Post-training

Aaron Mueller & Ge Gao
CAS CS 505: Natural Language Processing
Boston University
Spring 2026

Overview of Concepts

Post-training aligns a trained model to be more helpful or better at following instructions.

Instruction tuning supervises an LM on data
derived from many NLP tasks.

Reinforcement learning (RL) involves learning via experience, rather than via copying training data.

RLHF, DPO, and **RLVR** are common RL-based
post-training methods.

GPT

- Let's talk about GPT.
- These things are gigantic!

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

GPT4 has roughly **1.8 trillion** parameters !!!

GPT

- Let's talk about GPT.

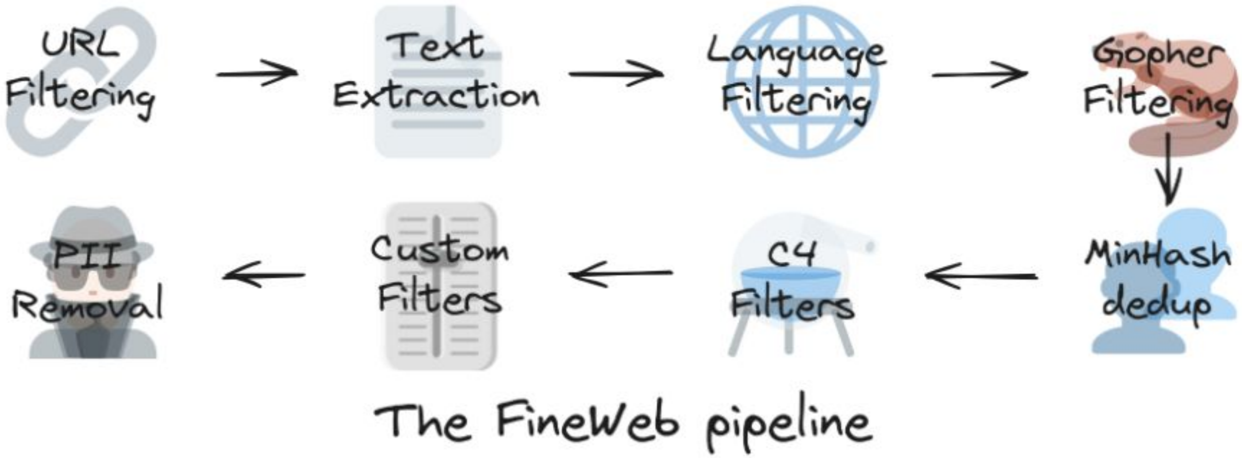
- Trained on an extremely large chunk of text.
  - Example: https://raw.githubusercontent.com/karpathy/char-rnn/master/data/tinyshakespeare/input.txt

  - Reality: entire internet

GPT

- Let's talk about GPT.

- Trained on an extremely large chunk of text.
  - Reality: entire internet
  - https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1



raw text

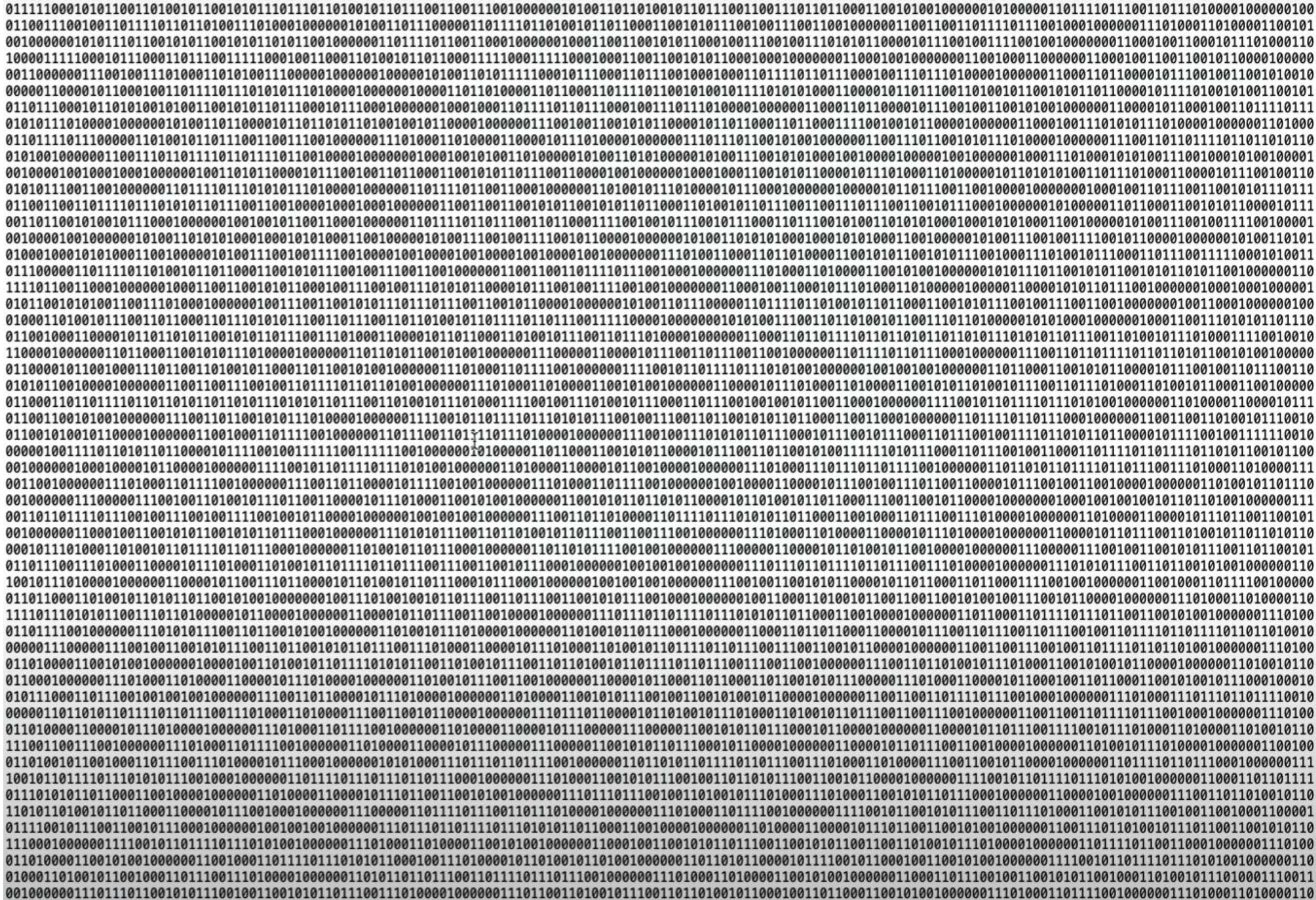
but, computers don’t get text

basis points (BPS). At $16 billion in revenue, every 10 bps of margin reduction results in a $16 million reduction in operating income, or about 2.5¢ in earnings per share reduction. While 10 bps may not sound like a lot, 2.5¢
represents 28% of the analysts $0.09 consensus estimate for Q1. So a minor 2% shift up in gasoline costs can shift earnings lower by 28% in Q1.\nLet's look at this in the wider context of future earnings potential. Amazon bulls
like to point to 2009 Q4 as a shining example of Amazon profit potential as they achieved 4.0% net margin and $0.85 per share of earnings. They claim that once all of the current investments subside, Amazon will either return to
or exceed those peak margin levels on much higher revenue. Here is a dose of reality for the bulls: If gasoline prices per gallon in 2009 Q4 were what it is today, Amazon would have earned a measly $0.02 per share instead of
$0.85! That means given today's energy environment Amazon faces nearly 4.0% net margin headwinds from its prior peak performance.\nHas The Fulfillment Center Build Out Strategy Had Any Positive Impact At All?\nThere is one other
interesting result of the analysis to note: the model generated an additive constant term of negative 8.8%. That means that with zero unit growth and no gasoline price change, Amazon shipping expense would shrink by 8.8% Y/Y. Th
is the positive impact of the distributed fulfillment center strategy and is already baked into the current numbers - you just can't see it in Amazon's financial results due to higher gasoline prices and much higher unit shipmen
than revenue growth over the past four years. The end result is that although Amazon management recognized they need to do something to stem these two major headwinds, they have only lost ground. While they have increased
fulfillment expense as a percentage of revenue by 270 bps, shipping expense has increased 120 bps as a percentage of revenue, thus reducing net margin by 400 bps over the past four years. Based on these trends it seems Amazon wi
need to continue to invest heavily in fulfillment expenses over the next several years just to return to 2009 shipping expense percentage levels of e-commerce revenue.\nIs There Hope For Gasoline Prices to Return to 2009 Levels?
\nThis is highly unlikely as the global financial crisis scared the energy markets into a severe dip (60% drop) at the end of 2008/early 2009. Gasoline prices have begun an exponential run up over the past few weeks and have wel
surpassed both the prior quarter and year ago quarter's level seven weeks into the quarter (Figure 2). Oil prices tend to lead gasoline prices by 4 to 6 weeks, so look for gasoline prices to continue to rise significantly into t
end of Q1 to near all-time high levels ($4.11/gal) as oil continues to print new 9-month highs. That could mean all time high shipping expense as a percentage of e-commerce revenue for Amazon in Q2. The current White House
Administration's tough policy approach on fossil fuels coupled with the Fed's QE infinity program should continue to pressure oil prices higher through at least 2015, leading me to wonder why Amazon is considered a "risk-on" tra
with its strong inverse earnings sensitivity to oil prices. Not only does an increase in gasoline cost increase Amazon's shipping expense, it also lowers consumer's discretionary income, thus providing an additional headwind to
revenue - a true double whammy.\nFigure 2: Weekly US Gasoline Prices for Q/Q and Y/Y Comparisons\nCore to Amazon's value proposition is the convenience of shipping product directly to your door. That service currently reduces
operating margin by 870 bps (up from 750 bps three years ago) and is directly correlated to the cost of gasoline. For every 2% increase in gasoline cost, Amazon gross margin declines by 10 bps. This relation has reduced Amazon n
margin by 400 bps to zero over the last four years. While spending billions on new distributed fulfillment centers has attempted to reduce the impact of soaring gasoline prices, shipping expenses as a percentage of e-commerce
revenue continues to climb as Amazon paid units shipped far outpace revenue growth and gasoline price increases continue to override any Amazon efficiency gains. In the current money printing and political environment, energy
costs are skyrocketing, thus having a direct and meaningful impact on Amazon profit potential for 2013 and beyond.\nNote: Additional information on the model, data and analysis can be made available through my consulting
practice.\nDisclosure: I am short AMZN. Tips for Preventing Medicare Fraud\nThe Department of Health and Human Services (HHS) Office of the Inspector General (OIG) has created a new web site to provide tips for preventing Medica
fraud and medical identity theft. See HHS news release.\nOIG’s new web site includes a brochure containing tips, where to report fraud, and other resources. In addition, CMS issued a Medicare Fraud & Abuse Fact Sheet earlier thi
year. Polish-born Chicago boxer Andrew Golota will plead not guilty at a court date July 19 to charges of unlawful possession of 12 guns, his lawyer said Friday. "There are two sides to every story, and we look forward to a full
exploration of the most recent allegations against Andrew," Matthew P. Walsh II said. Golota, 38, was released from police custody Thursday evening after being charged with the misdemeanor counts of possess-ing unregistered
firearms. His owner's ID card to possess them had been revoked, according to police, who came to Golota's Northwest Side home Wednesday after two women had left there and called 911. No charges have been filed related to other
allegations brought by the women. I’ve a new release of NYTProf ready to upload but I’m stuck.\nThe CPAN Testers service is reporting a failure on a number of systems but I can’t reproduce it locally or work out the cause.\nCan
you reproduce the failure with Devel::NYTProf 2.07_94? If so, could you give me remote access via ssh? (Or spare some time to investigate yourself – I’ll happily clue you in if you can reproduce the problem.)\nUpdate: No one cou
reproduce it. It seems that the failures was not what it appeared to be. A clue was that only one tester was affected. Devel-NYTProf-2.07_94.tar.gz unpacked itself into a directory called Devel-NYTProf-2.07. It seems that when
using CPANPLUS if the user had already got an old Devel-NYTProf-2.07 directory its contents got merged and tests would fail. I’m not convinced that’s the whole story, but Devel-NYTProf-2.07_95.tar.gz unpacked into a Devel-NYTPro
2.07_95 directory and didn’t run into the problem.\nUpdate: More usefully, Andreas made my wish come true by pointing out the -solve parameter to the ctgetreports utility in his CPAN::Testers::ParseReports distribution. It “trie
to identify the best contenders for a blame using Statistics::Regression. [...] The function prints the [...] top 3 candidates according to R^2 with their regression analysis.” Cool. Reprinted from the renowned Bach-Gesellschaft
edition, this work features the complete Sonatas and Partitas for Unaccompanied Violin and the six Sonatas for Violin and Clavier. The music has been reproduced in a size large enough to read easily, with large noteheads, wide
margins for notes, and lay-f... read more\nCustomers who bought this book also bought:\nOur Editors also recommend:\nThe Art of the Fugue & A Musical Offering by Johann Sebastian Bach 19 canons and fugues (complete with a piano
reduction). Also included is a set of canonical variations on a theme by Frederick the Great, complete with solutions and a realization of the keyboard part.\nAn Encyclopedia of the Violin by Alberto Bachmann First published in
1925, this renowned reference remains unsurpassed as a source of essential information, from construction and evolution to repertoire and technique. Includes a glossary and 73 illustrations.\nThe Three Violin Concerti in Full
Score by Johann Sebastian Bach The Concerto in A Minor (BWV 1041), the Concerto in E Major (BWV 1042), and the Concerto for Two Violins in D Minor (BWV 1043) – reproduced from an authoritative edition.\nCaprices and Etudes for
Solo Violin by Nicol Paganini, Henryk Wieniawski Here, in one large-format playing edition, are Paganini's innovative Caprices and Wieniawski's much-admired collections, L'Ecole Moderne and Etudes Caprices. The three scores offe
hours of rigorous practice and rehearsal.\nMass in B Minor in Full Score by Johann Sebastian Bach The crowning glory of Bach's achievements in the field of sacred music, this work is a powerful statement of Christian faith. This
affordable miniature score is a reprint of an authoritative edition.\nReprinted from the renowned Bach-Gesellschaft edition, this work features the complete Sonatas and Partitas for Unaccompanied Violin and the six Sonatas for
Violin and Clavier. The music has been reproduced in a size large enough to read easily, with large noteheads, wide margins for notes, and lay-flat pages.\nReprinted selections from Volume 9 and Volume 27 of Johann Sebastian
Bach's Werke, Bach-Gesellschaft, Leipzig, 1860 and 1879.\nThis book was printed in the United States of America.\nDover books are made to last a lifetime. Our US book-manufacturing partners produce the highest quality books in t
world and they create jobs for our fellow citizens. Manufacturing in the United States also ensures that our books are printed in an environmentally friendly fashion, on paper sourced from responsibly managed forests. Informatio
contained on this page is provided by companies via press release distributed through PR Newswire, an independent third-party content provider. PR Newswire, WorldNow and this Station make no warranties or representations in
connection therewith.\nSOURCE Urban School Food Alliance (USFA)\nUrban School Food Alliance Shows Solidarity by Serving the Same Menu to 2.9 Million Students\nLOS ANGELES, March 20, 2013 /PRNewswire/ -- In a show of solidarity i
providing healthy meals to U.S. students, the Urban School Food Alliance (USFA), which consists of some of the largest school districts in the country, will be serving the same lunch menu to its more than 2.9 million students
today. The six school districts, New York City Department of Education, Los Angeles Unified School District, Chicago Public Schools, Dallas Independent School District, Miami-Dade County Public Schools and Orange County Public
Schools in Orlando, officially formed USFA earlier this year to use their purchasing power to continue to improve food quality for students while driving down costs.\n"This show of solidarity is unprecedented," said Los Angeles
Unified School District Food Services Director David Binkle. "It demonstrates that all the school districts in the alliance can work together to implement the same programs while serving nutritious meals to our students."\nThe
lunch meal, consisting of savory roasted chicken, brown rice with seasoned black or red beans, steamed green broccoli, fresh seasonal fruit and milk, will be served more than 2.6 million times today in the six districts.\n"We
created this menu based on the most popular items we commonly serve in each of our Districts," said Leslie Fowler, Director of Nutrition Support Services at Chicago Public Schools. "Our goal moving forward is to identify these
commonalities and work with our vendors to capitalize on our purchasing power so that we're providing the best and freshest foods possible to our students at the lowest costs possible."\nThe Urban School Food Alliance first met
Denver in 2012 and has since kept in communication regularly via telephone. The food services directors from each of the school districts share and review menu items to ensure that they provide access to meals that meet the USDA
nutrient recommendations including whole grain products, low fat dairy, fresh produce and lean protein that when prepared are calorie conscious and low in fat, sugar and sodium. In addition to creating a purchasing powerhouse, t
coalition is working to be more ecologically friendly in their respective districts. As a whole, the alliance procures more than $530 million in food and food supplies annually.\nIn recent years, the school districts in the
alliance have implemented innovative ways to provide access to school meals including expanded options for reaching more students with breakfast on campus as well as supper after school in order to meet the needs of
students.\n©2012 PR Newswire. All Rights Reserved. Mastodon Two India shows cancelled\nTwo India shows cancelled\nDue to circumstances out of the bands control, Mastodon's appearance in India this month has been canceled; Percep
Live, the organizer of FLY Music Festival, India’s first ever mutli-genre music events announced that the festival stands cancelled due to unforeseen circumstances. All ticket holders who booked their tickets viaBook My Show or
Kya Zoonga will be contacted immediately for a full refund.\nFor Mastodon news that is customized to you and your hometown such as local concerts and appearances, please fill out the form. At Record Label, we are very concerned
about our fans' privacy and do not share or sell your information to other third parties.\nYou are signed up for email and/or mobile alerts Chat Box Chicks ~ Welcoming the New Year!\nView Single Post\n02-10-2013, 03:02 PM\nJoin
Date: Sep 2006\nDoes anyone have the long rectangle nesties? They're on my wish list but I'm wondering if I'd really make use out of them.\nView Public Profile\nSend a private message to laurlynn\nVisit laurlynn's Gallery\nFind
More Posts by laurlynn Tallahassee, FL (Sports Network) - An appearance in the 2013 MEAC-SWAC Challenge, a trip to Ohio State and five home games highlight the Florida A&M football schedule announced on Thursday.\nThe Rattlers,
who have a new head coach in Earl Holmes, will play 12 games, including five at home for the first time since 2006.\nFlorida A&M will face Mississippi State in the MEAC-SWAC Challenge on Sept. 1 in Orlando. The Rattlers will bot
open and close their regular season at the Florida Citrus Bowl because they also will face rival Bethune-Cookman in the Florida Classic on Nov. 23.\nThe Sept. 21 game with Ohio State is the first between the two programs and wil
net Florida A&M an $850,000 guarantee.\nFlorida A&M's home games at Bragg Memorial Stadium will be against Tennessee State (Sept. 7) and Samford (Sept. 14) in non-conference action, and Howard (Oct. 19, Homecoming), North Caroli
A&T (Oct. 26) and Delaware State (Nov. 16) within the MEAC.\nThe Rattlers also will play conference games on the road against Morgan State (Oct. 5), Savannah State (Oct. 12), Norfolk State (Nov. 2) and South Carolina State (Nov.
9).\nHolmes is trying to lift Florida A&M from a 4-7 finish last season.\n2013 Florida A&M Football Schedule\nSept. 1, Mississippi Valley State, MEAC-SWAC Challenge in Orlando, Fla.\nSept. 7, Tennessee State\nSept. 14,
Samford\nSept. 21, at Ohio State\nOct. 5, at Morgan State*\nOct. 12, at Savannah State*\nOct. 19, Howard* (Homecoming)\nOct. 26, North Carolina A&T*\nNov. 2, at Norfolk State*\nNov. 9, at South Carolina State*\nNov. 16, Delaware
State*\nNov. 23, Bethune-Cookman*, Florida Classic in Orlando, Fla.\n* - Mid-Eastern Athletic Conference game - TV & Video\n- Renovation & Decorating\n- How To\n- Be Healthy\n- home beautiful\nYou can also make one large shakshu
to share. Just spoon the sauce into a large ovenproof ceramic dish. Crack eggs into the dish, spaced evenly apart, then cook following the same recipe.\n1. Preheat oven to 180°C. Put six ¼-cup capacity ovenproof ramekins on an
oven tray. Heat oil in a large frying pan over a medium-high heat. Add onion, garlic and chilli and cook, stirring, for 1 minute. Add chorizo and cook for a further 2 minutes or until onion is soft. Add cumin and paprika and coo
for a further 30 seconds.\n2. Add tomatoes, sugar and salt and bring to the boil. Reduce heat to low and cook for 5 minutes or until sauce is thickened.\n3. Spoon sauce evenly among prepared ramekins. Carefully crack an egg into
centre of each. Bake for 12 minutes or until whites have set and yolks are still soft.4. Sprinkle each with dukkah and parsley and serve with sliced baguette. Oklahoma Deer Hunters Hit Record Harvest\nOklahoma deer hunters have
already achieved a record harvest for the state in 2006 – 104,000 and counting – but their opportunities to take deer are far from being over. The state's special antlerless deer season is just around the corner. Scheduled for

