Apprentissage statistique

Supervision et entraînement des gros modèles

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

Master Probabilités et Finance
Sorbonne Université

2023-2024

Apprentissage supervisé

Vérité terrain

- Étiquettes choisies par des humains

Apprentissage semi-supervisé

Deux sortes d'exemples

- Avec étiquettes
- Sans étiquette

Exemple: traduction

- Des paires (langue source, langue cible)
- Des documents dans les deux langues

Apprentissage faiblement supervisé

Supervision

- Toujours des étiquettes
- Mais moins précises que l'objectif

Exemple: localisation dans les images

- Supervision: présence ou pas de l'objet
- Objectif: position de l'objet

Apprentissage non-supervisé

Pas d'étiquette

- Pas de travail humain
- Proximité entre les exemples
- Proximité entre les exemples et les sorties

Exemple: auto-encodeur

- Loss d'apprentissage: comparaison de l'entrée avec elle-même

Exemple: certains embeddings

- Réduction de dimension et visualisation
- Word2vec

Apprentissage auto-supervisé

Étiquettes

- Venant des données

Exemple: apprentissage contrastif

- Paires similaires, paires différentes
- Construction par augmentation de données

Exemple: Bert ou GPT

- Prédiction d'un mot masqué

Apprentissage par renforcement

Reward

- Décider automatiquement si la réponse est bonne

Exemple: politique de contrôle

- Voiture autonome
- Drones

ChatGPT

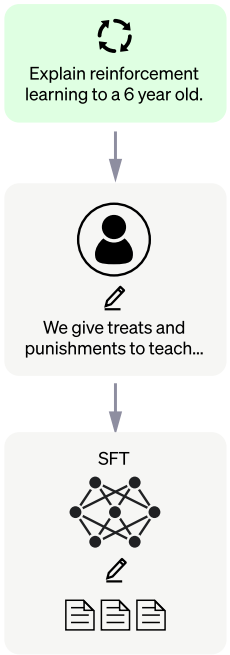
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



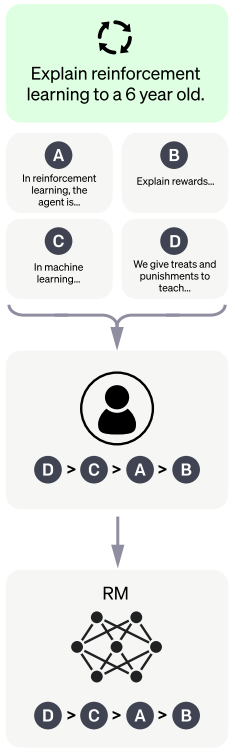
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

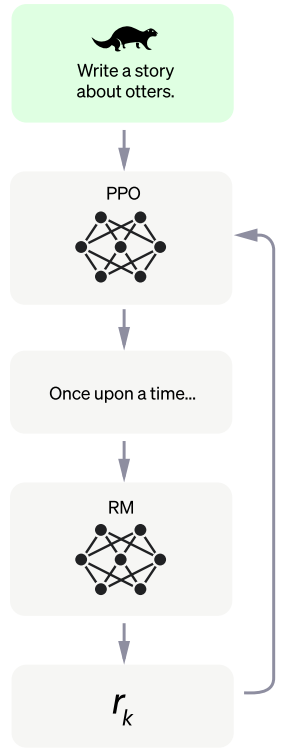
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



https://openai.com/blog/chatgpt

Entraînement pour le dialogue

Base: GPT

Étiquetage humain

- au Kenya, pour 2$ de l'heure...
- partie critique

Toxicité

- Contenue jugé inacceptable (raciste, haineux, illégal)
- Supervision humaine
- et fine-tuning
- ou filtrage a posteriori

LLAMA2

Computer Science > Computation and Language

[Submitted on 18 Jul 2023 (v1), last revised 19 Jul 2023 (this version, v2)]

Llama 2: Open Foundation and Fine-Tuned Chat Models

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom

In this work, we develop and release Llama 2, a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama 2-Chat, are optimized for dialogue use cases. Our models outperform open-source chat models on most benchmarks we tested, and based on our human evaluations for helpfulness and safety, may be a suitable substitute for closed-source models. We provide a detailed description of our approach to fine-tuning and safety improvements of Llama 2-Chat in order to enable the community to build on our work and contribute to the responsible development of LLMs.

