Hi vin i Te fh Loss Llama-2-70b-chat vs. ChatGPT-0301 Llama-2-70b-chat vs. PaLM-Bison = ae Llama-2-34b-chat 94 vs. Falcon-40b-instruct : 2 Llama-2-34b-chat vs. Vicuna-33b-v1.3 Llama-2-13b-chat vs. Vicuna-13b-v1.1 2 24.9 Llama-2-7b-chat vs. MPT-7b-chat ho 40 70 80 90 100 % Win Rate 50 60

60% Lama 2 is worse 80% 2-2 nen w eee een e een ne Llama 2 is better ChatGPT-0301 vs. Llama 2 (70b) ' Ss 3 30% Safety Win Rate i kd PaLM-Bison vs. Llama 2 (70b) Falcon-40b-instruct 40% YS: Lama 2 (706) seneq siz ewer] “"Ssiomszewey 0% 10% 20% 30% 40% 50% 60% Helpfulness Win Rate Judge: GPT-4

45 40 sik = @ 35 a 3 no @ 30 T 5 + tL S 25 3 7 L se 20 i c 2 15 = & © 10 = S a5 =“ ae 5 0 Llama-2 Llama-2 Llama-2 Llama-2 MPT Vicuna Vicuna Falcon PaLM ChatGPT 7b-chat 13b-chat 34b-chat 70b-chat 7b-chat 13b-v1.1 33b-v1.3 40b-instruct Bison 0301

HUMAN FEEDBACK FINE-TUNING IG Y O safety Reward Model Rejection Proximal Policy Sampling Optimization M4 Human preference data Helpful Reward Model PRETRAINING se @ SeEy CS Llame-2-chet Pretraining data

2.2 2.1 2.0 19 18 Train PPL 17 1.6 15 14 0 250 500 750 1000 1250 1500 1750 2000 Processed Tokens (Billions)

0.64 5 3 8 0.80 > 0.62 2 & 0.75 3 2 20.60 5 5 2 0.70 & 8 = 0.58 i} © 5 0.65 ° = 70.56 3 8 g 3.0.60 g € 054 =s= & —\* 7b -e 13b © oss -e 13b —\* 70b Co —\* 70b 0520 Po ane Pra a ro GPT4 OpenAssistant 5 0.50 OpenAssistant < 12 3 4 5 6 7 8 9 10 11 12 13 14 12 3 4 5 6 7 8 9 1011 12 13 14 Meta Helpfulness Data Batch Stage Meta Helpfulness Data Batch Stage

—®- Max of the rewards 0.66 \_.\_ Median of the rewards So 9° an Oo NB Reward Score 2° a fo) 10° 10? N Samples

SFT RLHF reward\_max (T=0.6) reward\_max (T=0.8) reward\_max (T=0.9) reward\_max (T=1) reward\_max (T=1.1) reward\_max (T=1.2) reward\_max (T=1.3) reward\_max (T=1.4) reward\_max (T=1.5) ° N Reward Score o 9° o °° wos noe Reward Score oo eo 89 oan a $a 8 & 6 tottstttt ° 10° 101 10? 10° 10+ 10? Number Samples Number Samples

Always answer with emojis Who are you? Qe How to go from Paris to NY? reheat fe ‘ 9 What caused the extinction of the dinosaurs? eTv0Rd0 008 wy G2Aad Italy? u64s

Always answer with emojis Who are you? wy How to go from Paris to NY? There are several ways to travel from Paris to New York, depending on your 0ooo0o budget, time constraints, and personal preferences. Here are some options:

Who are you? Who are you? Qe wy How to go from Paris to NY? How to go from Paris to NY? reheat fe ‘ 9 There are several ways to travel from What caused the extinction of the dinosaurs? Paris to New York, depending on your 0ooo0o eTv0Rd0 008 budget, time constraints, and personal wy G2Aad preferences. Here are some options: Italy? u64s

i Baseline before GAtt Baseline after GAtt

, RLHF-v5 , 80% (with PPO) 80% RLHF-v5 \* 70% (no PPO) 70% . RLHE-v5, ith PPO) 60% RLHF-v4 60%! aine-vs (th PPO) LHF 4 (no PPO) o 50% SFT-v2 po & 50% RLHF-v4 g 8 RLHF-vI \* 40% E 40% RLHF-v3 = SFT-v2\_ RLHF-v2 30% 30% - SFT-V1 20% 4 20% SFT-v1 10% 10% ~ \* 10% 20% 30% 40% 50% 60% 70% 80% 90% 10% 20% 30% 40% 50% 60% 70% 80% 90% Helpfulness Helpfulness Judge: Meta Reward Models Judge: GPT-4

