

TDM	729.89	915.51	185.62	▲25.43%	FLR	660.27	745.28	85.01	▲12.88%
HUM	749.73	924.29	174.56	▲23.28%	UVD	155.59	181.57	25.98	▲16.70%
DMW	833.72	1004.01	170.29	▲20.43%	QUV	440.55	540.21	99.66	▲22.62%
YZJ	903.49	1127.46	223.97	▲24.79%	HZT	285.51	344.98	59.47	▲20.83%
GLY	982.07	1219.39	237.32	▲24.17%	PCW	811.44	1029.66	218.22	▲26.89%
VDA	113.74	143.41	29.67	▲26.09%	AIK	361.77	451.39	89.62	▲24.77%
UVV	468.08	535.41	67.33	▲14.38%	ZJJ	858.36	994.57	136.21	▲15.87%
HJS	545.49	659.05	113.56	▲20.82%	RHJ	894.79	1046.68	151.89	▲16.97%
EGC	566.96	664.69	97.73	▲17.24%	VOV	425.08	509.95	84.87	▲19.97%

PPJ	912.63	1038.36	125.73	▲13.78%	ZBK	391.59	491.48	99.89	▲25.51%
UAQ	1309.55	1655.62	346.07	▲26.43%	BNY	969.21	1130.65	161.44	▲16.66%
DAQ	1295.17	1641.66	346.49	▲26.75%	SDM	735.44	913.39	177.95	▲24.20%
PNR	654.33	775.84	121.51	▲18.57%	TQQ	1323.91	1646.42	322.51	▲24.36%
ZTM	181.89	220.52	38.63	▲21.55%	OIS	543.42	667.24	123.82	▲22.79%
...	...	...	...	...	...	...	...	...	...