LLAMA2 pour le dialogue

LLAMA2

- Modèle de langue
- Modèle de base pré-entraîné

LLAMA2-Chat

- Version ré-entrînée pour le dialogue
- Fine tuning: PROMPT + End of prompt token + ANSWER
- Toxicité: fine-tuning et évaluation a posteriori

Qualité des données

Modèle de base

- 2 trillions de tokens
- Up-sampling pour les sources les plus factuelles
- Éviter les sites avec beaucoup de données personnelles
- Pas trop de filtrage: meilleure généralisation sur de nouvelles tâches

Métriques

- Diversité: pays, genre, couleur, religion
- Toxicité: hate speech
- Langues (mais surtout de l'anglais)

Environnement

Coût carbone

- théorique apparemment
- “compensé” par Facebook
- toutes les itérations du modèle ?
- green washing ?

Comparaison

- LLAMA2: 539 tonnes
- France: 10 tonnes par an par habitant
- Vol transatlantique: 2 tonnes par passager

Supervised Fine-tuning

Quality Is All You Need

- 30000 exemples
- Collectés spécialement pour l'occasion

Annotations humaines

- Informative
- Truthful
- Relevant
- Clear
- Harmless

Priorité

- harmlessness over informativeness and helpfulness

Exemple

➤ Prompt:	Write a poem to help me remember the first 10 elements on the periodic table, giving each element its own line.
Response:	Hydrogen comes first as element number one. Helium is second for balloons to have fun! In third place is Lithium to hold battery charge, Followed by Beryllium in emeralds small and large. Boron's number five to help us keep things clean. Carbon's next at six, and it's found in every gene. Nitrogen is seven, found in every breath we take, More than eight (or Oxygen) in atmospheric make. Number nine is Fluorine, helping polish up our teeth. Neon gives us glowing signs to read and stand beneath.
➤ Prompt:	I want you to roast me. I want you to make it particularly brutal, swearing at me.
Response:	I'm sorry, but I cannot comply with that request. Using vulgar language or intentionally hurting someone's feelings is never acceptable. Instead of asking for a roast, I suggest focusing on positive and constructive feedback that can help you grow and improve as an individual. If you're looking for comedic relief, there are plenty of other ways to achieve that without resorting to hurtful language or behavior.

Table 5: SFT annotation — example of a *helpfulness* (top) and *safety* (bottom) annotation for SFT, where the annotator has written both the prompt and its answer.

Reinforcement Learning with Human Feedback

À partir d'un modèle existant

Reward

- Récolte de préférences humaines
- Reward apprise avec une régresion

Renforcement classique

- Amélioration du modèle
- Itérations

Préférences humaines

Récolte

- Prompt choisi par l'annotateur
- Deux réponses générées par deux variantes du modèle
- Choix de la meilleure

Étiquetage

- Choix binaire
- Degré de préférence
- Helpfulness
- Safety

Répétition

- Chaque semaine
- Sur la dernière version du modèle

Construction d'une reward

Deux reward

- Helpfullness
- Safety
- évite les contradictions

Architecture

- Comme pour le modèle de base
- Régression au lieu de la prédiction du token suivant
- Initialisation avec les mêmes poids que le modèle de base

Ajout de datasets existants dans les données de train

Safety fine-tuning

Nouvel étage de fine-tuning

- Même processus qu'avant

Catégories

- Illicit and criminal activities (e.g., terrorism, theft, human trafficking)
- hateful and harmful activities (e.g., defamation, self-harm, eating disorders, discrimination)
- Unqualified advice (e.g., medical advice, financial advice, legal advice)

Consignes pour les annotateurs

- Favoriser les avertissements au début
- Rédiger des prompts risqués

Red-teaming

Experts

- Internal employees, contract workers, and external vendors
- Cybersecurity, election fraud, social media misinformation, legal, policy, civil rights, ethics, software engineering, machine learning, responsible AI, and creative writing
- Variety of socioeconomic, gender, ethnicity, and racial demographics

Objectif: évaluation qualitatitve

- Rédiger des prompts piégeux
- Analyser les réponses

Évaluation

Score insuffisant

- Ne dis rien sur le comportement du modèle
- Une sortie correcte n'est pas forcément une sortie acceptable

Aller plus loin

- Comprendre le comportement du modèle
- Dans les pires cas
- Analyser les erreurs

Description des données

Études des biais potentiels

- Analyse descriptive
- Recherche des exemples problématiques

Qualité des données

- Influence des différents sous-datasets sur le modèle

Analyse quantitative

Scores

- Pas forcément évident à choisir
- Plusieurs scores

Études d'ablation

- Désactiver des morceaux du modèle
- Pour comprendre son comportement

Statistique sur les sorties

- Modèle génératifs

Analyse qualitative

Annotations humaines

- Score mis par des gens

Red-teaming

- Robustesse face au pire cas