, ambos modelos compactos que suportam o mecanismo *Chain-of-Thought* (CoT).
- **Aplicação de Reward Model:** Após o treinamento do reasoning com Search Tool, para alinhar as respostas do Qwen3 0.6B às preferências humanas, utilizei um *Reward Model* pré-treinado, o **Skywork/Skywork-Reward-V2-Qwen3-0.6B,** um LLM treinado em dados de LLMs anotados por humanos.
    - A recompensa é a soma da recompensa verificável com a recompensa do LLM Judge
- **Disponibilidade dos Resultados:** Os resultados detalhados desta fase de experimentação estão acessíveis em Resultados - Search Tool.
    - Modelos melhoraram significativamente.
    - Modelos de 0.6B ficaram próximos à modelos 5x maior.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Observação: [caso precise fazer alguma observação, de qualquer "natureza"]

Agradeço ao Artur, pela sugestão do Reward Model.
É muito bom conseguir entender as "letrinhas" do Aprendizado por Reforço..

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go!

Modelo	Média Exact Match %	Média Partial Match %	Média F1 Score
Qwen2.5-3b-it	29.42	33.02	0.2305
Qwen3-06b	25.38	28.41	0.1702
Qwen2.5-3B-Instruct	23.23	23.43	0.1028