# GPT-1 and GPT-2 Review

Click to add text

Amin Saied

2021-01-15



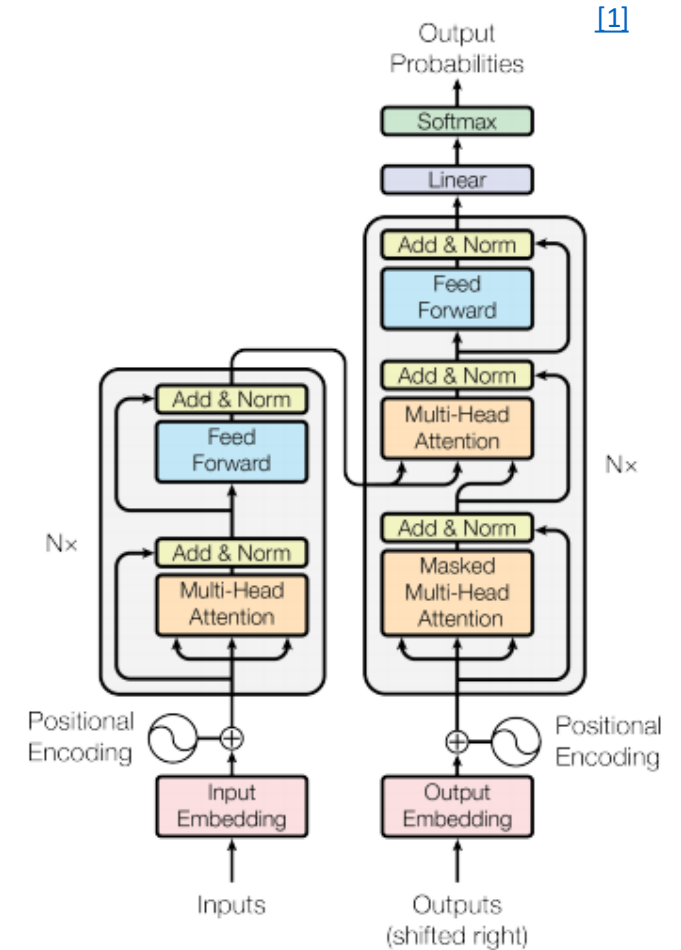


Background



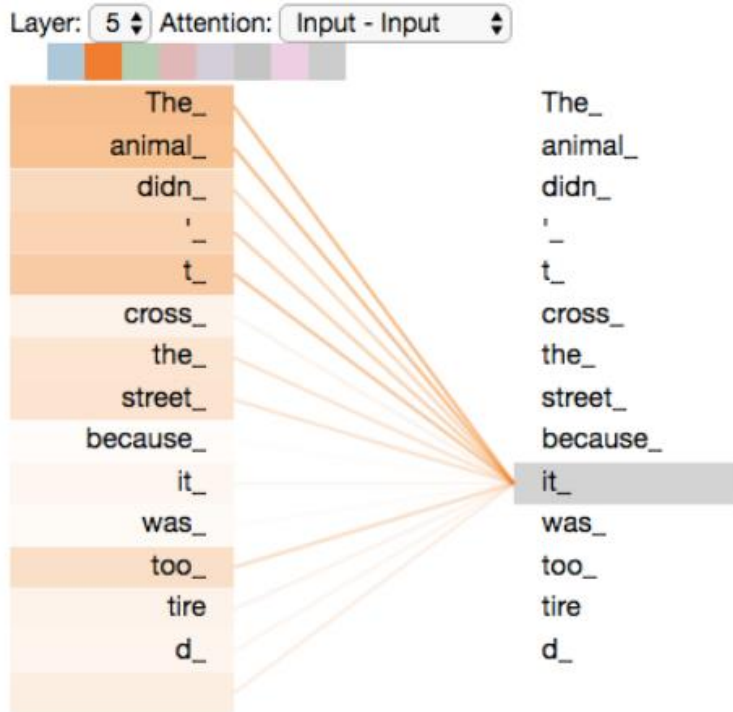
# Transformers: Life before GPT-1

- Sequence-to-sequence model
- Evolution of RNNs
- Review:
  - Self-attention
  - Multi-headed attention
  - Encoder/decoder



# Breaking down self-attention

"The animal didn't cross the street because it was too tired"



$$Z := \text{softmax} \left( \frac{Q \cdot K^T}{\sqrt{d}} \right) \cdot V$$

Input

Embedding

Queries

Keys

Values

Score

Divide by  $8 (\sqrt{d_k})$

Softmax

Softmax

X

Value

Sum

Thinking

$x_1$  [ ] [ ] [ ] [ ]

$q_1$  [ ] [ ] [ ]

$k_1$  [ ] [ ] [ ]

$v_1$  [ ] [ ] [ ]

$q_1 \cdot k_1 = 112$

14

0.88

$v_1$  [ ] [ ] [ ]

$z_1$  [ ] [ ] [ ]

Machines

$x_2$  [ ] [ ] [ ] [ ]

$q_2$  [ ] [ ] [ ]

$k_2$  [ ] [ ] [ ]

$v_2$  [ ] [ ] [ ]

$q_1 \cdot k_2 = 96$

12

0.12

$v_2$  [ ] [ ] [ ]

$z_2$  [ ] [ ] [ ]

# Breaking down multi-headed self-attention

1) This is our input sentence\*

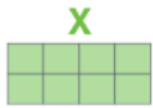
2) We embed each word\*

3) Split into 8 heads. We multiply  $X$  or  $R$  with weight matrices

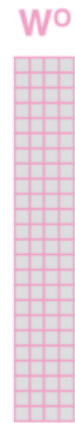
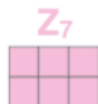
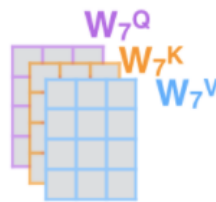
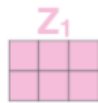
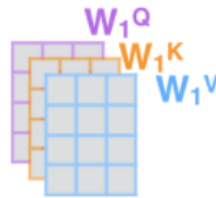
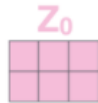
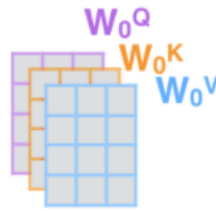
4) Calculate attention using the resulting  $Q/K/V$  matrices

5) Concatenate the resulting  $Z$  matrices, then multiply with weight matrix  $W^O$  to produce the output of the layer

