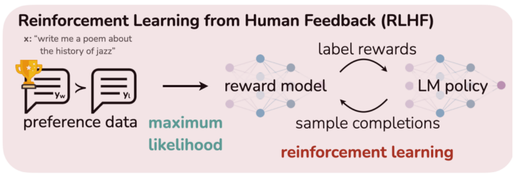
رياضيات المغرب MATH MAROC

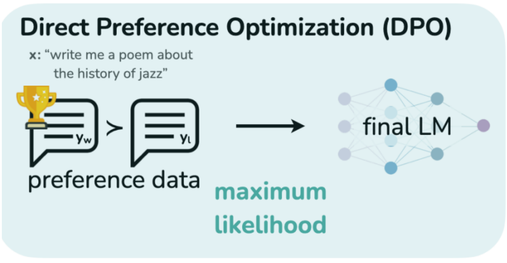
لاصق ديال AI

1 Février 2024

Self-Rewarding Language Models

- الوصفة اللي خدمات بيها OpenAI باش صيبات ChatGPT أو لي سميتها Reinforcement Learning from Human Feedback أو RLHF كاتعتامد على تدريب Reward Model على مجموعة ديال الأسئلة والأجوبة والمقارنات بينات الأجوبة باش من بعد نستعملوه فالتدريب ديال LLM ديالنا بحيث كيولي هو لي كيعطي Feedback على الأجوبة ديال LLM ف RL Pipeline باش كيتعلم كيفاش يجاوب بطريقة لي Aligned مع Human Preference بحال مثلا يجاوب Objectively بلا كثرة الفهامة أو ميخصرش الهضرة، إلخ.
- من بعد جا بحث من Stanford بطريقة جديدة سميتها Direct Preference Optimization أو DPO لي عاودات الصياغة ديال RL Problem باش استخرجات منو Optimal Policy أو هاكا رجع بحال شي Problem ديال Fitting واحد على Model واحد Dataset بلا RL، حيت LLM بحد ذاتو كيلعب دور Reward Model فهاد Formulation.
- دابا جات Meta باش تبين لينا بلي غير LLM راه كافي باش تا ديك Human Feedback ديال Dataset نصيبوها منو بواحد طريقة سميتها LLM-as-a-Judge فين كانخليو Model ديالنا يعطي Evaluation ديال الأجوبة ديالو. من وراها كانديرو واحد DPO لي كيعطينا Model حسن ومنو كانصيبو Preference Data أحس، إلخ.
- هادشي خلاهم من بعد 3 ديال iterations ديال التدريب يحصلو على نتائج حسن من GPT4, Claude2, Gemini Pro فاش قارنوه معاهم باستعمال Alpaca Eval لي كاتعيط على GPT4 يحكم فالمقارنة ديال الأجوبة ويعطينا Winrate.





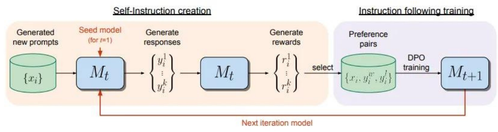
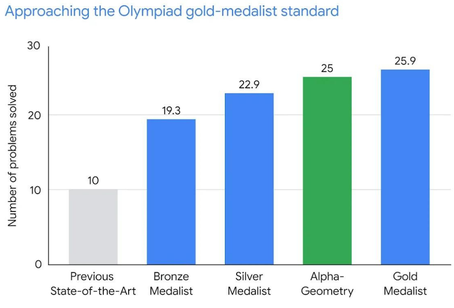
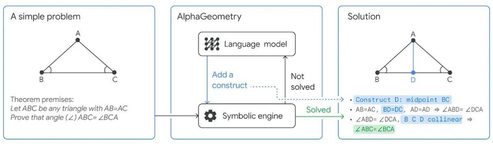