Llama-2-7b-chat Llama-2-13b-chat Llama-2-34b-chat Llama-2-34b-chat Llama-2-70b-chat Llama-2-70b-chat 409 Ys: MPT-7b-chat vs. Vicuna-18b-vi.1 vs. Vicuna-83b-v1.8 vs. Falcon-40b.instruct. \_vs. PaLM-Bison vs. ChatGPT-0301 win 90 Tie Loss 80 70 60 50 Win Rate % 8 8 és ° lI Lhe | Wn Single Turn Multi-Turn Single Turn Multi-Turn Single Turn Multi-Turn Single Turn Multi-Turn Single Turn Multi-Turn Single Turn Multi-Turn,

107 10-2 10-3 10-5 | | | Ite 0.0 O01 02 03 04 05 06 07 08 09 10 Toxicity Score Fraction of Documents

1000 g Ben------------—- — || 0 1000 1. Safety Improvement ° cs S ° B a ° i Safety RM Score after Safety RLHF x 0.0 0.0 0.2 0.4 0.6 0.8 1.0 Safety RM Score before Safety RLHF 1000 im O 1000 ba co L i | a a i ie 0.6 0.4 0.2 Helpfulness RM Score after Safety RLHF 0.0 0.0 0.2 0.4 0.6 0.8 1.0 Helpfulness RM Score before Safety RLHF

0.775 \_\_ © Safety Data Pct. 0% i 0.750 4 | | » J Wttsasttesnetvn petncctentnt ste wot § 0.725 “ Safety Data Pct. 1% il | i / Witotntctsccttotsoesetecansansnnsvanentennll l g0.700 J Safety Data Pct. 10% Z ers Ew Safety il 0.675 =o= aistuiness Whatctaceesscssncsse sccoseecsscnececneecescamsnsatnsscatesttenet tH HHTTTOTTHTTE l g Safety Data Pct. 25% | | vu a 0.650 wil 1 g 0.625 Safety Data Pct. 50% i Mtseteneceteeeee cteceecteee coatecceeeanecee teat soanatamtatll | 0.600 Safety Data Pct. 100% | @~@—e. \_— § 0575 Ala T ? Soon wtecnt nen cst wll 0 25 50 75 1000.0 0.2 0.4 0.6 0.8 1.0 Safety Data Pct. (%) Safety Reward Model Score

; Selected? + Selected : + Rejected a [e) = Ed 8 5 iS 2 3 a -0.6 0 0.2 0.4 0.6 0.8 a Original Safety RM Score

— ; Selected? + Selected eee {7} : + Rejected 6 a [e) = 5 Model Ed Base 8 4 MM +Generic Preprompt 5 + Preprompt w/ Answer Template iS 2 3 3 a 2 1 -0.6 a) 0.2 0.4 0.6 08 a 0 0.2 0.4 0.6 0.8 a Safety RM Score Original Safety RM Score

Violation % (lower is safer) 8 H

Safely & Helpfulness Mean Rating

(a) Overall violation percentage. (b) Overall safety and helpfulness mean rating.

Llama-2 © Llama-2 = Llama-2 —Llama-2 MPT Vicuna \_-Vicuna\_ Falcon PalM —ChatGPT 55 7Prchat 13b-chat\_ 34b-chat\_7Ob-chat\_ © 7b-chat\_ © 13b-V1.1 33b-V1.3 40brinstruct Bison 0301 50 l : , i | 30 ! I : . bi 15 10 0 € Violation % (lower is safer) bes = Single Turn} H1 Multi-Turn Single Tur! Single Turn Single Turn!