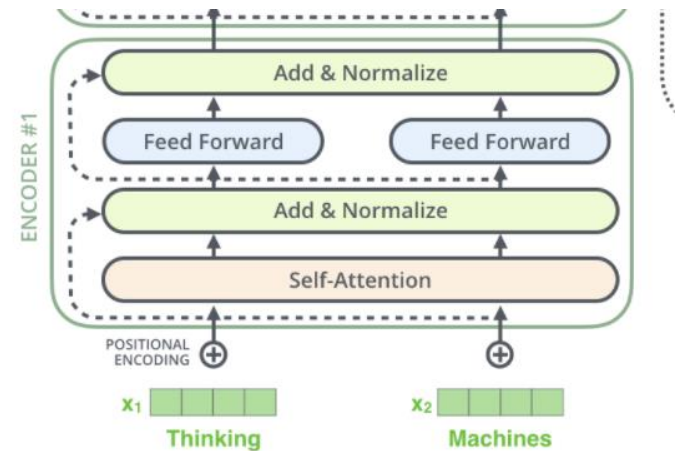
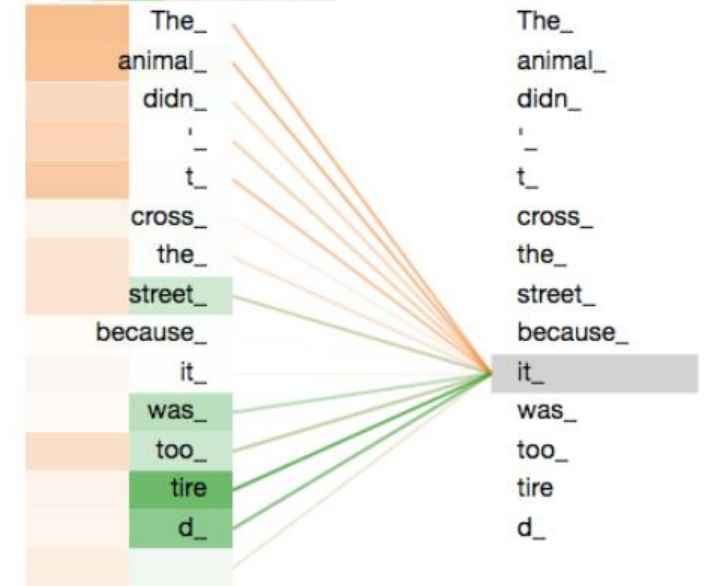
Thinking  
Machines



\* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one



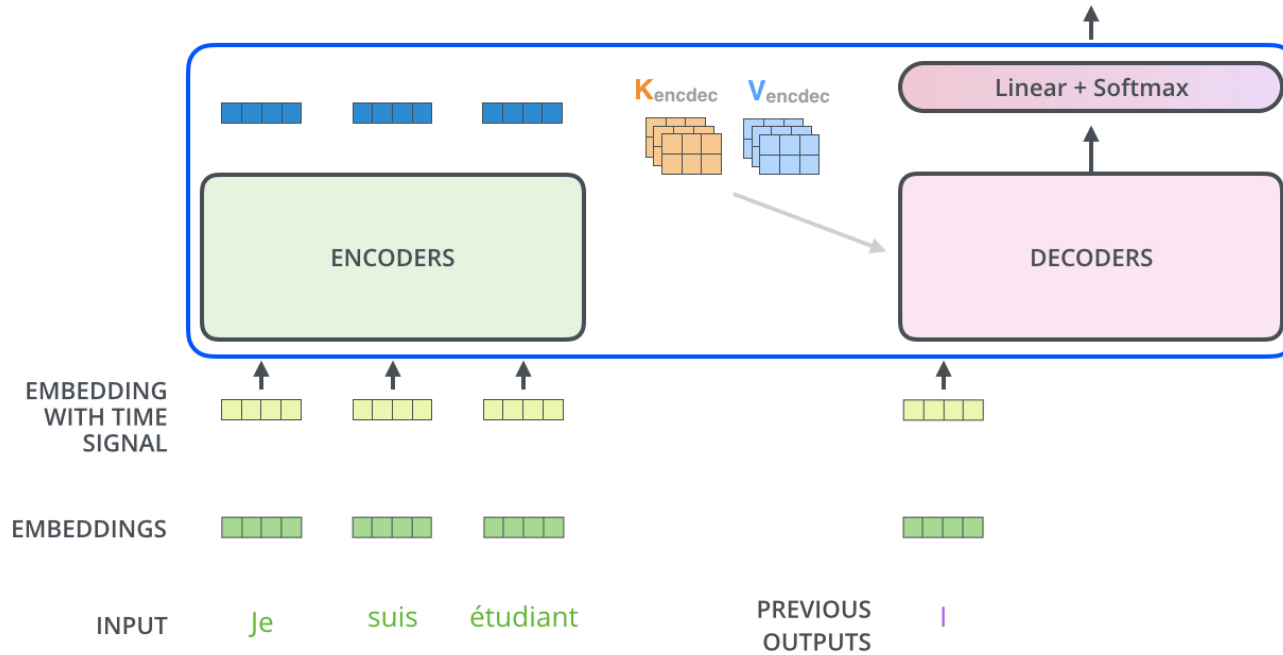
Layer: 5 Attention: Input - Input



# Encoder / decoder

Decoding time step: 1 2 3 4 5 6

OUTPUT |

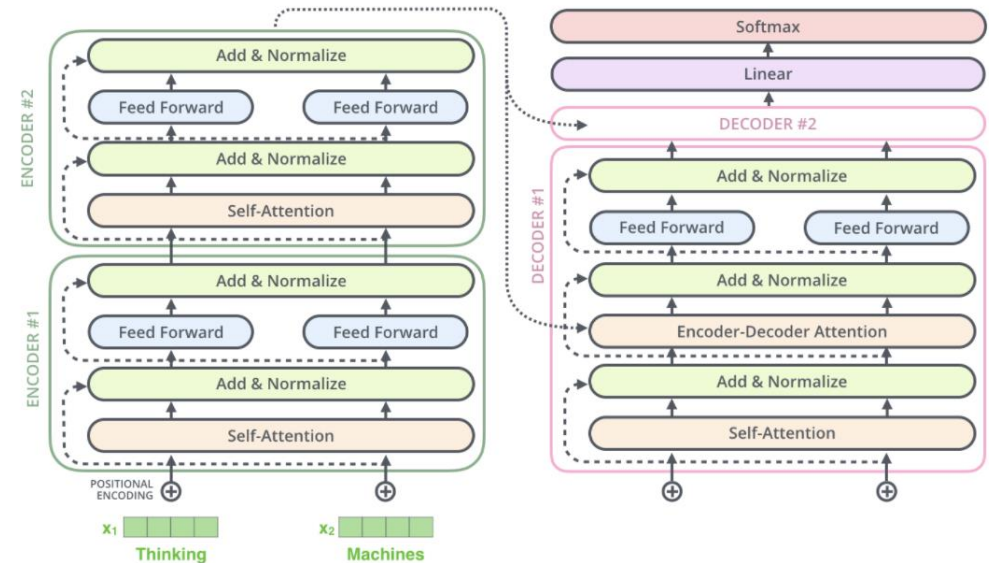


## Encoder

- Full-context
- (Contextual) Word-embeddings

## Decoder

- Left-context (masking)
- => predict next word



# Glossary

- Tokens
- Attention (self-attention, multi-headed)
- Transformer
- Encoder / decoder





GPT 1



# GPT-1

Improving Language Understanding by Generative Pre-Training – Radford et al

## Key takeaways

- Semi-supervised learning with transformers
  - Pretraining / finetuning
- Decoder-only architecture
- Simplified approach to transfer learning

=>

- SOTA in 9/12 tasks studied



# Language modelling (unsupervised approach)

## 3.1 Unsupervised pre-training

Given an unsupervised corpus of tokens  $\mathcal{U} = \{u_1, \dots, u_n\}$ , we use a standard language modeling objective to maximize the following likelihood:

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \Theta) \quad (1)$$

where  $k$  is the size of the context window, and the conditional probability  $P$  is modeled using a neural network with parameters  $\Theta$ . These parameters are trained using stochastic gradient descent [51].

- Different from above: this is unsupervised!
- Training data:
  - Data: Edon Lulzim Zhegrova (born 31 March 1999) is a Kosovan professional footballer who plays as a right winger for Swiss club Basel
  - Input: Edon Lulzim Zhegrova (born 31 March 1999) is a Kosovan professional
  - Output: Edon Lulzim Zhegrova (born 31 March 1999) is a Kosovan professional **footballer**
- Jargon: Auto-regressive language modelling
- Transfer learning in NLP!



# Decoder-only architecture

- Based on previous work [2] using decoder-only transformer to generate Wikipedia articles
- Key-insight [2]: convert seq-to-seq task into language modelling task
  - Seq-to-seq:  $(x_1, \dots, x_m) \mapsto (y_1, \dots, y_n)$
  - LM:  $(x_1, \dots, x_m, \delta, y_1, \dots, y_n)$ , where  $\delta$ =separator token

$$p(w^1, \dots, w^{n+\eta}) = \prod_{j=1}^{n+\eta} p(w^j | w^1, \dots, w^{j-1})$$

- [1]: Semi-supervised approach!

[2] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer. Generating wikipedia by summarizing long sequences. ICLR, 2018.