bytes

vocab size: 2

seq length:
too long



bytes

vocab size: 256

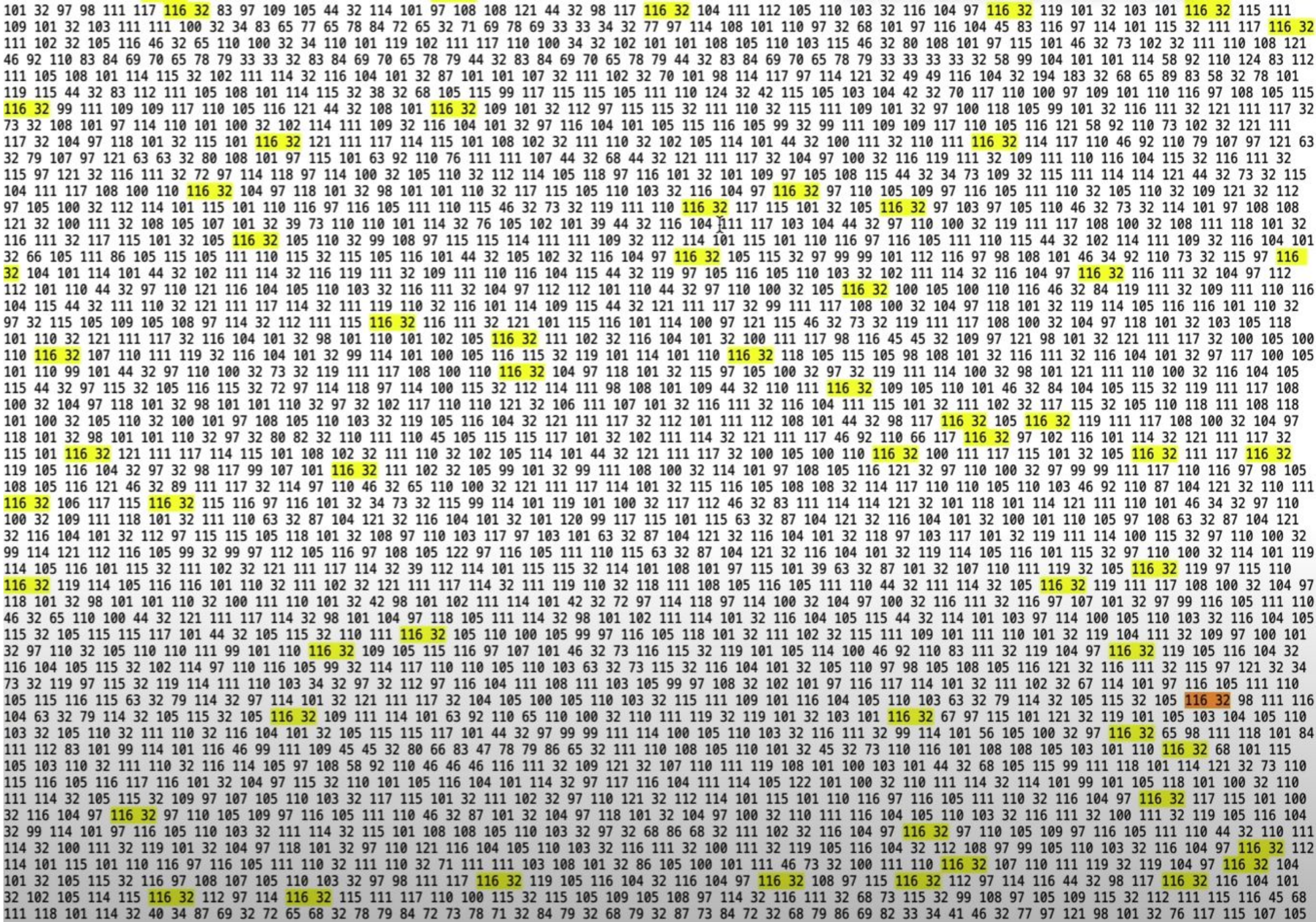
seq length: long

101 32 97 98 111 117 116 32 83 97 109 105 44 32 114 101 97 108 108 121 44 32 98 117 116 32 104 111 112 105 110 103 32 116 104 97 116 32 119 101 32 103 101 116 32 115 111
109 101 32 103 111 111 100 32 34 83 65 77 65 78 84 72 65 32 71 69 78 69 33 33 34 32 77 97 114 108 101 110 97 32 68 101 97 116 104 45 83 116 97 114 101 115 32 111 117 116 32
111 102 32 105 116 46 32 65 110 100 32 34 110 101 119 102 111 117 110 100 34 32 102 101 101 108 105 110 103 115 46 32 80 108 101 97 115 101 46 32 73 102 32 111 110 108 121
46 92 110 83 84 69 70 65 78 79 33 33 32 83 84 69 70 65 78 79 44 32 83 84 69 70 65 78 79 44 32 83 84 69 70 65 78 79 33 33 33 33 32 58 99 104 101 101 114 58 92 110 124 83 112
111 105 108 101 114 115 32 102 111 114 32 116 104 101 32 87 101 101 107 32 111 102 32 70 101 98 114 117 97 114 121 32 49 49 116 104 32 194 183 32 68 65 89 83 58 32 78 101
119 115 44 32 83 112 111 105 108 101 114 115 32 38 32 68 105 115 99 117 115 115 105 111 110 124 32 42 115 105 103 104 42 32 70 117 110 100 97 109 101 110 116 97 108 105 115
116 32 99 111 109 109 117 110 105 116 121 44 32 108 101 116 32 109 101 32 112 97 115 115 32 111 110 32 115 111 109 101 32 97 100 118 105 99 101 32 116 111 32 121 111 117 32
73 32 108 101 97 114 110 101 100 32 102 114 111 109 32 116 104 101 32 97 116 104 101 105 115 116 105 99 32 99 111 109 109 117 110 105 116 121 58 92 110 73 102 32 121 111
117 32 104 97 118 101 32 115 101 116 32 121 111 117 114 115 101 108 102 32 111 110 32 102 105 114 101 44 32 100 111 32 110 111 116 32 114 117 110 46 92 110 79 107 97 121 63
32 79 107 97 121 63 63 32 80 108 101 97 115 101 63 92 110 76 111 111 107 44 32 68 44 32 121 111 117 32 104 97 100 32 116 119 111 32 109 111 110 116 104 115 32 116 111 32
115 97 121 32 116 111 32 72 97 114 118 97 114 100 32 105 110 32 112 114 105 118 97 116 101 32 101 109 97 105 108 115 44 32 34 73 109 32 115 111 114 114 121 44 32 73 32 115
104 111 117 108 100 110 116 32 104 97 118 101 32 98 101 101 110 32 117 115 105 110 103 32 116 104 97 116 32 97 110 105 109 97 116 105 111 110 32 105 110 32 109 121 32 112
97 105 100 32 112 114 101 115 101 110 116 97 116 105 111 110 115 46 32 73 32 119 111 110 116 32 117 115 101 32 105 116 32 97 103 97 105 110 46 32 73 32 114 101 97 108 108
121 32 100 111 32 108 105 107 101 32 39 73 110 110 101 114 32 76 105 102 101 39 44 32 116 104 111 117 103 104 44 32 97 110 100 32 119 111 117 108 100 32 108 111 118 101 32
116 111 32 117 115 101 32 105 116 32 105 110 32 99 108 97 115 115 114 111 111 109 32 112 114 101 115 101 110 116 97 116 105 111 110 115 44 32 102 114 111 109 32 116 104 101
32 66 105 111 86 105 115 105 111 110 115 32 115 105 116 101 44 32 105 102 32 116 104 97 116 32 105 115 32 97 99 99 101 112 116 97 98 108 101 46 34 92 110 73 32 115 97 116
32 104 101 114 101 44 32 102 111 114 32 116 119 111 32 109 111 110 116 104 115 44 32 119 97 105 116 105 110 103 32 102 111 114 32 116 104 97 116 32 116 111 32 104 97 112
112 101 110 44 32 97 110 121 116 104 105 110 103 32 116 111 32 104 97 112 112 101 110 44 32 97 110 100 32 105 116 32 100 105 100 110 116 46 32 84 119 111 32 109 111 110 116
104 115 44 32 111 110 32 121 111 117 114 32 111 119 110 32 116 101 114 109 115 44 32 121 111 117 32 99 111 117 108 100 32 104 97 118 101 32 119 114 105 116 116 101 110 32
97 32 115 105 109 105 108 97 114 32 112 111 115 116 32 116 111 32 121 101 115 116 101 114 100 97 121 115 46 32 73 32 119 111 117 108 100 32 104 97 118 101 32 103 105 118
101 110 32 121 111 117 32 116 104 101 32 98 101 110 101 102 105 116 32 111 102 32 116 104 101 32 100 111 117 98 116 45 45 32 109 97 121 98 101 32 121 111 117 32 100 105 100
110 116 32 107 110 111 119 32 116 104 101 32 99 114 101 100 105 116 115 32 119 101 114 101 110 116 32 118 105 115 105 98 108 101 32 116 111 32 116 104 101 32 97 117 100 105
101 110 99 101 44 32 97 110 100 32 73 32 119 111 117 108 100 110 116 32 104 97 118 101 32 115 97 105 100 32 97 32 119 111 114 100 32 98 101 121 111 110 100 32 116 104 105
115 44 32 97 115 32 105 116 115 32 72 97 114 118 97 114 100 115 32 112 114 111 98 108 101 109 44 32 110 111 116 32 109 105 110 101 46 32 84 104 105 115 32 119 111 117 108
100 32 104 97 118 101 32 98 101 101 110 32 97 32 102 117 110 110 121 32 106 111 107 101 32 116 111 32 116 104 111 115 101 32 111 102 32 117 115 32 105 110 118 111 108 118
101 100 32 105 110 32 100 101 97 108 105 110 103 32 119 105 116 104 32 121 111 117 32 112 101 111 112 108 101 44 32 98 117 116 32 105 116 32 119 111 117 108 100 32 104 97
118 101 32 98 101 101 110 32 97 32 80 82 32 110 111 110 45 105 115 115 117 101 32 102 111 114 32 121 111 117 46 92 110 66 117 116 32 97 102 116 101 114 32 121 111 117 32
115 101 116 32 121 111 117 114 115 101 108 102 32 111 110 32 102 105 114 101 44 32 121 111 117 32 100 105 100 110 116 32 100 111 117 115 101 32 105 116 32 111 117 116 32
119 105 116 104 32 97 32 98 117 99 107 101 116 32 111 102 32 105 99 101 32 99 111 108 100 32 114 101 97 108 105 116 121 32 97 110 100 32 97 99 99 111 117 110 116 97 98 105
108 105 116 121 46 32 89 111 117 32 114 97 110 46 32 65 110 100 32 121 111 117 114 32 115 116 105 108 108 32 114 117 110 110 105 110 103 46 92 110 87 104 121 32 110 111
116 32 106 117 115 116 32 115 116 97 116 101 32 34 73 32 115 99 114 101 119 101 100 32 117 112 46 32 83 111 114 114 121 32 101 118 101 114 121 111 110 101 46 34 32 97 110