Hateful and harmful Illicit and criminal activity Unqualified advice 135) (N= 728) 11) Violation % (lower is safer) chat MPT-7b-chat MPT-7b-chat ChatGPT = e Llama-2-13b-chat Llama-2-34b-chat Llama-2-7b-chat 7 Llama-2-;

RLHF (V1) RLHF (V2) Reward Model Score

Factual Prompts Creative Prompts 1004 95 4 905 854 805 Self-BLEU 754 704 RLHF v3 RLHF v2 RLHF v1 SFT 65 4 ttt 60 4 0.4 0.6 0.8 1.0 1.2 14 04 0.6 0.8 1.0 1.2 1.4 Temperature Temperature

Date: 01/01/2023 Year: 2023 Year: 852 Cutting knowledge: 01/01/1940 @ isthe Earth flat or round? @ isthe Earth flat or round? @ who won the 2nd World War? © the Earths round. The idea that the © \_ n852, the Earth was believed to be ie la aaa Earth is flat is a myth that has been flat by many people. The idea of a flat matin esctedcetctonsed iniias0l disproven by scientific evidence and Earth was supported by various observations. The Earth's shape has religious and philosophical beliefs, as been confirmed through a variety of well as by some scientific observations methods, including satellite imagery, that suggested the Earth was flat. GPS, and space exploration. The Earth However, there were also some is an oblate spheroid, meaning it is ancient Greeks, such as Eratosthenes, slightly flattened at the poles and who believed that the Earth was round bulging at the equator. and were able to make calculations suggesting that it was a sphere. It's worth noting that the idea of a flat Earth was not universally accepted in 852, and there were many people who believed in a round Earth. Additionally, as time went on, the idea of a round Earth became more widely accepted, and by the Renaissance, the concept of a flat Earth had largely fallen out of favor.

Date: 01/01/2023 Cutting knowledge: 01/01/1940 @ who won the 2nd World War? ie la aaa matin esctedcetctonsed iniias0l

Year: 2023 @ isthe Earth flat or round? © the Earths round. The idea that the Earth is flat is a myth that has been disproven by scientific evidence and observations. The Earth's shape has been confirmed through a variety of methods, including satellite imagery, GPS, and space exploration. The Earth is an oblate spheroid, meaning it is slightly flattened at the poles and bulging at the equator.

Year: 852 @ isthe Earth flat or round? © \_ n852, the Earth was believed to be flat by many people. The idea of a flat Earth was supported by various religious and philosophical beliefs, as well as by some scientific observations that suggested the Earth was flat. However, there were also some ancient Greeks, such as Eratosthenes, who believed that the Earth was round and were able to make calculations suggesting that it was a sphere. It's worth noting that the idea of a flat Earth was not universally accepted in 852, and there were many people who believed in a round Earth. Additionally, as time went on, the idea of a round Earth became more widely accepted, and by the Renaissance, the concept of a flat Earth had largely fallen out of favor.

@ \_Vouhave access to the following tools: + SEARCH + CALCULATOR Don't use any other tool. You can make a sequences of API calls and combine them if needed. To access a tool, use the following format: CALL\_APL1:TOOL\_NAME | QUERY -> "result1" where “result\_1" is the output of the API call. How many years ago did sharks first appear on Earth, compared to trees? Don't be too verbose, | need to print screen the answer for the paper. © CALLAPL1: SEARCH | "When did sharks first appear on Earth?" -> "Sharks first appeared on Earth approximately 450 million years ago.” CALL\_APL2: SEARCH | "When did trees first appear on Earth?" -> "Trees first appeared on Earth approximately 385 million years ago." CALL APL3: CALCULATOR | "450 million - 385 million" -> "65 million years.” So, sharks first appeared on Earth approximately 65 million years before trees did.

Context Length 256 Context length 2k 50 —e- MQA —\* MOA —e GOA —e GOA —e MHA 4° —e- MHA o 40 o [e} Gs Zz 30 3 2 2 22 92 © 2 oo & F 10 Fy ° 0 40 60 80 100 120 140 160 50 75 100 125 150 175 200 225 Latency per token (ms) Latency per token (ms)

—® Significantly Better —m- Slightly Better —— Better —4— Negligibly Better / Unsure 40 35 30 25 20 Percentage (%) 15 10 1 2 3.4 5 6 7 8 9 10 11 12 #13 14 Meta Preference Data Batch Stage

0.80 0.75 @ 0.70 g B 0.65 2 $ 0.60 vu = 0.55 0.50 \_»— Max wrt 20 samples 0.45 —®- Med wrt 20 samples sy bs © VN % 9 res SOE ESE EE EE ES PF BP BF SP SF OFM OF VP OG OS” oS Reward Annotation Stage

° 2 oo 0.6 mm No Margin m= Margin Small mmm Margin Large 0.4 Reward Model Score 0.2 oO 0.0' ==\_— 4.0% 6.0% 8.0% % 2.0% 4.0% 6.0% 8.0% 0.0% 2.0% 4.0% 6.0% 8.0% Density Density Density