[1] Improving Language Understanding by Generative Pre-Training – Radford et al



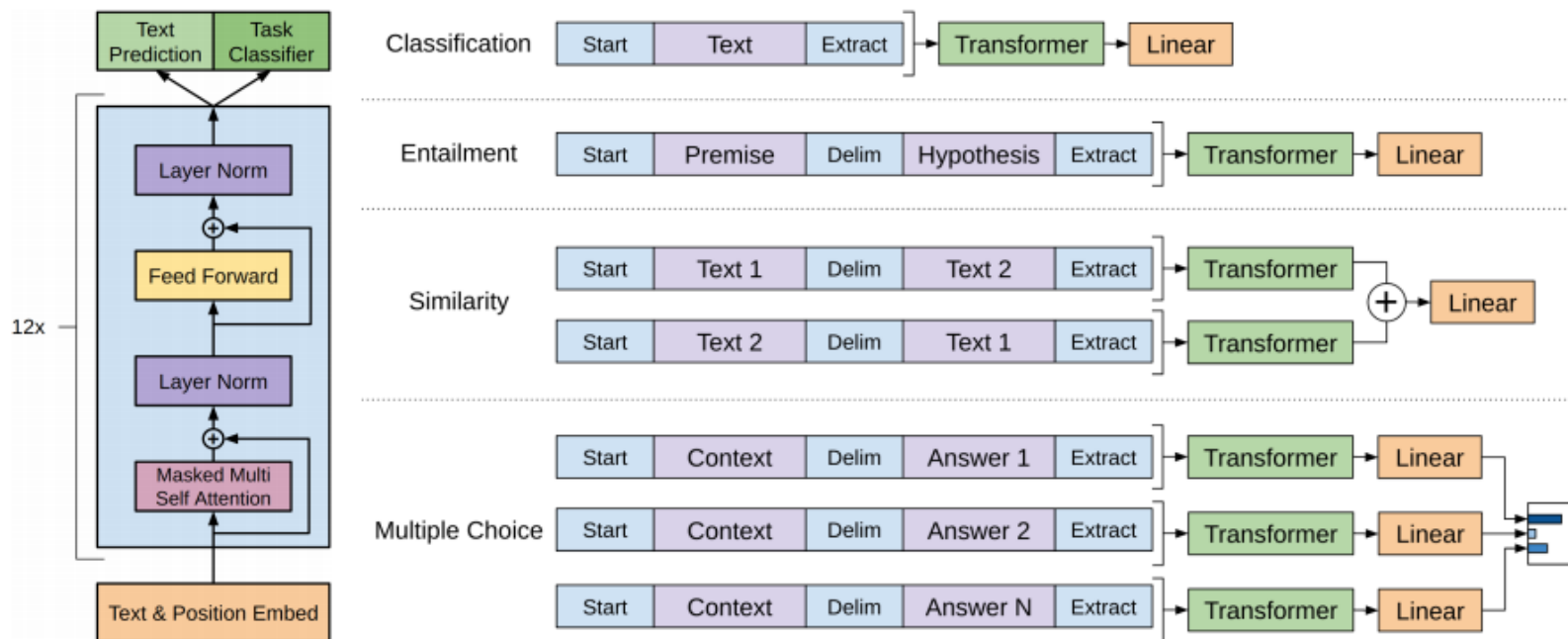
### 3.2 Supervised fine-tuning

After training the model with the objective in Eq. 1, we adapt the parameters to the supervised target task. We assume a labeled dataset  $\mathcal{C}$ , where each instance consists of a sequence of input tokens,  $x^1, \dots, x^m$ , along with a label  $y$ . The inputs are passed through our pre-trained model to obtain the final transformer block's activation  $h_i^m$ , which is then fed into an added linear output layer with parameters  $W_y$  to predict  $y$ :

$$P(y|x^1, \dots, x^m) = \text{softmax}(h_i^m W_y). \quad (3)$$

This gives us the following objective to maximize:

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \dots, x^m). \quad (4)$$



# Glossary

- Tokens
- Attention (self-attention, multi-headed)
- Transformer
- Encoder / decoder
- **Pretrain / Finetune**
- **Language modelling**
- **Auto-regressive**