100 32 109 111 118 101 32 111 110 63 32 87 104 121 32 116 104 101 32 101 120 99 117 115 101 115 63 32 87 104 121 32 116 104 101 32 100 101 110 105 97 108 63 32 87 104 121
32 116 104 101 32 112 97 115 115 105 118 101 32 108 97 110 103 117 97 103 101 63 32 87 104 121 32 116 104 101 32 118 97 103 117 101 32 119 111 114 100 115 32 97 110 100 32
99 114 121 112 116 105 99 32 99 97 112 105 116 97 108 105 122 97 116 105 111 110 115 63 32 87 104 121 32 116 104 101 32 119 114 105 116 101 115 32 97 110 100 32 114 101 119
114 105 116 101 115 32 111 102 32 121 111 117 114 32 39 112 114 101 115 115 32 114 101 108 101 97 115 101 39 63 32 87 101 32 107 110 111 119 32 105 116 32 119 97 115 110
116 32 119 114 105 116 116 101 110 32 111 102 32 121 111 117 114 32 111 119 110 32 118 111 108 105 116 105 111 110 44 32 111 114 32 105 116 32 119 111 117 108 100 32 104 97
118 101 32 98 101 101 110 32 100 111 110 101 32 42 98 101 102 111 114 101 42 32 72 97 114 118 97 114 100 32 104 97 100 32 116 111 32 116 97 107 101 32 97 99 116 105 111 110
46 32 65 110 100 44 32 121 111 117 114 32 98 101 104 97 118 105 111 114 32 98 101 102 111 114 101 32 116 104 105 115 44 32 114 101 103 97 114 100 105 110 103 32 116 104 105
115 32 105 115 115 117 101 44 32 105 115 32 110 111 116 32 105 110 100 105 99 97 116 105 118 101 32 111 102 32 115 111 109 101 111 110 101 32 119 104 111 32 109 97 100 101
32 97 110 32 105 110 110 111 99 101 110 116 32 109 105 115 116 97 107 101 46 32 73 116 115 32 119 101 105 114 100 46 92 110 83 111 32 119 104 97 116 32 119 105 116 104 32
116 104 105 115 32 102 114 97 110 116 105 99 32 114 117 110 110 105 110 103 63 32 73 115 32 116 104 101 32 105 110 97 98 105 108 105 116 121 32 116 111 32 115 97 121 32 34
73 32 119 97 115 32 119 114 111 110 103 34 32 97 32 112 97 116 104 111 108 111 103 105 99 97 108 32 102 101 97 116 117 114 101 32 111 102 32 67 114 101 97 116 105 111 110
105 115 116 115 63 32 79 114 32 97 114 101 32 121 111 117 32 104 105 100 105 110 103 32 115 111 109 101 116 104 105 110 103 63 32 79 114 32 105 115 32 105 116 32 98 111 116
104 63 32 79 114 32 105 115 32 105 116 32 109 111 114 101 63 92 110 65 110 100 32 110 111 119 32 119 101 32 103 101 116 32 67 97 115 101 121 32 119 101 105 103 104 105 110
103 32 105 110 32 111 110 32 116 104 101 32 105 115 115 117 101 44 32 97 99 99 111 114 100 105 110 103 32 116 111 32 99 114 101 56 105 100 32 97 116 32 65 98 111 118 101 84
111 112 83 101 99 114 101 116 46 99 111 109 45 45 32 80 66 83 47 78 79 86 65 32 111 110 108 105 110 101 32 45 32 73 110 116 101 108 108 105 103 101 110 116 32 68 101 115
105 103 110 32 111 110 32 116 114 105 97 108 58 92 110 46 46 46 116 111 32 109 121 32 107 110 111 119 108 101 100 103 101 44 32 68 105 115 99 111 118 101 114 121 32 73 110
115 116 105 116 117 116 101 32 104 97 115 32 110 101 105 116 104 101 114 32 97 117 116 104 111 114 105 122 101 100 32 110 111 114 32 114 101 99 101 105 118 101 100 32 110
111 114 32 105 115 32 109 97 107 105 110 103 32 117 115 101 32 111 102 32 97 110 121 32 112 114 101 115 101 110 116 97 116 105 111 110 32 116 104 97 116 32 117 115 101 100
32 116 104 97 116 32 97 110 105 109 97 116 105 111 110 46 32 87 101 32 104 97 118 101 32 104 97 100 32 110 111 116 104 105 110 103 32 116 111 32 100 111 32 119 105 116 104
32 99 114 101 97 116 105 110 103 32 111 114 32 115 101 108 108 105 110 103 32 97 32 68 86 68 32 111 102 32 116 104 97 116 32 97 110 105 109 97 116 105 111 110 44 32 110 111
114 32 100 111 32 119 101 32 104 97 118 101 32 97 110 121 116 104 105 110 103 32 116 111 32 100 111 32 119 105 116 104 32 112 108 97 99 105 110 103 32 116 104 97 116 32 112
114 101 115 101 110 116 97 116 105 111 110 32 111 110 32 71 111 111 103 108 101 32 86 105 100 101 111 46 73 32 100 111 110 116 32 107 110 111 119 32 119 104 97 116 32 104
101 32 105 115 32 116 97 108 107 105 110 103 32 97 98 111 117 116 32 119 105 116 104 32 116 104 97 116 32 108 97 115 116 32 112 97 114 116 44 32 98 117 116 32 116 104 101
32 102 105 114 115 116 32 112 97 114 116 32 115 111 117 110 100 115 32 115 105 109 105 108 97 114 32 116 111 32 68 73 115 32 99 108 97 105 109 115 32 112 111 115 116 45 68
111 118 101 114 32 40 34 87 69 32 72 65 68 32 78 79 84 72 73 78 71 32 84 79 32 68 79 32 87 73 84 72 32 68 79 86 69 82 33 34 41 46 32 77 97 121 98 101 32 76 117 115 107 105

unique tokens

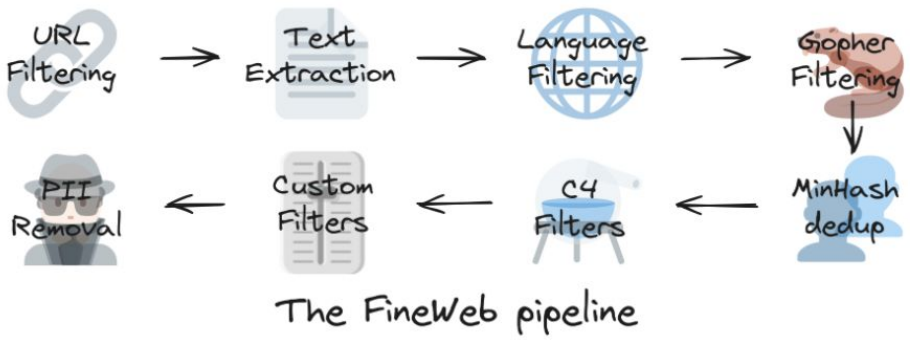
vocab size:
100k

seq length:
short



GPT

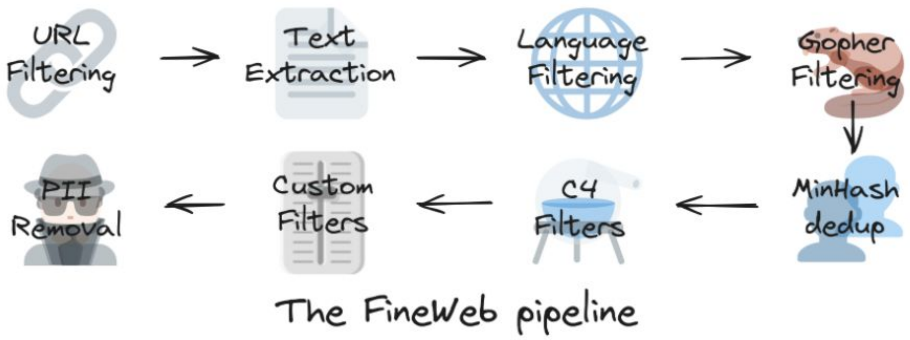
- Step 1: download and preprocess the internet



- Step 2: tokenization
  - convert raw text into sequences of symbols (tokens)
  - https://tiktokenizer.vercel.app/

GPT

- Step 1: download and preprocess the internet

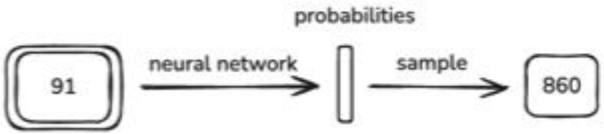


- Step 2: tokenization
  - convert raw text into sequences of symbols (tokens)
  - https://tiktokenizer.vercel.app/

- Step 3: training neural network (Decoder of the Transformer)
  - visualization: https://bbycroft.net/llm

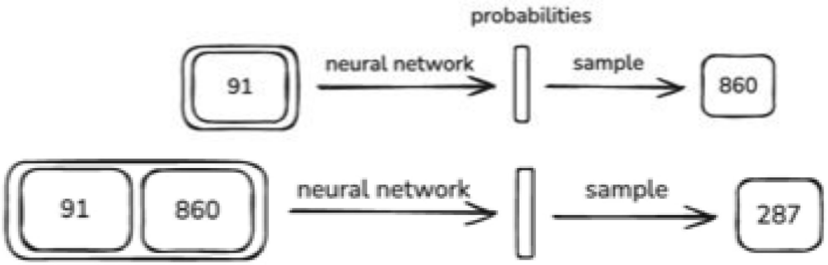
GPT

- Step 3: training neural network (Decoder of the Transformer)
  - visualization: https://bbycroft.net/llm
  - inference:



GPT

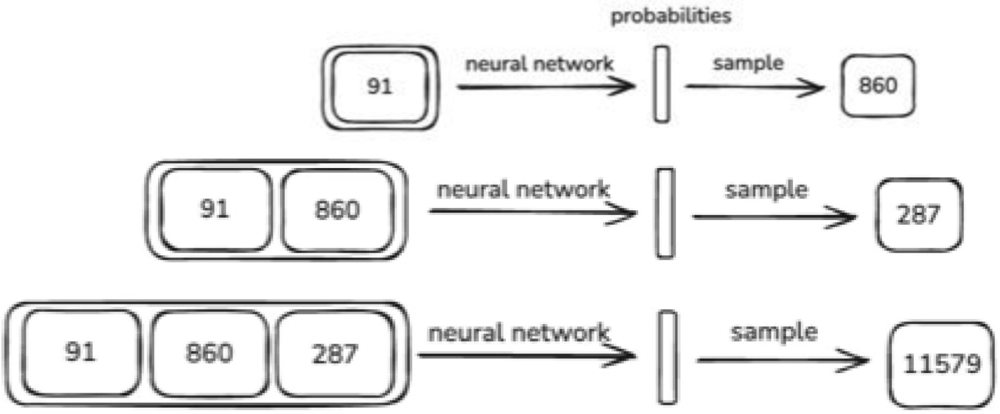
- Step 3: training neural network (Decoder of the Transformer)
  - visualization: https://bbycroft.net/llm
  - inference:



GPT

- Step 3: training neural network (Decoder of the Transformer)
  - visualization: https://bbycroft.net/llm

  - inference:



GPT

- So that's GPT. What did we achieve when we successfully trained a GPT?