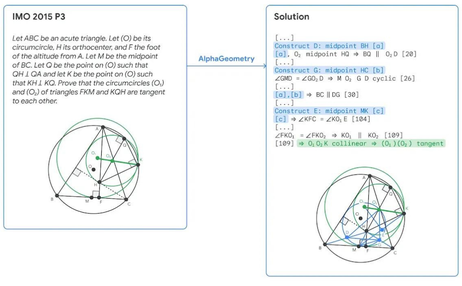
Figure 1: **Self-Rewarding Language Models.** Our self-alignment method consists of two steps: (i) *Self-Instruction creation*: newly created prompts are used to generate candidate responses from model M_t, which also predicts its own rewards via LLM-as-a-Judge prompting. (ii) Instruction following training: preference pairs are selected from the generated data, which are used for training via DPO, resulting in model M_{t+1}. This whole procedure can then be iterated resulting in both improved instruction following and reward modeling ability.

- RLHF: https://arxiv.org/abs/2204.05862
- DPO: https://arxiv.org/abs/2305.18290
- SRM: https://arxiv.org/abs/2401.10020

AlphaGeometry: An Olympiad-level AI system for geometry







- من بعد ما ربحو بطل العالم فلعبة Go باستعمال AlphaGo، أو قدرو يلقاو طرق باش تحصل عل Produit Matriciel بعدد أقل ديال Multiplications من لي كنا كانستعملو بواسطة AlphaTensor، زائد AlphaFold لي كيقدر يتوقع 3D Protein Folding من ترتيب ديال Amino Acids، دابا جات Google DeepMind او خرجات AlphaGeometry لي كيقدر يحل مسائل رياضية ديال الهندسة من درجة IMO يعني لي كيتحطو فالأولمبياد الدولي للرياضيات.
- باش يصولو لهاد المستوى احتاجو أنهم يصنعو Dataset فيها 100 مليون مثال أو تمرين دربو عليه واحد LLM أو هادشي كامل Synthetic Data.
- الفكرة هي أن واحد Symbolic Engine كيحاول يحل المشكل الهندسي باستعمال Brute Force فين كيعدد جميع Statements لي ممكن يتقالو على الشكل الهندسي. من بعد فاش كيوحل كيرجع LLM لي كيقتارح إضافة عنصر جديد ماكانش فالشكل قبل، استنادا على "Intuition" لي طور من التدريب ديالو على ملايين الأمثلة، يعني مثلا يزيد منتصف قطعة محددة.
- هاد التقسيمة لي دارو ماشي جديدة حيت Logic Theorist لي تكتب في 1956 كمحاولة أولية لبرمجة نظام مفكر، كان كيخدم بطريقة مشابهة. وأنظمة أخرى اعتامدو على نفس النهج من قبل غير هوما كانو كيستعملو Heuristics فديك المرحلة ديال الاقتراح، وهاد الحلول ماشي بقوة LLM أو أحيانا كيكونو Random. الفصل مابين جوج أنواع ديال التفكير كانلقاوه تا فكتاب Thinking, Fast and Slow لي كيسميهم System 1 & System 2 أي المرحلة ديال Intuition (Fast) ومرحلة ديال Reasoning (Slow).

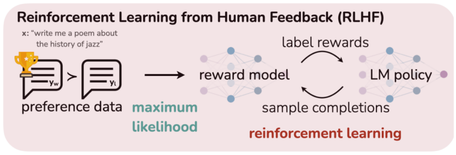
- Blog:
  - deepmind.google/discover/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/
- Code:
  - github.com/google-deepmind/alphageometry