# Experimental Results

**Model specifications** Our model largely follows the original transformer work [62]. We trained a 12-layer decoder-only transformer with masked self-attention heads (768 dimensional states and 12 attention heads). For the position-wise feed-forward networks, we used 3072 dimensional inner states. We used the Adam optimization scheme [27] with a max learning rate of  $2.5e-4$ . The learning rate was increased linearly from zero over the first 2000 updates and annealed to 0 using a cosine schedule. We train for 100 epochs on minibatches of 64 randomly sampled, contiguous sequences of 512 tokens. Since layernorm [2] is used extensively throughout the model, a simple weight initialization of  $N(0, 0.02)$  was sufficient. We used a bytepair encoding (BPE) vocabulary with 40,000 merges [53] and residual, embedding, and attention dropouts with a rate of 0.1 for regularization. We also employed a modified version of L2 regularization proposed in [37], with  $w = 0.01$  on all non bias or gain weights. For the activation function, we used the Gaussian Error Linear Unit (GELU) [18]. We used learned position embeddings instead of the sinusoidal version proposed in the original work. We use the *ffly* library<sup>2</sup> to clean the raw text in BooksCorpus, standardize some punctuation and whitespace, and use the *spaCy* tokenizer.<sup>3</sup>

- 12 layer decoder
- 768 dim hidden states
- 12 attention heads (multi-headed attention)

Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	<b>93.2</b>	-	-	-	-
TF-KLD [23]	-	-	<b>86.0</b>	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	<u>81.0</u>	-	-
Single-task BiLSTM + ELMo + Attn [64]	<u>35.0</u>	90.2	80.2	55.5	<u>66.1</u>	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	<u>68.9</u>
Finetuned Transformer LM (ours)	<b>45.4</b>	91.3	82.3	<b>82.0</b>	<b>70.3</b>	<b>72.8</b>

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

## Question Answering

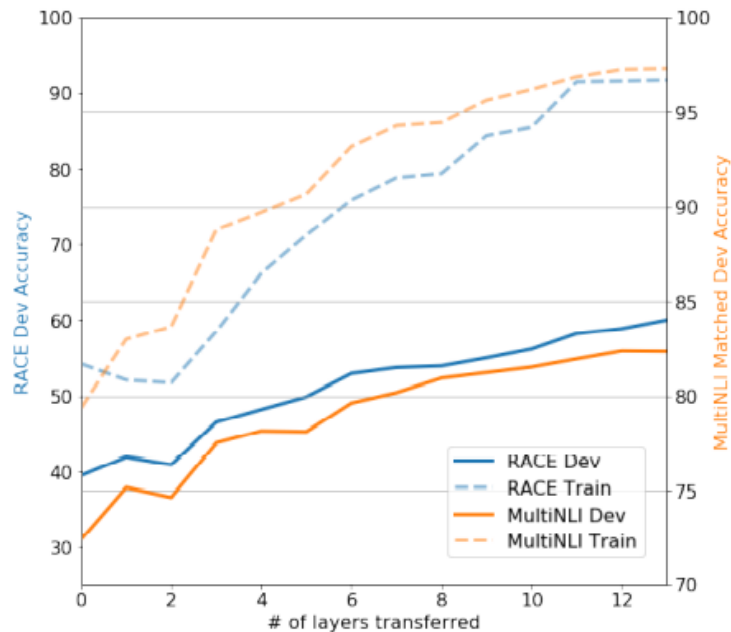
Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	<u>77.6</u>	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	<u>60.2</u>	<u>50.3</u>	<u>53.3</u>
Finetuned Transformer LM (ours)	<b>86.5</b>	<b>62.9</b>	<b>57.4</b>	<b>59.0</b>

## Natural language inference

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	<u>89.3</u>	-	-	-
CAFE [58] (5x)	80.2	79.0	<u>89.3</u>	-	-	-
Stochastic Answer Network [35] (3x)	<u>80.6</u>	<u>80.1</u>	-	-	-	-
CAFE [58]	78.7	77.9	88.5	<u>83.3</u>	-	-
GenSen [64]	71.4	71.3	-	-	<u>82.3</u>	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	<b>61.7</b>
Finetuned Transformer LM (ours)	<b>82.1</b>	<b>81.4</b>	<b>89.9</b>	<b>88.3</b>	<b>88.1</b>	56.0