GPT

- So that's GPT. What did we achieve when we successfully trained a GPT?
    - It's just a blabber that keeps generating text that looks like human languages.
    - Or rather, it is just "dreaming" the internet that it has seen.
    - https://huggingface.co/openai-community/gpt2

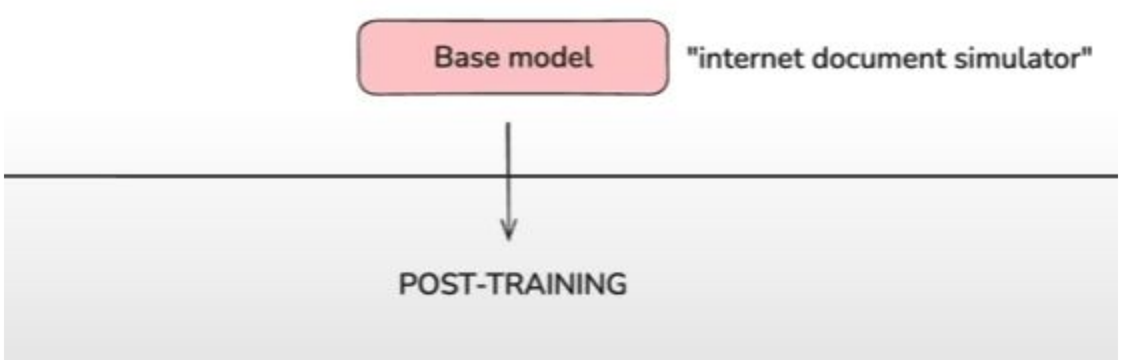
- It's not the magical ChatGPT yet
    - It doesn't interact with you.

GPT

- So that's GPT. What did we achieve when we successfully trained a GPT?
  - It's just a blabber that keeps generating text that looks like human languages.
  - Or rather, it is just "dreaming" the internet that it has seen.
  - https://huggingface.co/openai-community/gpt2

- It's not the magical ChatGPT yet
  - It doesn't interact with you.
  - we call these models the base models (GPT2,3,4, Llama..)

ChatGPT



Post-Training:

- Train on conversations between human and AI assistant

Instruction Tuning

Language models don't know what you want.

- Pre-trained models simply predict the most probable next token.
  - This is not the same thing as following instructions.

Prompt: What year was Boston University founded?

LM output: Better yet, what year was Northeastern University founded? Or Boston College?

- A large model has probably seen "Boston University was founded in 1839" before. How can we get it to actually respond to this query in a helpful way?

Instruction Tuning Data

Translate this from English to Thai:

Cats sleep quietly.
แมวนอนหลับอย่างเงียบ ๆ

Summarize this report.

studies of massive galaxy clusters and groups at @xmath5 typically find environments with little - to - no star formation activity…

Answer the question given the context.

Context: Boston University is a private research university in Boston, Massachusetts. BU was founded in 1839 by…

Question: When was BU founded?

Answer: 1839

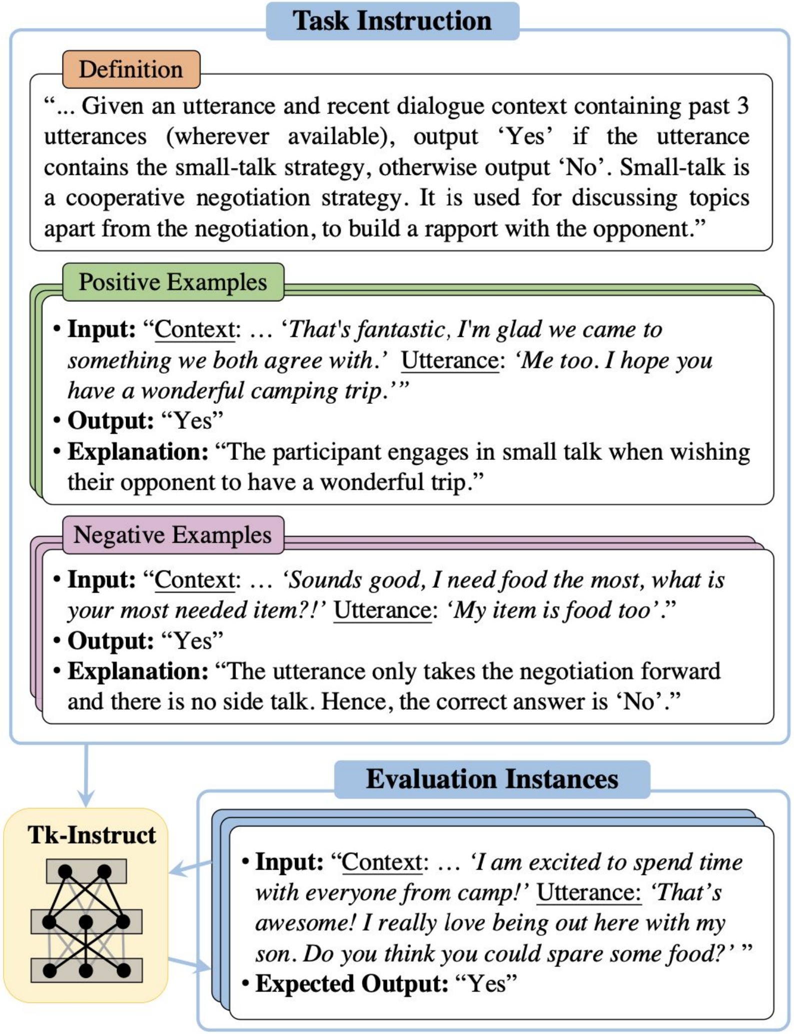
Concatenate instructions to supervised responses. Train on these with typical self-supervised language modeling objective.

Generating Instruction Tuning Examples

- You can hand-write instructions, but this would take a lot of effort.
- It's more common to design templates, and slot in examples from existing datasets.

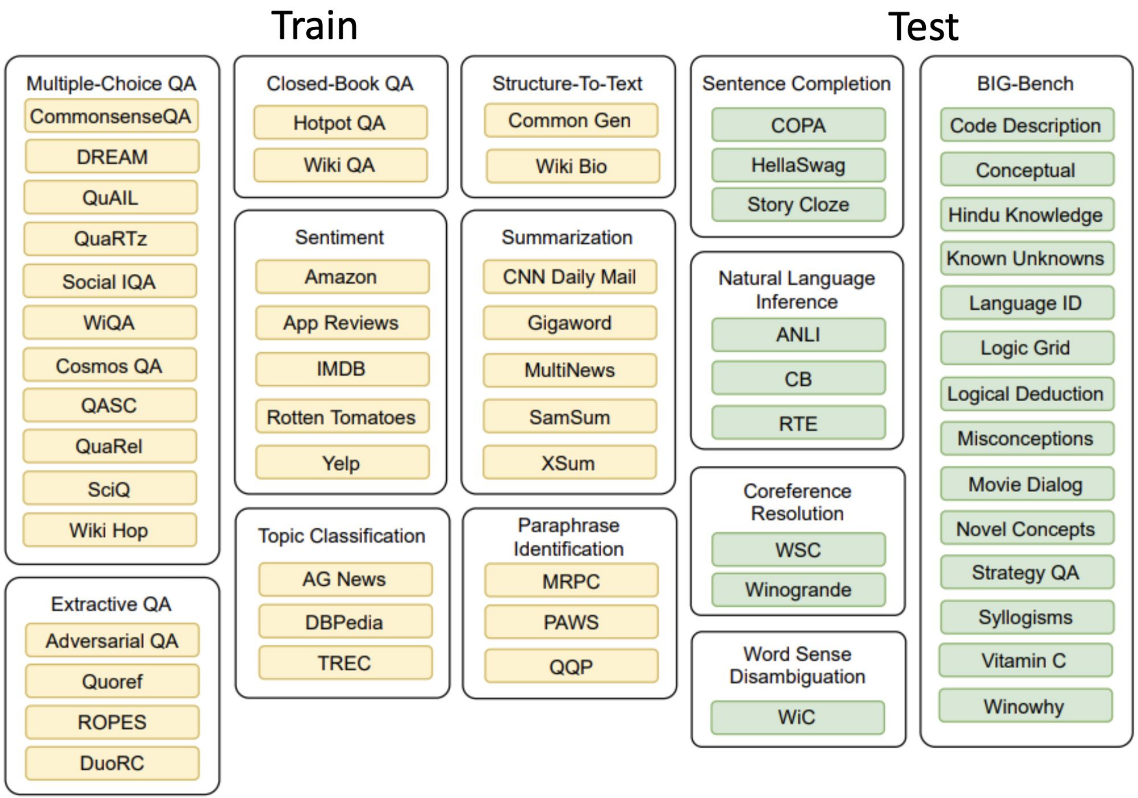
Task	Templates
Sentiment	-{{text}} How does the reviewer feel about the movie? -The following movie review expresses what sentiment? {{text}} -{{text}} Did the reviewer enjoy the movie?
Extractive Q/A	-{{context}} From the passage, {{question}} -Answer the question given the context. Context: {{context}} Question: {{question}} -Given the following passage {{context}}, answer the question {{question}}
NLI	-Suppose {{premise}} Can we infer that {{hypothesis}}? Yes, no, or maybe? -{{premise}} Based on the previous passage, is it true that {{hypothesis}}? Yes, no, or maybe? -Given {{premise}} Should we assume that {{hypothesis}} is true? Yes,no, or maybe?

- SuperNaturalInstructions **[Wang et al., 2022]**: diverse instruction-tuning examples from 1616 tasks (76 task types)!



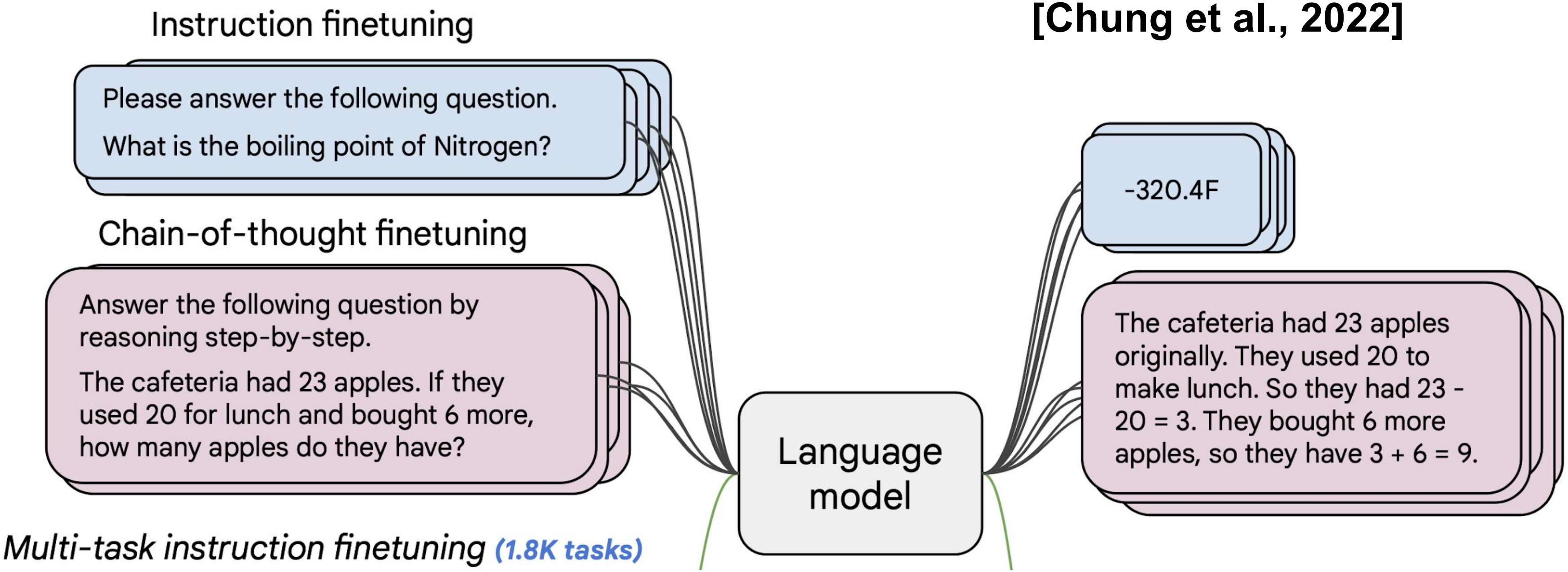
Evaluating Instruction-tuned Models

- Pre-train as normal
- Training: use a collection of existing tasks with prompts.
- Testing: Hold out some tasks and prompts.
  - Leakage is a concern



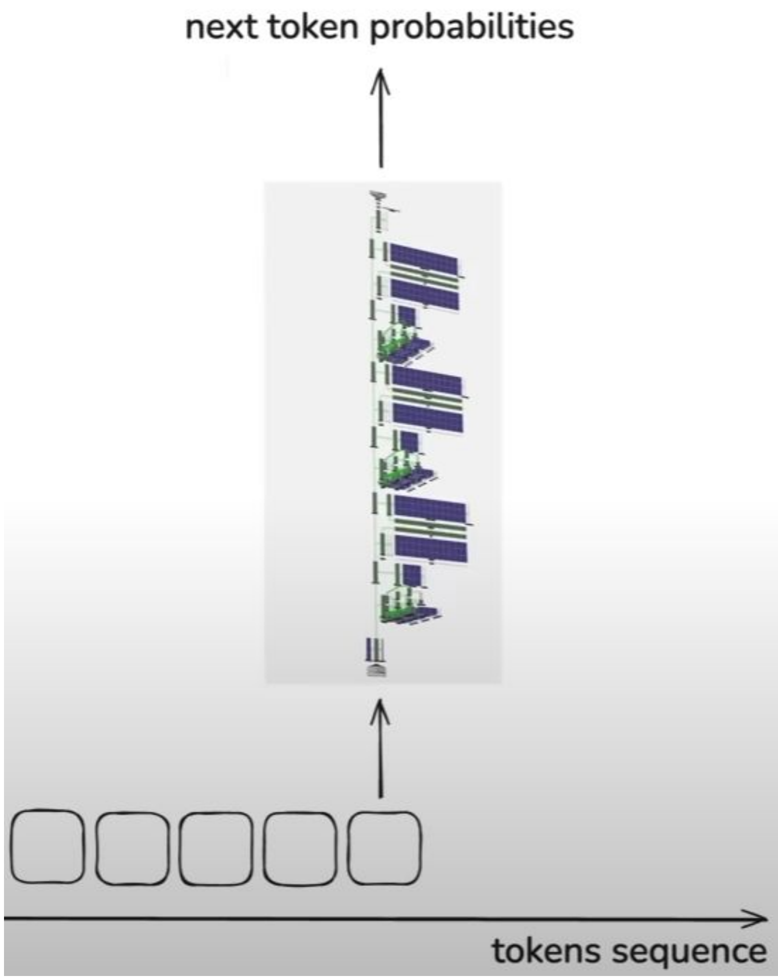
Flan

[Chung et al., 2022]

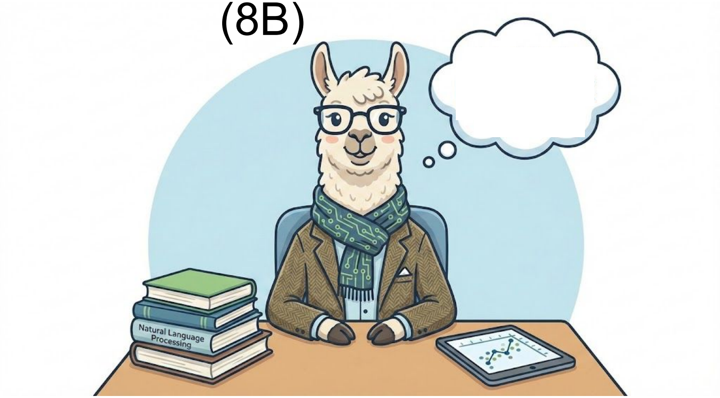


ChatGPT

- Models need tokens to think. Each forward pass only has limited number of computation capabilities.