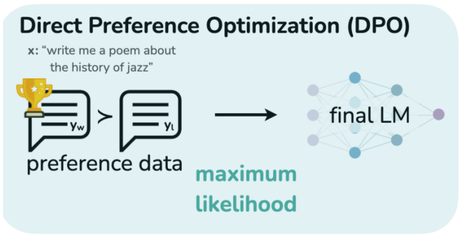
Laseq of AI

1 Février 2024

Self-Rewarding Language Models

- The approach employed by OpenAI to enhance ChatGPT, known as Reinforcement Learning from Human Feedback (RLHF), relies on training a Reward Model on a dataset comprising questions, answers, and comparisons between those responses. This model is utilized in training the Large Language Model (LLM), providing feedback on the LLM's answers within the framework of Reinforcement Learning (RL). Consequently, the model learns how to consistently respond in alignment with human preferences, such as delivering objective responses without unnecessary complexity or inappropriate language, and so forth.
- Subsequently, a study from Stanford University introduces the concept of Direct Preference Optimization (DPO), which reframes the RL problem to extract the Optimal Policy. This approach transforms the problem into a model-fitting problem on a dataset without employing Reinforcement Learning, with LLM playing the role of the reward model in this formulation.
- Now, Meta presents new research suggesting that the Language Model (LLM) alone is sufficient to create the Human Feedback dataset. The said dataset is created using a method called "LLM-as-a-Judge," where the model proposes evaluation for its own responses. DPO is then applied to obtain an improved and more aligned model, that in itself is used to generate Preference Data,
- After 3 training cycles, they found the results using Alpaca Eval, a method that calls CGT4 to compare the answers and provide a Winrat, to be outperforming GPT-4, Claude2, and Gemini Pro!





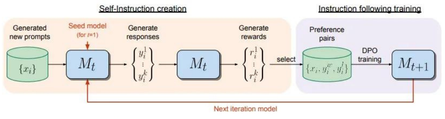
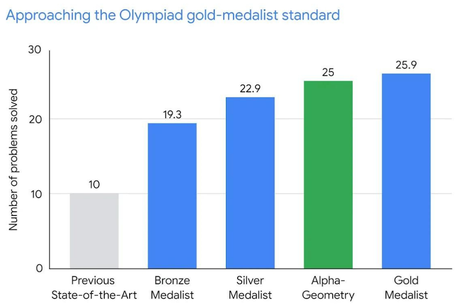
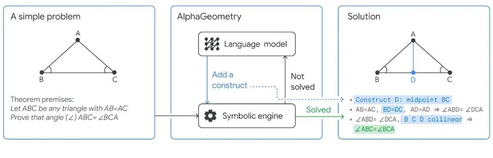


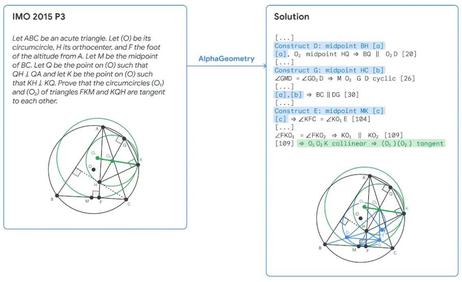
Figure 1: **Self-Rewarding Language Models.** Our self-alignment method consists of two steps: (i) *Self-Instruction creation*: newly created prompts are used to generate candidate responses from model M_t, which also predicts its own rewards via LLM-as-a-Judge prompting. (ii) Instruction following training: preference pairs are selected from the generated data, which are used for training via DPO, resulting in model M_{t+1}. This whole procedure can then be iterated resulting in both improved instruction following and reward modeling ability.

- RLHF: https://arxiv.org/abs/2204.05862
- DPO: https://arxiv.org/abs/2305.18290
- SRM: https://arxiv.org/abs/2401.10020