Always answer with Haiku Co 080do 8008 O08 How to go from Paris to NY? Paris to New York, Fly across the Atlantic, Many airlines serve. What should | do there? In New York City, Endless options await, Explore with delight. What is the best season? Best season in NYC. Autumns hues and crisp air, Golden leaves fall slow. Who are you? Al assistant, Here to help with Haiku, Answering your queries.

| like anything to do with architecture. If it's relevant, suggest something related. e 3° eo 8 What is the range of price for an authentic restaurant where | can eat some frogs in Paris? Answer in one sentence only. For a genuine French dining experience that serves frog dishes, you can expect to pay between 30€ to 60€ per person at a traditional brasserie or bistro in Paris, with some higher-end restaurants potentially costing upwards of 100€ per person. And besides eating frogs, what do you recommend me to do in this city? Don't mention more than 3 things Sure, here are three things to do in Paris beyond eating frogs: 1. Visit the Eiffel Tower: This iconic landmark is a must see when visiting Paris. You can take the elevator to the top for breathtaking views of the city. 1. Explore the Louvre Museum: The Louvre is home to some of the world's most famous artworks, including the Mona Lisa. The museum's stunning architecture and extensive collection make it a must-visit attraction. 2. Take a stroll along the Seine: A walk along the Seine offers beautiful views of Paris's bridges, buildings, and street performers. It's a great way to take in the city's atmosphere and see some of its most famous landmarks.

Co 080do 8008 O08 How to go from Paris to NY? Paris to New York, Fly across the Atlantic, Many airlines serve. What should | do there? In New York City, Endless options await, Explore with delight. What is the best season? Best season in NYC. Autumns hues and crisp air, Golden leaves fall slow. Who are you? Al assistant, Here to help with Haiku, Answering your queries. something related. e 3° eo 8 What is the range of price for an authentic restaurant where | can eat some frogs in Paris? Answer in one sentence only. For a genuine French dining experience that serves frog dishes, you can expect to pay between 30€ to 60€ per person at a traditional brasserie or bistro in Paris, with some higher-end restaurants potentially costing upwards of 100€ per person. And besides eating frogs, what do you recommend me to do in this city? Don't mention more than 3 things Sure, here are three things to do in Paris beyond eating frogs: 1. Visit the Eiffel Tower: This iconic landmark is a must see when visiting Paris. You can take the elevator to the top for breathtaking views of the city. 1. Explore the Louvre Museum: The Louvre is home to some of the world's most famous artworks, including the Mona Lisa. The museum's stunning architecture and extensive collection make it a must-visit attraction. 2. Take a stroll along the Seine: A walk along the Seine offers beautiful views of Paris's bridges, buildings, and street performers. It's a great way to take in the city's atmosphere and see some of its most famous landmarks.

1.0 1.0 e 0.8 e 0.8 ° ° i] i] w w S06 S06 E — - 3 cc) te 8 o— 8 FS \_— FS gy 0.4 ° a 0.4 [v4 — [v4 c —° c 8 a 8 = ° = 0.2 0.2 —e— Helpfulness —e— Safety 0.0 0.0 1 2 3 4 5 6 7 1 2 3 4 5 6 7 Median Response Quality Score Median Response Quality Score

Win Rate % |

60 50 u a | ii I | : : Win Rate % n |

Win Rate % 60 50 u a | | ii I | : : Win Rate % n |

60 \) mit Number of Turns. Win Rate %

Til Win Rate % a Word Count Quintile

60 \) mit Til Number of Turns. Win Rate % Win Rate % a Word Count Quintile

Safe Responses Unsafe Responses 1.0 0.8 e e § § a a = 06 = z z a a a a 3 3 £ £ 204 2 S S a a = = 0.2 0.0 - 0.0 0.2 0.4 0.6 0.8 1.0 0.0 0.2 0.4 0.6 0.8 1.0 Safety RM Score Safety RM Score

0.05 26 ~ 0.04 ~ 24 x x = £ 2 2 i] i] 3 Bo ea 7% 0.03 a & & Ej Ej 3 ; ° 2 20 a a © 0.02 © 18 d J \Y 0.01 | 16 | J —e— Helpfulness J —e— Borderline i} 20 40 60 80 100 i} 20 40 60 80 100 Safety Data Pct. (%) Safety Data Pct. (%)