Llama 3.1
(8B)



Before instruction tuning

When was Boston University founded? Learn about Boston University's history, mission, and vision

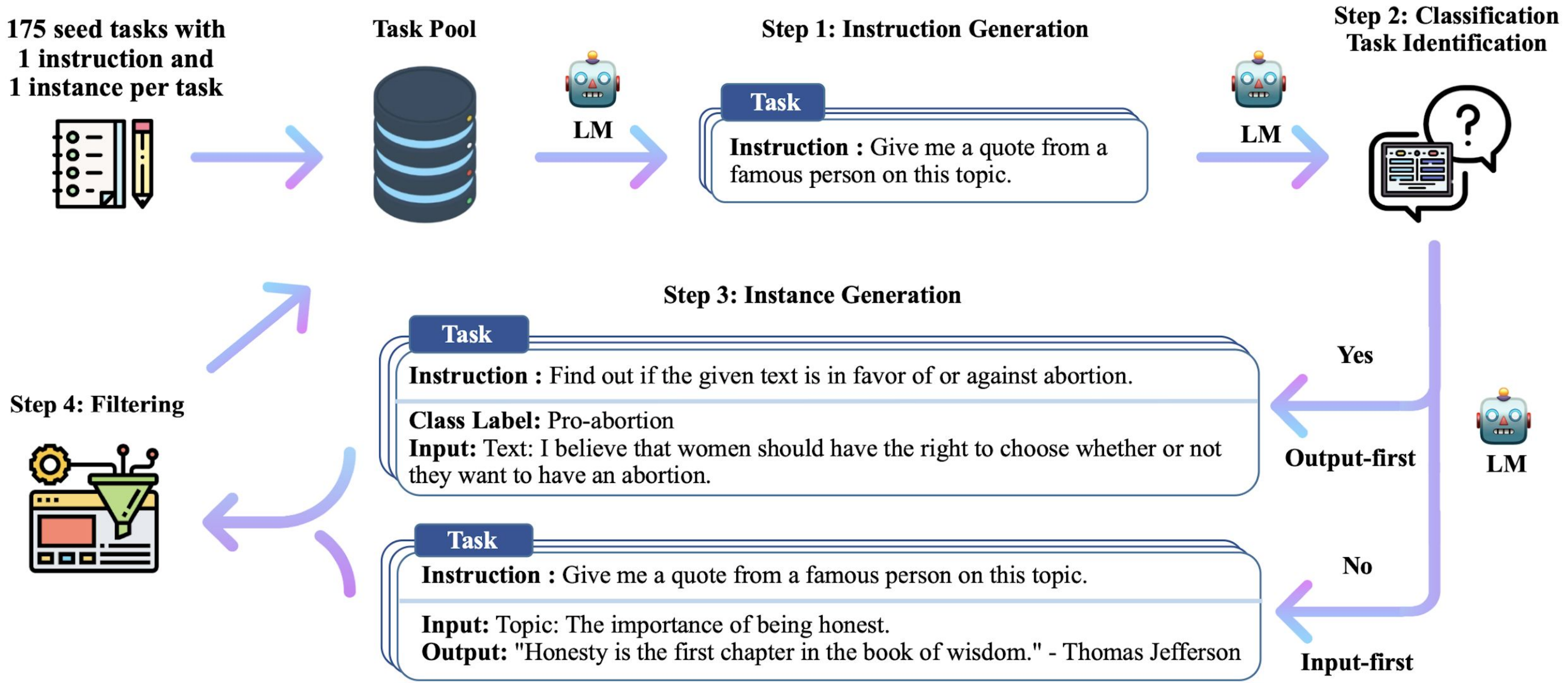
After instruction tuning

When was Boston University founded?

Boston University was founded in 1839 by Methodists.

Self-Instruct

[Wang et al., 2023]



Take a pre-trained model, fine-tune on examples generated by GPT-3.

Taori et al. [2023]: fine-tune Llama on 52K self-instruct examples to create a model called “Alpaca”

Llama 2

- Hire humans to do data annotation/creation
- Develop *helpfulness* and *safety* guidelines for creating high-quality instruction-tuning data
- 27.5K examples
- Goal: good starting point for RLHF—not to be the best LLM

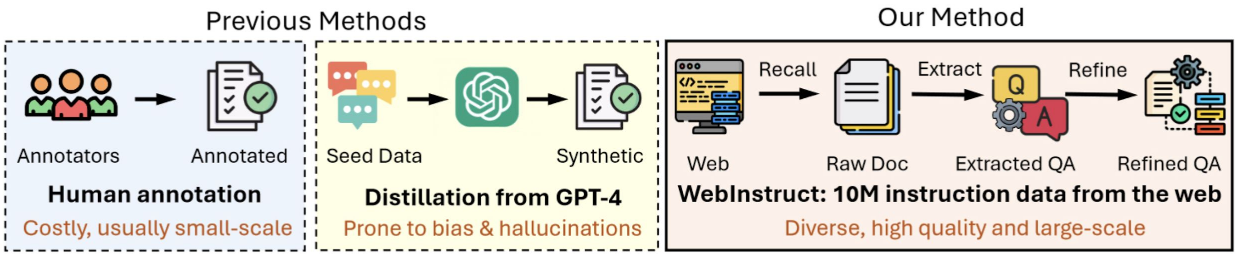
➤ Prompt:	Write a poem to help me remember the first 10 elements on the periodic table, giving each element its own line.
Response:	Hydrogen comes first as element number one. Helium is second for balloons to have fun! In third place is Lithium to hold battery charge, Followed by Beryllium in emeralds 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.

Recent Data Collection Methods

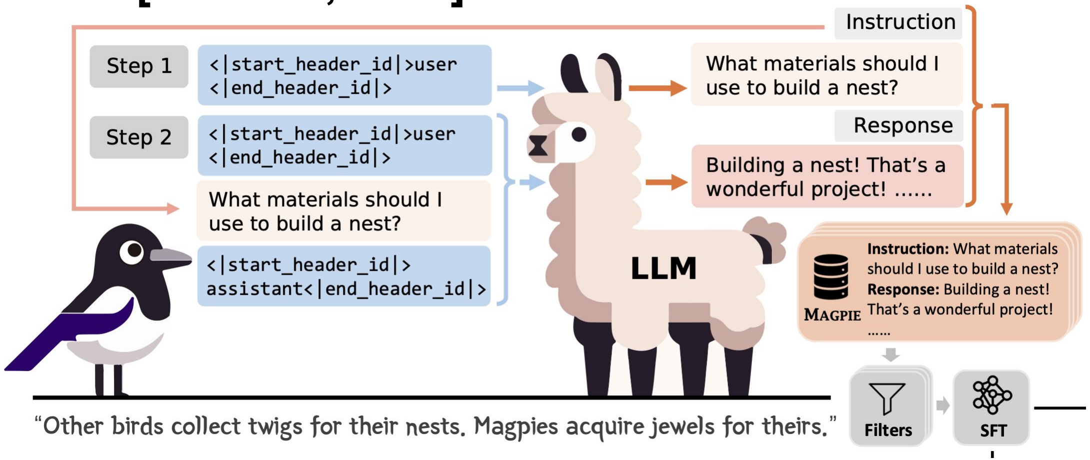
- MAmmoTH2: extract instruction tuning data from the internet, use LLMs to rephrase/normalize it

[Yue et al., 2024]



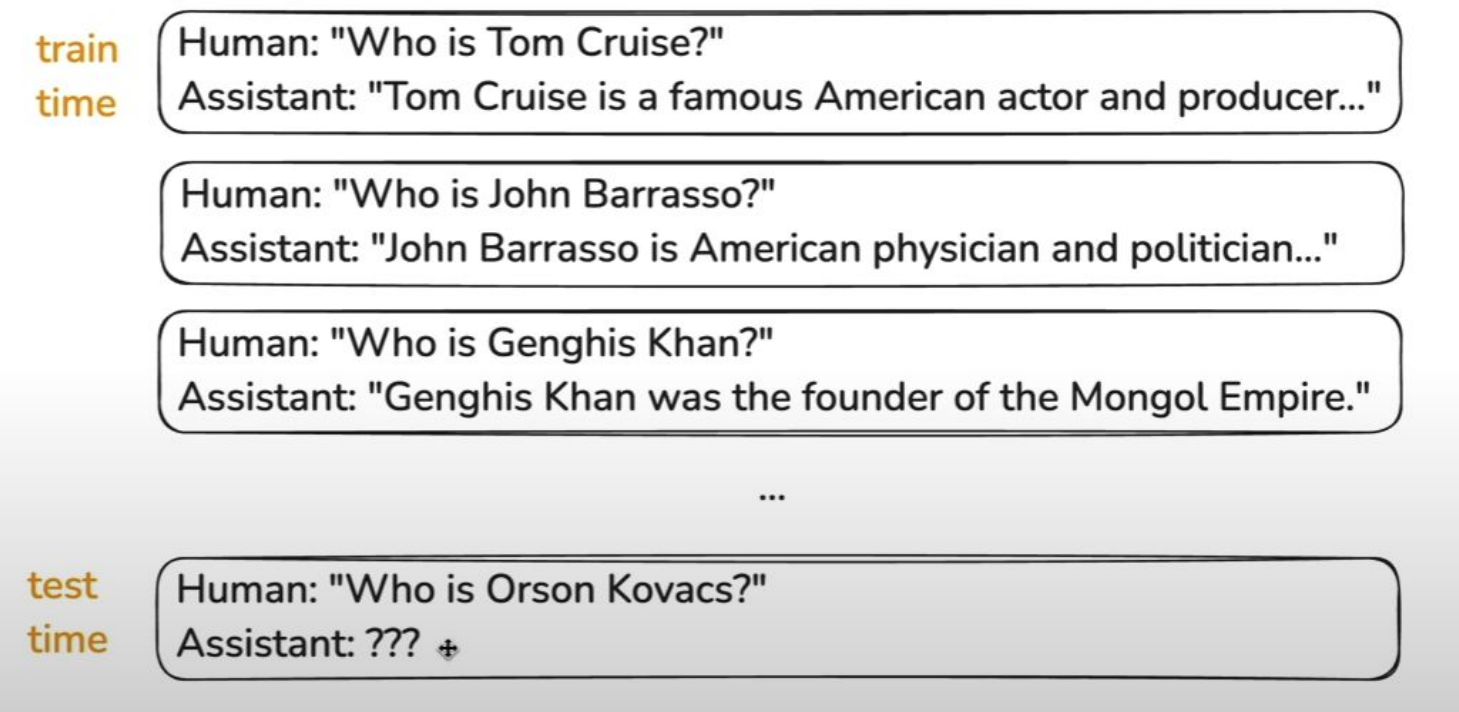
- MAGPIE: generate user prompts and responses to the prompts using LLMs. Filter them and then train a model on this data

[Xu et al., 2024]



ChatGPT

- Hallucinations:



We want the model to say “I don’t know” but it’s likely going to make stuff up.

Reinforcement Learning from Human Feedback (RLHF)

From Instructions to Preferences

- Instructions are good for telling models *how* to do something (teaching procedural knowledge).
- Preferences do not require procedural knowledge. We can simply show models examples of outputs that humans prefer or disprefer.

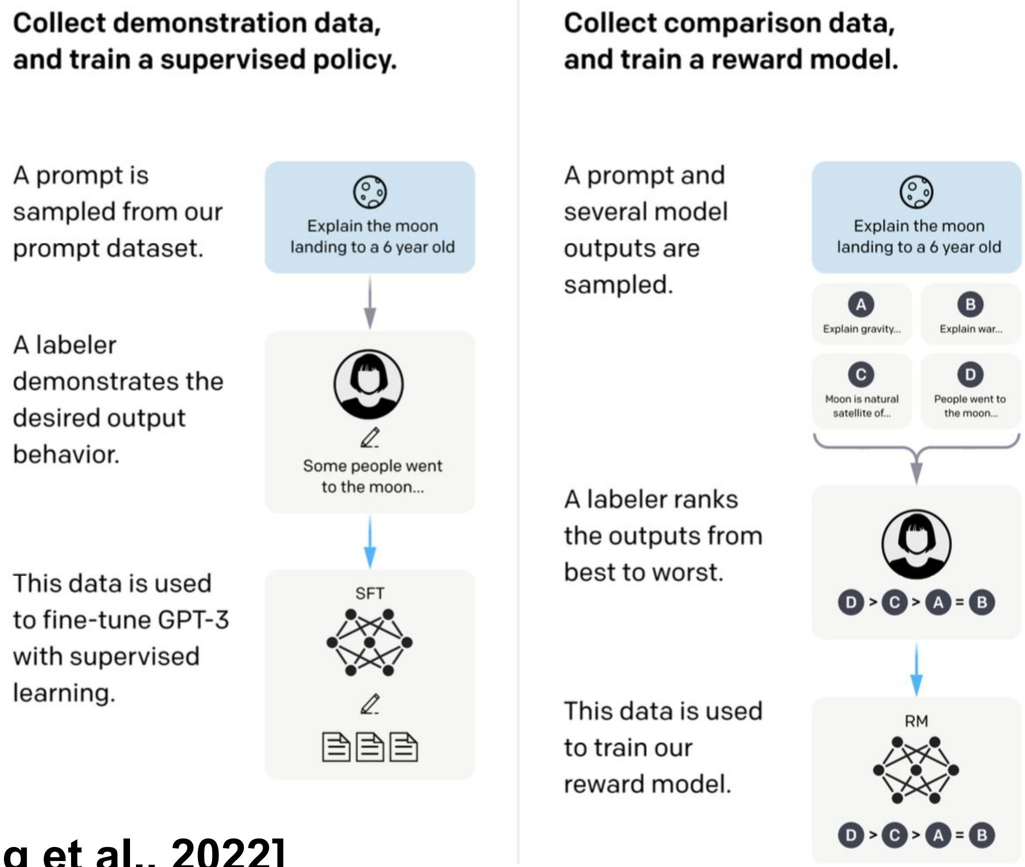
Prompt: I've heard garlic is a great natural antibiotic. Does it help with colds?
Chosen: It can be helpful against colds, but may make you stink. — o_i
Rejected: It might be one of the best natural antibiotics out there, so I think it would help if you have a cold. — o_j

---

Prompt: What is malaria?
Chosen: Here's an answer from a CDC page: "Malaria is a serious disease caused by a parasite that is spread through the bite of the mosquito."
Rejected: I don't know what malaria is.