# Details

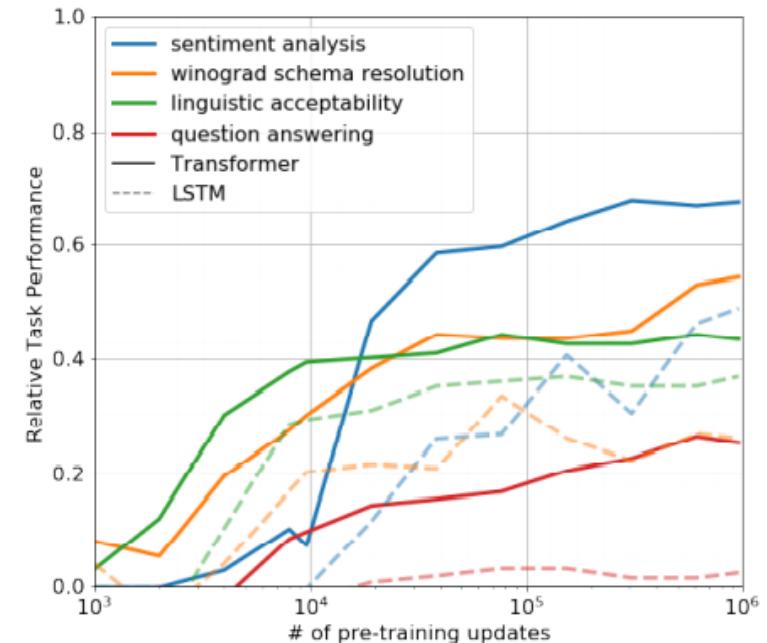
- Augmented objective function in finetuning
- More layers is better!
- Zero-shot



We additionally found that including language modeling as an auxiliary objective to the fine-tuning helped learning by (a) improving generalization of the supervised model, and (b) accelerating convergence. This is in line with prior work [50, 43], who also observed improved performance with such an auxiliary objective. Specifically, we optimize the following objective (with weight  $\lambda$ ):

$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C}) \quad (5)$$

Overall, the only extra parameters we require during fine-tuning are  $W_y$ , and embeddings for delimiter tokens (described below in Section 3.3).





# GP 2

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100  
 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200  
 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300  
 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400  
 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500  
 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600  
 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700  
 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800  
 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900  
 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000



# GPT-2

- Current paradigm => “narrow learners”
  - Don’t generalize well to out-of-distribution data
  - Hypothesis: Single task training
- Idea: Use LM and zero-shot => “general learners”
- + Make your models huge 😊
  
- $P(\text{output}|\text{input}) \rightarrow P(\text{output}|\text{input}, \text{task})$ 
  - (translate to french, english text, french text)
  - (answer the question, document, question, answer)



# WebText

- Common Crawl: big but low-quality
  - Don't use
- WebText:
  - Outbound links from Reddit (with karma  $\geq 3$ )
  - 45 million links
  - 40 GB of text
  - (Removed Wikipedia to avoid conflicts with other datasets)