AlphaGeometry: An Olympiad-level AI system for geometry



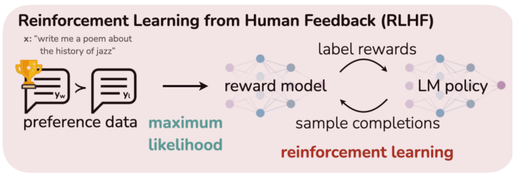




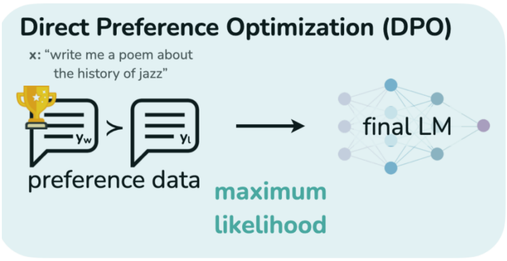
- After the triumph over the world champion in the game of Go using AlphaGo, and their success in finding a way to multiply two matrices with fewer multiplication operations using AlphaTensor, along with AlphaFold's impressive predictive capabilities on three-dimensional folding of proteins through the arrangement of amino acids. Google DeepMind presents us, now, with the AlphaGeometry, a model that can solve mathematical geometric problems at a difficulty level of the International Mathematical Olympiad.
- To achieve this, they created a dataset of 100 million examples or exercises. The dataset was then used to train a Large Language Model (LLM), all using synthetic data.
- The main idea is that there is a Symbolic Engine that attempts to solve the geometric problem using a brute force approach, where it enumerates everything that can be deduced from the geometric shape. When the problem becomes intractable, the Large Language Model (LLM) suggests adding a new element not present in the original shape, based on the "intuition" developed during training on millions of examples, such as increasing the midpoint of a certain piece.
- This approach is somewhat not new, in this regard, we mention the Logic Theorist developed in 1956, which worked in a similar way. However, it relied on professional intuition (Heuristics) or randomness (Random) in the proposal stage, instead of LLMs. There is a division between two types of thinking, namely System 1 and System 2 as explained in the book "Thinking, Fast and Slow," where System 1 happens quickly (Intuition), while System 2 occurs more slowly (Reasoning).
- Blog:
  - deepmind.google/discover/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/
- Code:
  - github.com/google-deepmind/alphageometry

النماذج ذاتية المكافأة

- الوصفة التي استخدمتها OpenAI لتحسين ChatGPT، والتي تُعرف باسم التعلم المقوى باستعمال ملاحظات بشرية (RLHF)، تعتمد على تدريب نموذج المكافآت (Reward Model) على مجموعة من الأسئلة والأجوبة والمقارنات بين تلك الإجابات. يتم استخدام هذا النموذج في تدريب النموذج اللغوي الضخم (LLM)، حيث يقوم بتقديم ملاحظات (Feedback) حول إجابات النموذج اللغوي (LLM) في إطار التعلم بالتقوية (RL). وهكذا يتعلم النموذج كيفية الاجابة بشكل متناسق مع التفضيلات البشرية، مثل الرد بشكلٍ موضوعي دون زيادة في التعقيد أو كلام نابي، وما إلى ذلك.



- ثم أتت دراسة من جامعة Stanford باسم التحسين المباشر من التفضيلات (DPO)، التي تقوم بإعادة صياغة مسألة التعلم بالتقوية (RL) لاستخراج السياسة الأمثل (Optimal Policy). وبهذا يصبح مشكل التحكم في الذكاء الاصطناعي (Alignment) كأي عملية تدريب على مجموعة من البيانات بدون استخدام التعلم بالتقوية، حيث يلعب LLM دور نموذج المكافآت في هذه الصيغة..



- والآن تأتينا شركة Meta ببحث يظهر أن النموذج اللغوي (LLM) فقط كاف لحل مشكلة التحكم. حيث قاموا باستعماله للحصول على مجموعة كبيرة من بيانات التفضيل باستعمال طريقة جديدة تُعرف بـ "النموذج-اللغوي-كحكم" حيث يقوم النموذج باقتراح تقييم لإجاباته. يتم بعد ذلك تطبيق (DPO) للحصول على نموذج محسن ومتحكم بيه يُعطي بيانات تفضيل أحسن من التي سبقتها، ثم نعيد تطبيق هاذه العملي عدة دورات.
- بعد 3 دورات من التدريب، يتحققون من النتائج باستخدام Alpaca Eval ويجدون أنه قد تفوق على GPT4 وClaude2 وGemini Pro !