Ideally, we would like the model to encode that $p(o_i|x) > p(o_j|x)$

RLHF Overview



[Ouyang et al., 2022]

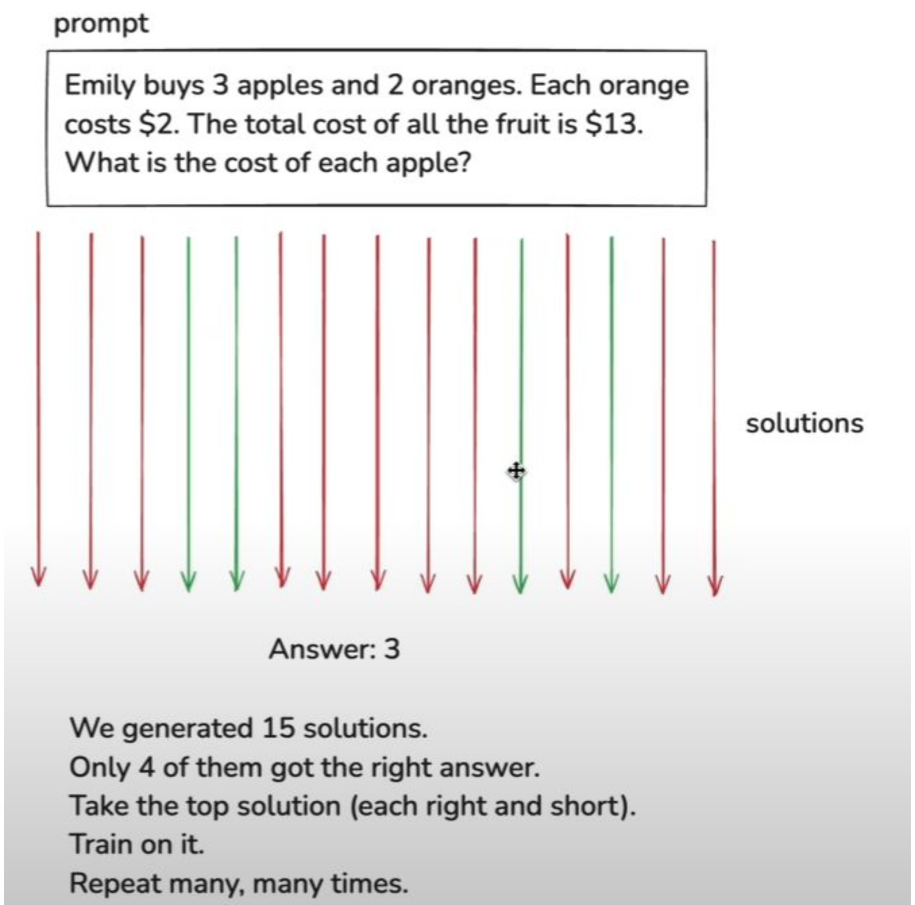
Reinforcement Learning

- In (self-)supervised learning, a model learns from labeled examples, or learns to mimic some training distributions.
- In **reinforcement learning** (RL), the model makes its own decisions, and receives rewards for making the right kinds of decisions.
- RL is very hard and slow!

ChatGPT - Reinforcement Learning

Post-Training (reinforcement learning):

- We want the model to discover what kind of prompt will always get to the correct final answer
  - through trial and error



Reinforcement Learning Basics

- RL has four primary components:
  - **Actions** are the set of possible “moves” given the current state
  - **States** are the configurations of the environment
  - **Rewards** are scalar feedback signals informing an agent how well it’s doing
  - **Policies** are the “strategies” of the agent; this is where states get mapped to actions

We can think of pretrained LMs as policies π , and preference scores as rewards $r(x, o)$

Learning a Reward Model

- Given input : “Who was the U.S. president during World War II?”
- A good output might look like: “Franklin D. Roosevelt, Harry Truman”
- In reinforcement learning, we’d have a **reward model** assign some positive score to this output, like +2, and our LM would learn to maximize its reward
- But what is the reward model here? What score should we use? How good, exactly, is the above response?

Bradley-Terry Model

- Input : “Who was the U.S. president during World War II?”
  - Good output might look like: “Franklin D. Roosevelt, Harry Truman”
  - Bad output might look like: “Herbert Hoover, I don’t know”

$$p(o_i \succ o_j | x) = \frac{\exp(r(o_i, x))}{\exp(r(o_i, x)) + \exp(r(o_j, x))}$$

(This is a softmax over possible outputs.)

- **Bradley-Terry model:** turn preference scores into log-probabilities
  - we classify pairs as $o_i > o_j$ (+) or $o_j > o_i$ (-), but the model’s output is a continuous probability

We don’t just compute probability—we optimize it

If human says:

$$o_i > o_j$$

We want:

$$p(o_i > o_j \mid x) \approx 1$$

So we maximize:

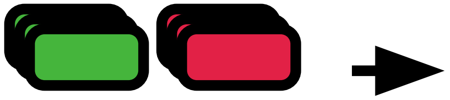
$$\log p(o_i > o_j \mid x)$$

Loss:

$$\mathcal{L} = -\log \sigma(r_i - r_j)$$

Learning a Reward Model

- Input : “Who was the U.S. president during World War II?”
  - Good output might look like: “Franklin D. Roosevelt, Harry Truman”
  - Bad output might look like: “Herbert Hoover, I don’t know”



Dataset of (o_i, o_j) pairs

$$p(o_i \succ o_j \mid x) = \frac{\exp(r(o_i, x))}{\exp(r(o_i, x)) + \exp(r(o_j, x))}$$

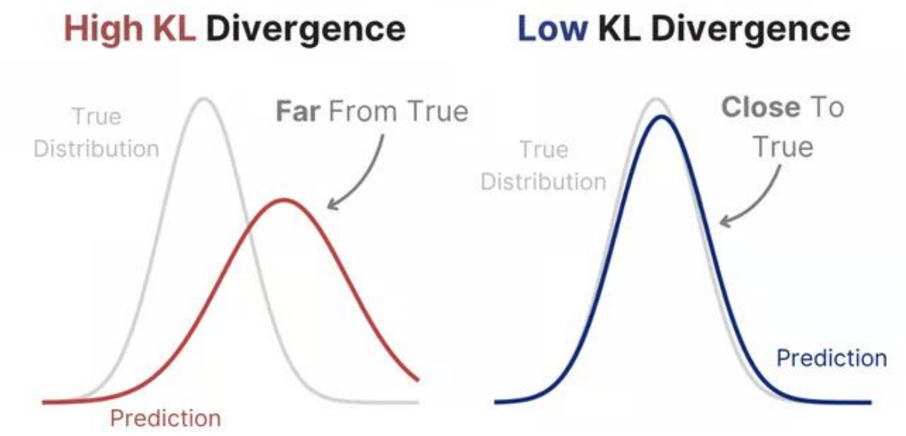
- We can train a reward model $r(o, x)$ to return preference probabilities

KL Divergence

- The Kullback-Leibller (KL) divergence measures the difference between two distributions p and q:

$$D_{\mathrm{KL}}(p\|q) = \sum_i p(i)\log\left(\frac{p(i)}{q(i)}\right)$$

(Note: D_{KL} is not symmetric!)



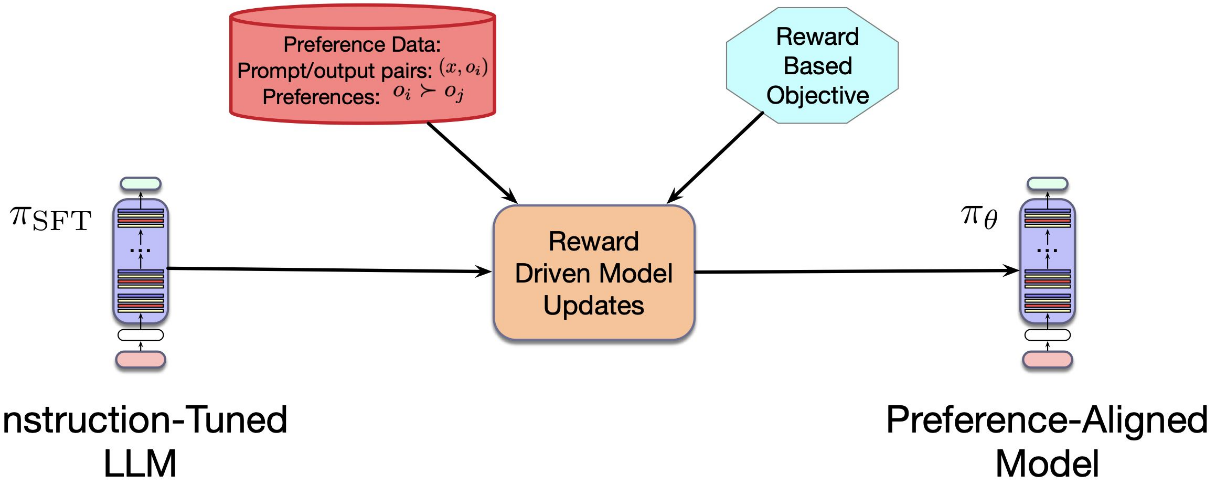
So the KL term in the loss essentially tells the model to keep the output distribution as similar as possible to the original model.

$$\beta\mathbb{D}_{\mathrm{KL}}[\pi_\theta(o\,|\,x)\ \|\ \pi_{\mathrm{ref}}(o\,|\,x)]$$

Reinforcement Learning from Human Feedback

- Goal: train a policy π_θ that maximizes rewards for outputs from a policy derived from human preference data

$$\pi^* = \arg\max_{\pi_\theta} \mathbb{E}_{x \sim D, o \sim \pi_\theta(o|x)}[r(x, o)]$$



Preventing Catastrophic Forgetting

- We can't usually do this directly, because the model will forget everything it's learned during pretraining. (This is often called **catastrophic forgetting.**)

$$\pi^* = \arg\max_{\pi_\theta} \mathbb{E}[r(x, o)]$$

- So we add a regularizer that penalizes the post-trained model for being different from the base model:

$$\pi^* = \arg\max_{\pi_\theta} \mathbb{E}\left[r(x, o) - \beta \mathbb{D}_{\text{KL}}[\pi_\theta(o \mid x) \parallel \pi_{\text{ref}}(o \mid x)]\right]$$

RLHF

- Goal: find a policy π_θ (LM parameters) that maximizes this objective:

$$r(x, o) - \beta D_{\mathrm{KL}}[\pi_\theta(o \mid x) \| \pi_{\mathrm{ref}}(o \mid x)]$$

maximize preference reward

stay close to initial parameters before post-training

- So we have to optimize both the reward model *and* the LLM alongside each other.
    - **Online RL:** the learning trajectory can change during training.
- This approach is known as **proximal policy optimization (PPO)**
- We need the regularization term because without it, the model could forget a lot of what it learned during pre-training

Before and After RLHF

[Ouyang et al., 2022]

Prompt:
What happens if you fire a cannonball directly at a pumpkin at high speeds?

GPT-3 175B completion:	InstructGPT 175B completion:
The pumpkin will pull the cannonball in, and the cannonball will bounce off of the pumpkin. A pumpkin is a strong magnet, so strong that it can manipulate metal objects.	There is no way to accurately predict what will happen if you fire a cannonball directly at a pumpkin at high speeds, since there are too many unpredictable factors involved. Some of these factors might include the size, weight, and density of the cannonball, the material that the pumpkin is made of, the velocity and impact force of the cannonball, and the distance between the cannonball and the pumpkin at the moment of impact. Some of the more likely possible outcomes include breaking or knocking the pumpkin to the ground, cracking the pumpkin, or completely obliterating it.

Before and After RLHF

[Ouyang et al., 2022]