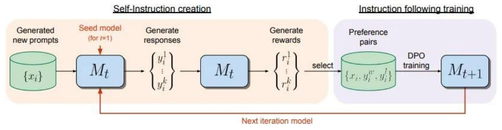
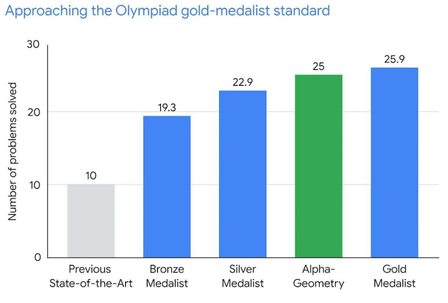


Figure 1: **Self-Rewarding Language Models.** Our self-alignment method consists of two steps: (i) *Self-Instruction creation*: newly created prompts are used to generate candidate responses from model M_t, which also predicts its own rewards via LLM-as-a-Judge prompting. (ii) Instruction following training: preference pairs are selected from the generated data, which are used for training via DPO, resulting in model M_{t+1}. This whole procedure can then be iterated resulting in both improved instruction following and reward modeling ability.

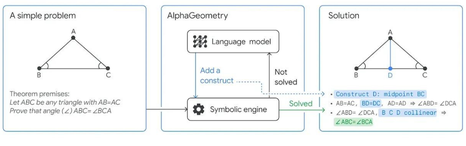
- RLHF: https://arxiv.org/abs/2204.05862
- DPO: https://arxiv.org/abs/2305.18290
- SRM: https://arxiv.org/abs/2401.10020

AlphaGeometry: نموذج ذكاء اصطناعي يَحُلّ المسائل الهندسية من أولمبياد الرياضيات الدولي

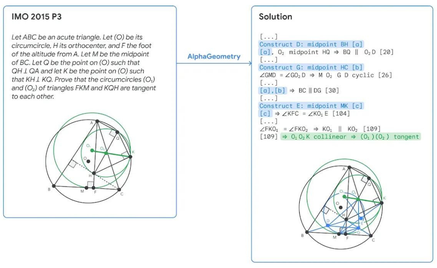
- بعد فوزهم على بطل العالم في لعبة الغو باستخدام AlphaGo، و تمكنهم من العثور على كيفية ضرب مصفوفتين بعدد أقل من عمليات الضرب بواسطة AlphaTensor. بالإضافة إلى AlphaFold القادر على توقع الطي ثلاثي الأبعاد للبروتين من خلال ترتيب أحماضه الأمينية. الآن، تقدم لنا Google DeepMind نموذج AlphaGeometry الذي يستطيع حل مشاكل رياضية هندسية بمستوى صعوبة الأولمبياد الدولية للرياضيات.



- للوصول إلى هذا المستوى، قاموا بإنشاء مجموعة بيانات تحتوي على 100 مليون مثال أو تمرين، ثم دربوا نموذج لغة (LLM) عليها، وهذا كله باستخدام بيانات اصطناعية (Synthetic Data).
- الفكرة الرئيسية هي أن هناك محرك رمزي (Symbolic Engine) يحاول حل المشكلة الهندسية باستخدام هجوم القوة العمياء (Brute Force) حيث يُعَدِّدُ كل ما يمكن استخلاصه من الشكل الهندسي. و بعد أن يستعصي حل المشكلة، يقوم نموذج اللغة الكبير (LLM) بتقديم اقتراح لإضافة عنصر جديد غير موجود في الشكل الأصلي، استناداً على "الحدس" الذي تم تطويره خلال التدريب على الملايين من الأمثلة، مثل زيادة منتصف قطعة معينة.



- هذا النهج ليس جديدًا إلى حد ما، نذكر في هذا الصدد،المَنْظَر المنطقي (Logic Theorist) الذي تم تطويره في عام 1956 حيث كان يعمل بطريقة مشابهة. لكنه اعتمد على الحدس المهني (Heuristics) أو العشوائية (Random) في مرحلة الاقتراح، بدل LLMs. هناك تقسيم بين نوعين من التفكير، وهما System 1 و System 2 كما تم توضيحه في كتاب "Thinking, Fast and Slow"، حيث يحدث System 1 بسرعة (Intuition)، بينما يحدث System 2 بشكل أبطأ (Reasoning).



- Blog:
  - deepmind.google/discover/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/
- Code:
  - github.com/google-deepmind/alphageometry