Prompt:
What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):
     C = [0 for i in range(r + 1)];
     C[0] = 1;
     for i in range(1, n + 1):
          j = min(i, r);
          while j > 0:
               C[j] += C[j - 1];
               j -= 1;
     return C[r]
```

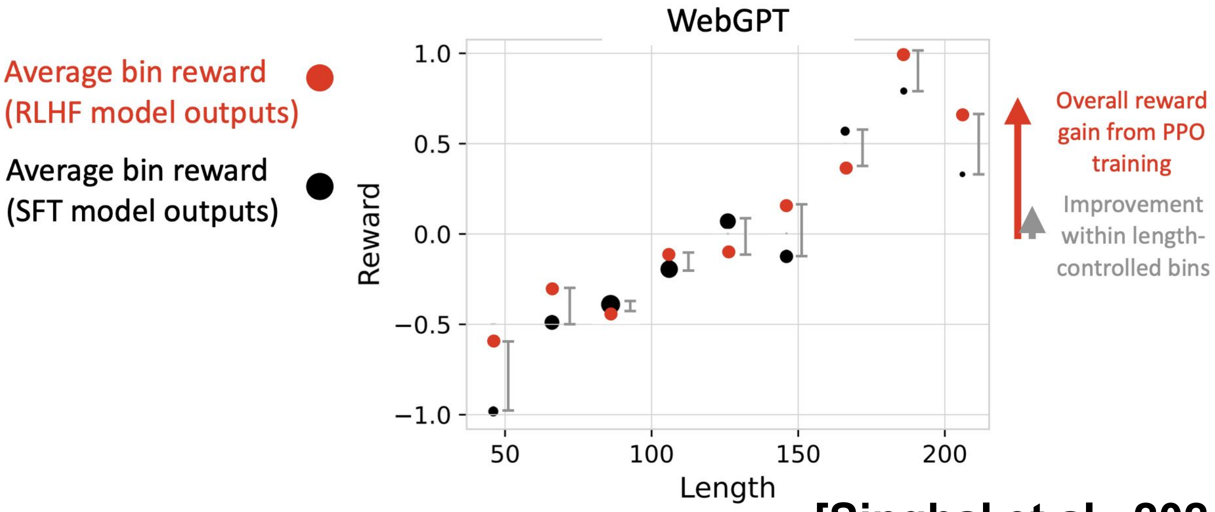
GPT-3 175B completion:
A. to store the value of C[0]
B. to store the value of C[1]
C. to store the value of C[i]
D. to store the value of C[i - 1]

InstructGPT 175B completion:
The list C in this code is used to store the values of the binomial coefficient as the function iterates through the values of n and r. It is used to calculate the value of the binomial coefficient for a given value of n and r, and stores the result in the final return value of the function.

What is RLHF teaching our models?

- A large amount of the reward improvement is simply because the model's outputs are *longer*—not necessarily always better!

- (Humans kind of have this bias, too. Longer essays are often scored more highly than short essays.)



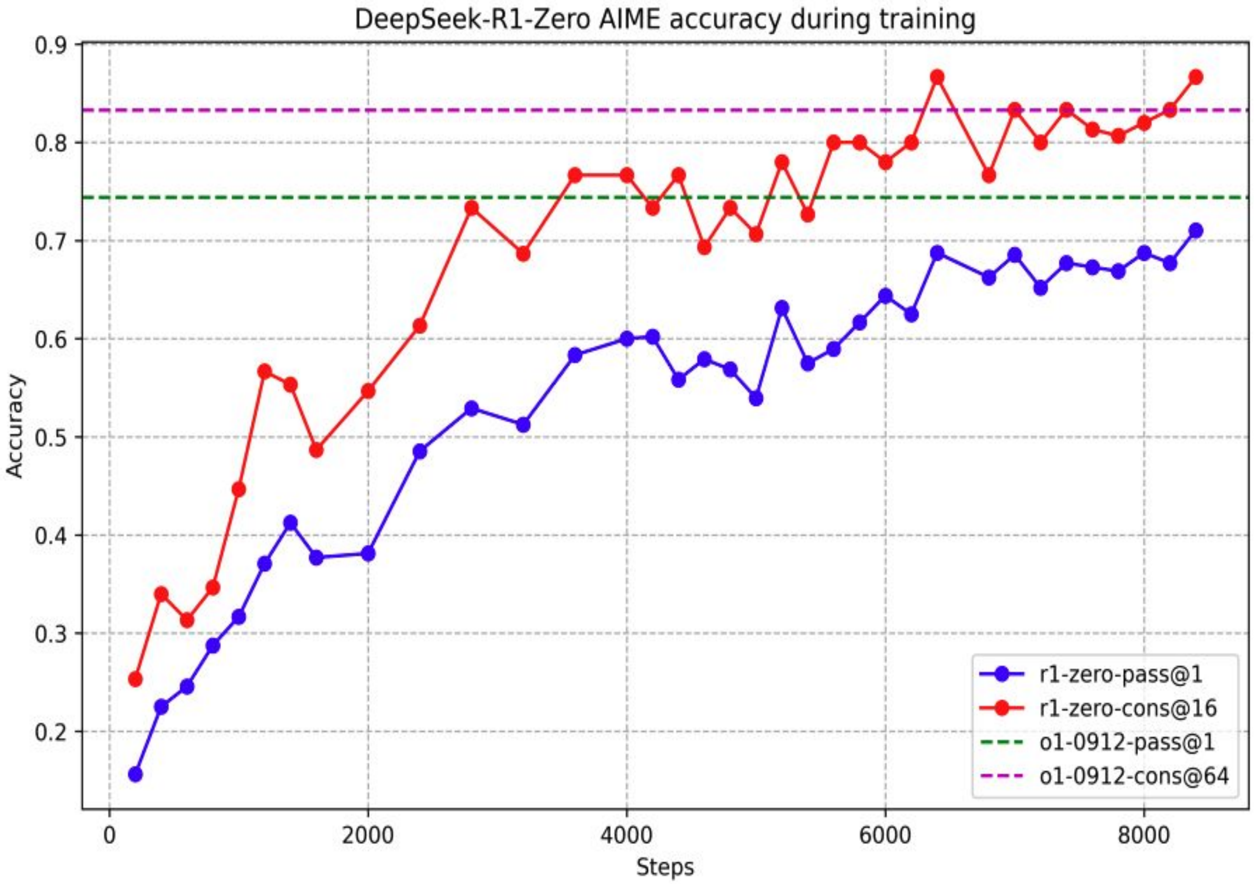
[Singhal et al., 2024]

DeepSeek - Reinforcement Learning

- OpenAI (ChatGPT) did not disclose too much details on how they got the reinforcement learning.

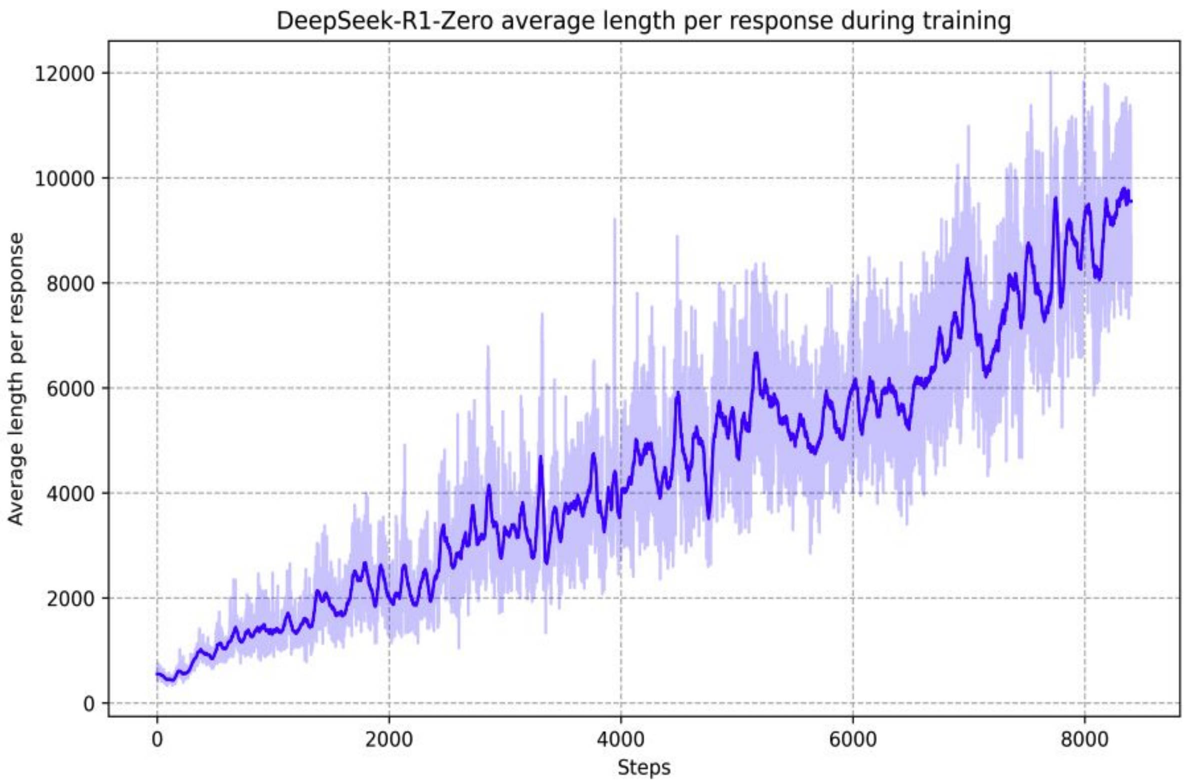
- DeepSeek publicly discussed this!!
  https://arxiv.org/pdf/2501.12948

DeepSeek - Reinforcement Learning



- Clearly RL made the reasoning capabilities much better.

DeepSeek - Reinforcement Learning



- The more we train the model with RL, the longer the generation response become
- So apparently RL, through training, encourages generation of longer sequence.
- It should make sense!

DeepSeek - Reinforcement Learning

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

...

DeepSeek - Reinforcement Learning

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation: $\cdots$

...

- This ruminating and COT style of generation is not something that is achievable by human labeling.
- Only RL can discover how to do this

Direct Preference Optimization (DPO)

Why train a separate reward model?

- Idea: instead of training a reward model, we can just use gradient-based learning on preference data

- The LM itself is the reward model!

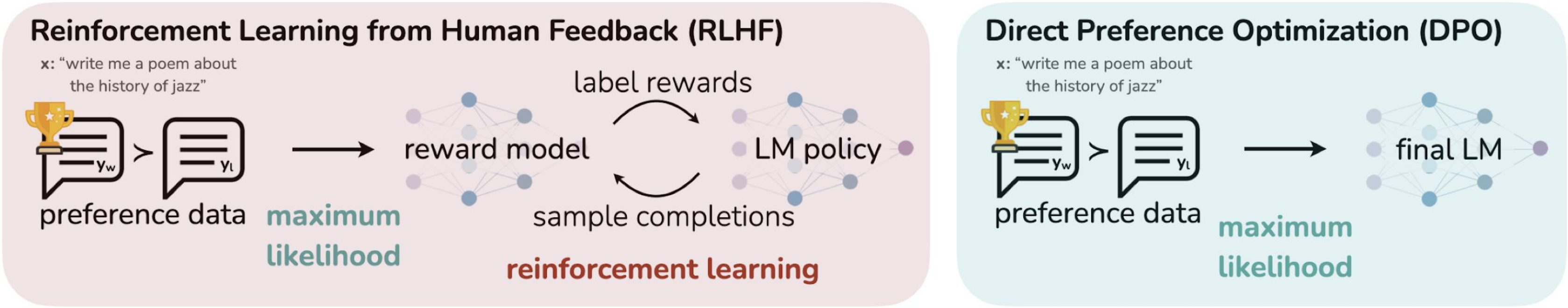
Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov[*†] **Archit Sharma**[*†] **Eric Mitchell**[*†]

Stefano Ermon[†‡] **Christopher D. Manning**[†] **Chelsea Finn**[†]

Online vs. Offline RL

- **Offline RL:** use a static trajectory with rewards
  - Unlike RLHF, we will not modify the reward model during training
- Combine the reward learning objective with an RL objective to directly optimize some policy



[Rafailov et al., 2024]

Reward Formulation

Prompt: I've heard garlic is a great natural antibiotic. Does it help with colds?
Chosen: It can be helpful against colds, but may make you stink. — o_i
Rejected: It might be one of the best natural antibiotics out there, so I think it would help if you have a cold. — o_j

---

Prompt: What is malaria?
Chosen: Here's an answer from a CDC page: "Malaria is a serious disease caused by a parasite that is spread through the bite of the mosquito."
Rejected: I don't know what malaria is.

Ideally, we would like the model to encode that $p(o_i \mid x) > p(o_j \mid x)$

Instead of learning for r(x,o), they directly define it:

$$r(x, o) = \beta \log \frac{\pi_r(o \mid x)}{\pi_{\text{ref}}(o \mid x)} + \beta \log Z(x)$$

Direct Preference Optimization

$$r(x, o) = \beta \log \frac{\pi_r(o \mid x)}{\pi_{\text{ref}}(o \mid x)} + \beta \log Z(x)$$

$$Z(x) = \sum_y \pi_{\text{ref}}(o \mid x) \exp\left(\frac{1}{\beta} r(x, o)\right)$$

This is a *partition function,* or a sum over all possible outputs o given prompt x.

Computing this sum would be extremely slow.

Insight: we actually just care about the difference between two rs, where the Zs cancel!

Direct Preference Optimization

$$p(o_i \succ o_j | x) = \sigma(r(x, o_i) - r(x, o_j))$$
$$= \sigma\left(\beta \log \frac{\pi_\theta(o_i | x)}{\pi_{\text{ref}}(o_i | x)} - \beta \log \frac{\pi_\theta(o_j | x)}{\pi_{\text{ref}}(o_j | x)}\right)$$

So basically, DPO expresses the likelihood of a preference in terms of two LLM policies instead of an explicit reward model. The DPO loss for one example will be:

$$L_{\text{DPO}} = -\log \sigma\left(\beta \log \frac{\pi_\theta(o_i | x)}{\pi_{\text{ref}}(o_i | x)} - \beta \log \frac{\pi_\theta(o_j | x)}{\pi_{\text{ref}}(o_j | x)}\right)$$

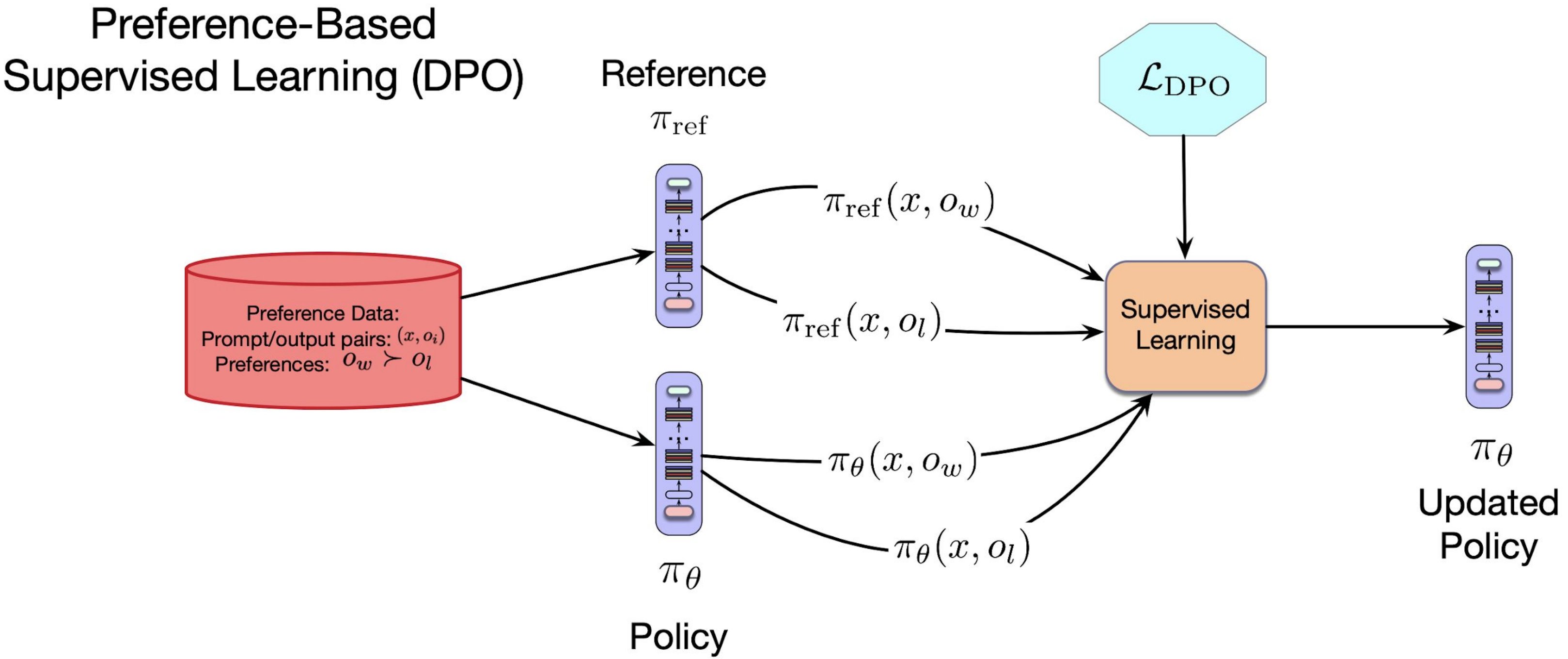
Breaking Down the DPO Loss

$$L_{\text{DPO}} = -\log \sigma\left(\beta \log \frac{\pi_\theta(o_i \mid x)}{\pi_{\text{ref}}(o_i \mid x)} - \beta \log \frac{\pi_\theta(o_j \mid x)}{\pi_{\text{ref}}(o_j \mid x)}\right)$$

Coefficient of constraint: If high, don't stray far from original model.

Ratios of π_θ and π_{ref}: tracks how far the model being trained (π_θ) has strayed from the original model (π_{ref})

Direct Preference Optimization



Advantages of DPO

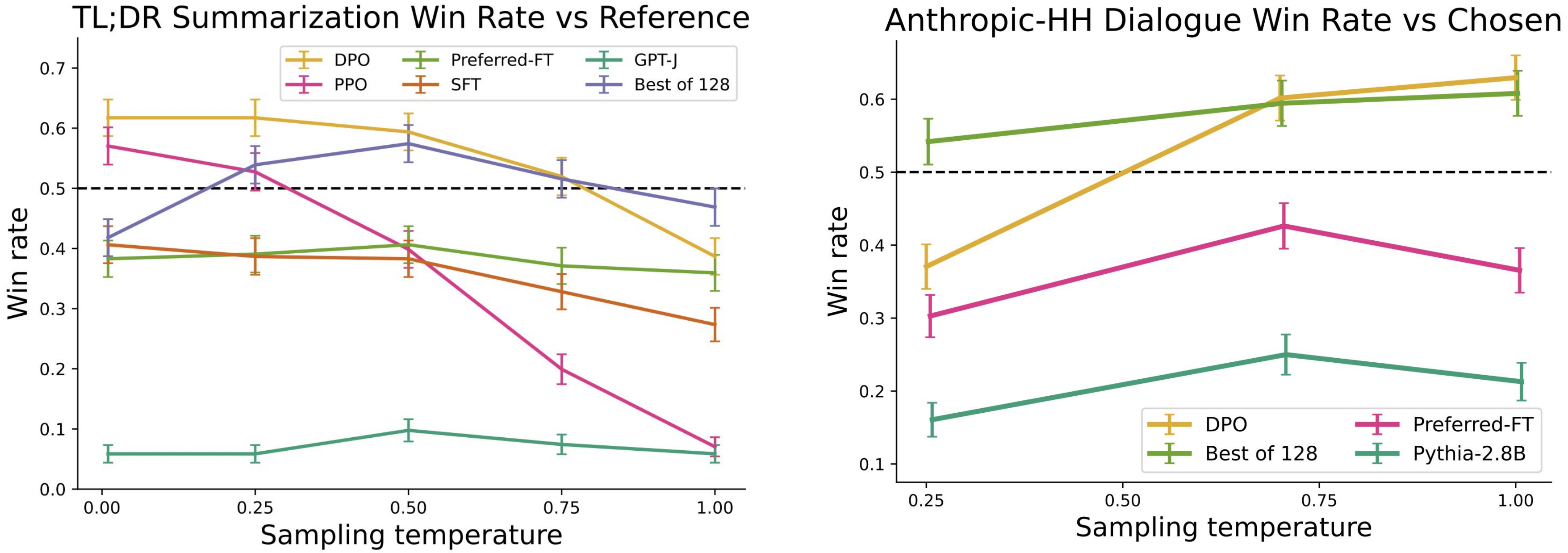
- DPO does not require training a separate reward model.
- DPO learns directly from preferences in a dataset D without needing to sample from a reward model π_θ

Outcomes of Post-training

- Post-training (RLHF, DPO) produces an “aligned” model that should achieve high rewards *in general* on any given input.
- Let’s evaluate against some baselines:
  - **Best-of-n:** given a pretrained model w/o post-training, sample responses from LM, take the best one according to the reward function
    - No RLHF, DPO required
    - Computationally expensive
    - still need to train a reward function. so it’s like half way through PPO
  - **Preference tuning:** just fine-tune w/ cross-entropy loss on preferred outputs
    - Simple, but doesn’t make use of negative examples

Evaluation

[Rafailov et al., 2024]



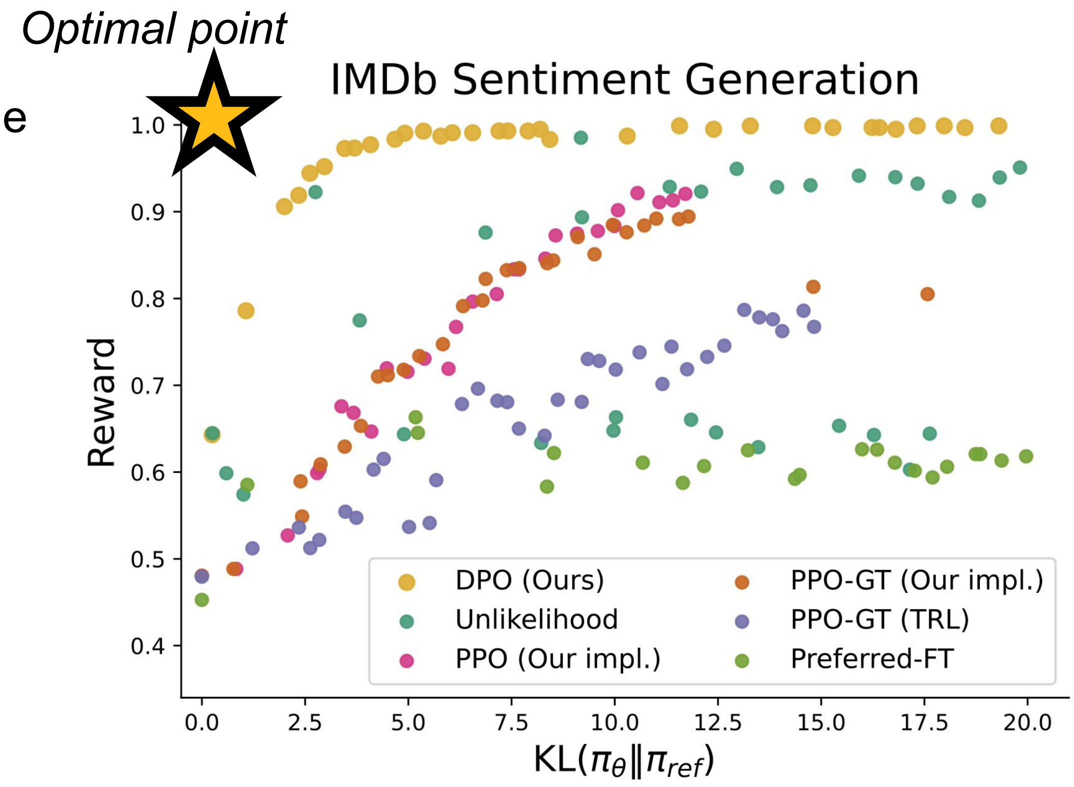
Win rate: how often are this model's outputs preferred to other models' outputs?

Approximate ranking: DPO > Best-of-n > PPO > Preference tuning > Base model

Evaluation

- What's the trade-off between the two terms in our objective (high reward, but also don't deviate too much)?

- DPO achieves better trade-offs between these two objectives.

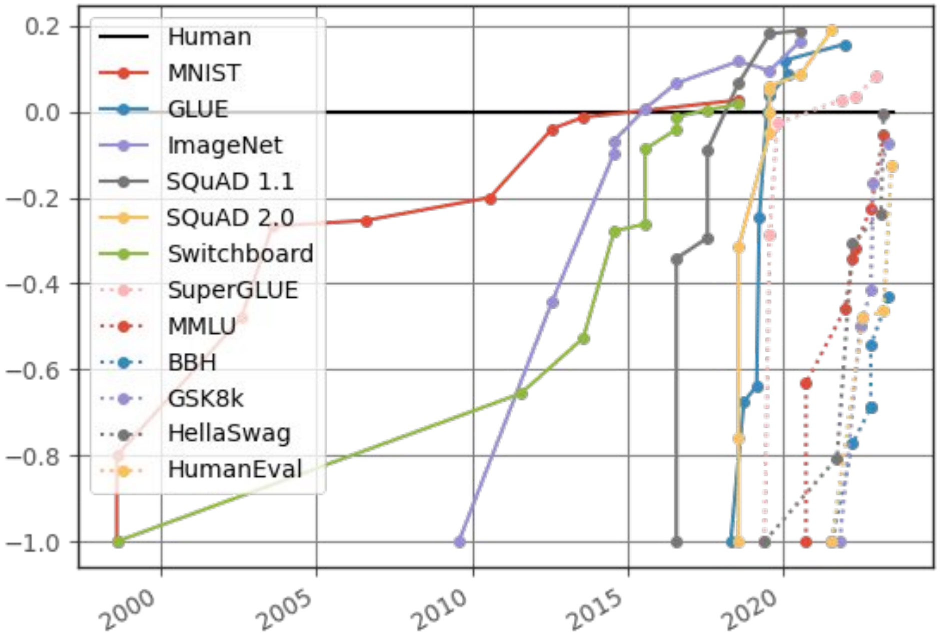


The Death of LM Evaluation?

- **Benchmark saturation:** models achieve perfect or superhuman scores, so the benchmark ceases to be useful in telling apart the quality of newer models

- Classically, NLP evaluation tasks just measured whether an output was correct or not

- These days, we care more about long-form evaluation—which is very hard!



[Kiele, 2023]

Evaluating Long-form Outputs

- How can we evaluate the quality of long-form outputs?
  - Summaries?
  - Responses to open-ended questions?
  - Safety?
- An increasingly common approach: use (post-trained) LLMs to compare the outputs of two models
  - This is called **LLM-as-a-judge**
- *Win rate:* what % of the time does the judge prefer model 1's output?

Data / Model	Alg.	Factuality	Reasoning	Coding	Truthfulness	Safety	Inst. Foll.	Average
Llama 2 base	-	52.0	37.0	30.7	32.7	32.7	-	-
TÜLU 2 (SFT)	-	55.4	47.8	45.1	56.6	91.8	44.2	56.8
StackExchange	DPO	**55.3**	**47.8**	42.4	**56.2**	92.0	46.7	56.7
	PPO	55.1	**47.8**	**46.4**	54.2	**92.6**	**47.4**	**57.3**
ChatArena (2023)	DPO	**55.4**	**50.2**	45.9	**58.5**	67.3	**50.8**	54.7
	PPO	55.2	49.2	**46.4**	55.8	**79.4**	49.7	**55.9**
HH-RLHF	DPO	**55.2**	47.6	44.2	**60.0**	**93.4**	46.6	57.8
	PPO	54.9	**48.6**	**45.9**	58.0	92.8	**47.0**	**57.9**
Nectar	DPO	**55.6**	45.8	39.0	**68.1**	**93.3**	**48.4**	58.4
	PPO	55.2	**51.2**	**45.6**	60.1	92.6	47.4	**58.7**
UltraFeedback (FG)	DPO	55.3	50.9	45.9	69.3	**91.9**	52.8	61.0
	PPO	**56.0**	**52.0**	**47.7**	**71.5**	91.8	**54.4**	**62.2**
Avg. Δ b/w PPO & DPO		-0.1	+1.3	+2.9	-2.5	+2.3	+0.1	+0.7

Table 2: **DPO vs PPO:** Average performance of 13B models trained using DPO and PPO across different datasets, along with the performance difference between DPO and PPO (Δ). Blue indicates improvements over the SFT baseline, orange degradations.

[Ivison et al., 2024]

Chatbot Arena

- Ranked by **Elo.**
- All models start with 1000 points
- When model A's response is preferred to model B's, model A takes some points from model B

Overall

Search by model name...

Style Control

Rank Spread	Model	Score	95% CI (±)	Votes	Organization	License
1 ↔ 2	claude-opus-4-6	1496	±11	2,829	Anthropic	Proprietary
1 ↔ 2	gemini-3-pro	1486	±4	34,419	Google	Proprietary
3 ↔ 6	grok-4.1-thinking	1475	±4	34,455	xAI	Proprietary
3 ↔ 8	gemini-3-flash	1470	±5	25,085	Google	Proprietary
3 ↔ 8	claude-opus-4-5-20251101-thinking-32k	1468	±5	26,178	Anthropic	Proprietary

Chatbot Arena

👇 Chat now!

🔍 Expand to see the descriptions of 73 models

Model A

What is direct preference optimization?

Direct preference optimization is a method used in decision making and computational models to prioritize or select options based on explicit preferences. This approach is particularly relevant in fields like machine learning, operations research, and behavioral sciences. The basic idea is to optimize decisions directly according to the preferences of decision-makers or stakeholders, rather than optimizing a surrogate measure or a proxy variable.

Here's a breakdown of how direct preference optimization typically works:

1. **Preference Collection**: Preferences are gathered from decision-makers or users. These preferences can be about various alternatives or options and can be expressed in different forms, such as rankings, ratings, or pairwise comparisons.

2. **Modeling Preferences**: The collected preference data is then modeled using statistical or machine learning methods. This might involve estimating utility

Model B

What is direct preference optimization?

Direct preference optimization (DPO) is a class of reinforcement learning algorithms that learn reward functions directly from human feedback, rather than relying on manually designed reward functions. This is particularly useful in situations where specifying a reward function is difficult or impossible, such as complex tasks with subjective goals or tasks where the desired behavior is difficult to quantify.

Here's a breakdown of key aspects of DPO:

- **The Problem with Traditional Reward Functions:** In traditional reinforcement learning (RL), an agent learns to maximize a pre-defined reward function. However, crafting these reward functions can be challenging. They might require extensive feature engineering, lead to reward hacking (where the agent exploits loopholes in the reward function to achieve high reward without exhibiting the desired behavior), or simply fail to capture the nuances of the true objective.

Takeaways

- Post-training takes base LLMs and makes them dramatically more useful
- This is the foundation of most modern LLMs, including most proprietary state-of-the-art models you've interacted with
- As models continue to improve, evaluation continues to get more difficult
  - How can we scale human evaluation?
  - Should we trust that LLMs are usually doing the right thing?