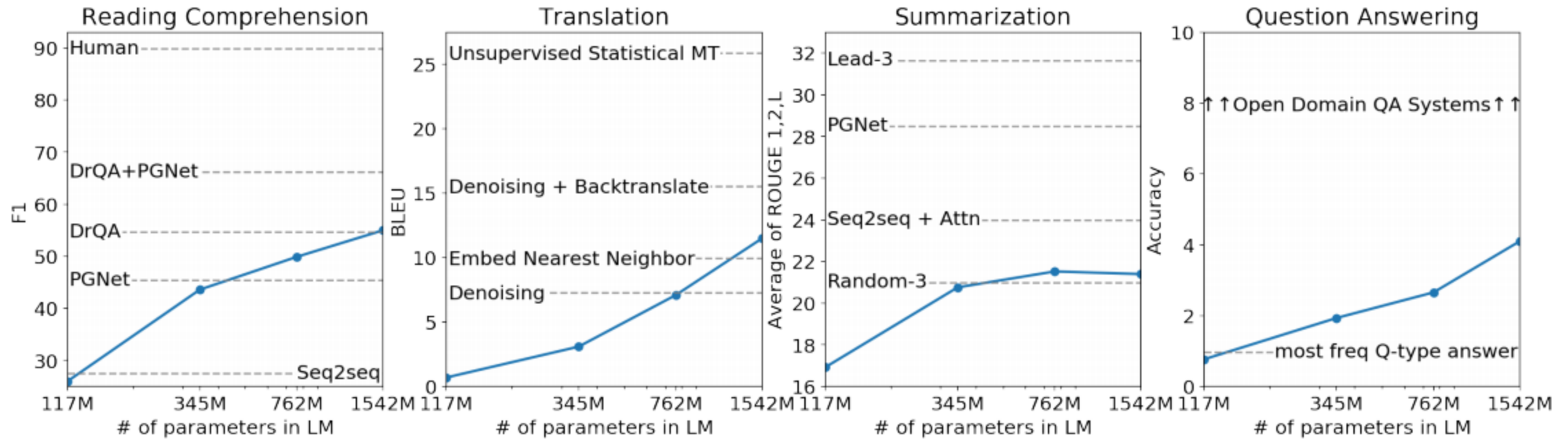
# Zero-shot Language Modelling

**Language Models are Unsupervised Multitask Learners**

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	<b>21.8</b>
117M	<b>35.13</b>	45.99	<b>87.65</b>	<b>83.4</b>	<b>29.41</b>	65.85	1.16	1.17	37.50	75.20
345M	<b>15.60</b>	55.48	<b>92.35</b>	<b>87.1</b>	<b>22.76</b>	47.33	1.01	<b>1.06</b>	26.37	55.72
762M	<b>10.87</b>	<b>60.12</b>	<b>93.45</b>	<b>88.0</b>	<b>19.93</b>	<b>40.31</b>	<b>0.97</b>	<b>1.02</b>	22.05	44.575
1542M	<b>8.63</b>	<b>63.24</b>	<b>93.30</b>	<b>89.05</b>	<b>18.34</b>	<b>35.76</b>	<b>0.93</b>	<b>0.98</b>	<b>17.48</b>	42.16

*Table 3.* Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

# Zero-shot Downstream



- Promising and impressive (compared to expectations)
- But far from SOTA



# Example: Natural Questions

- Top 30 most-confident answers
- Question: did these show up in the training data?

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%
What us state forms the western boundary of montana?	Montana	✗	52.3%
Who plays ser davos in game of thrones?	Peter Dinklage	✗	52.1%
Who appoints the chair of the federal reserve system?	Janet Yellen	✗	51.5%
State the process that divides one nucleus into two genetically identical nuclei?	mitosis	✓	50.7%
Who won the most mvp awards in the nba?	Michael Jordan	✗	50.2%
What river is associated with the city of rome?	the Tiber	✓	48.6%
Who is the first president to be impeached?	Andrew Johnson	✓	48.3%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

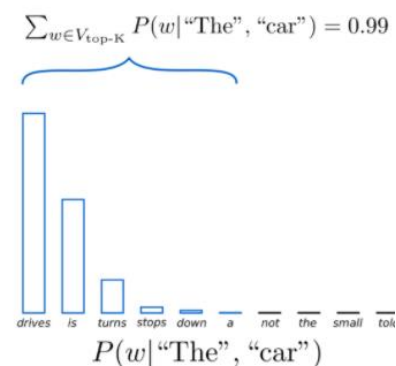
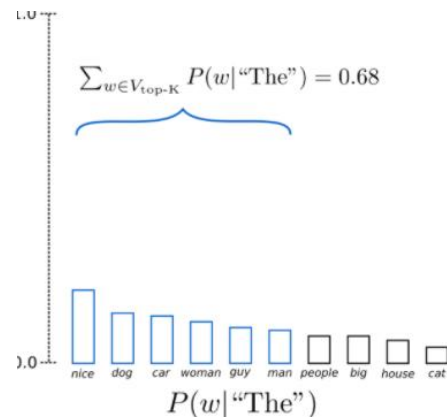
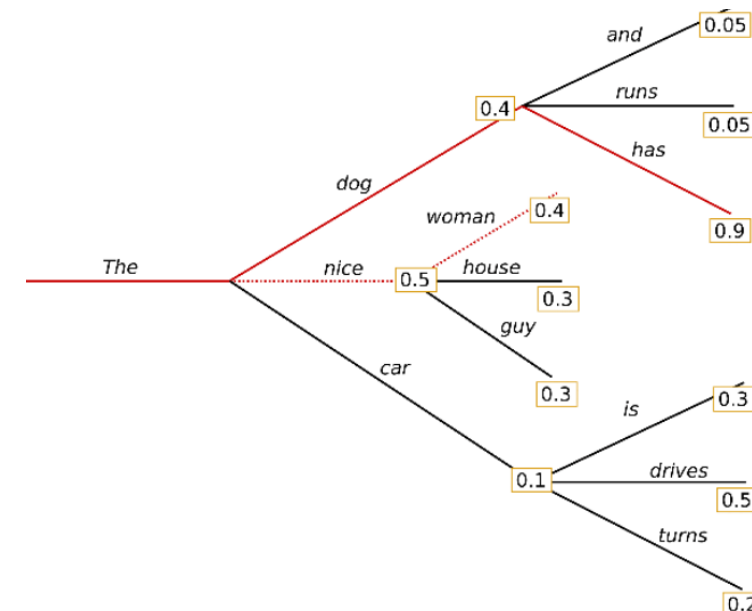
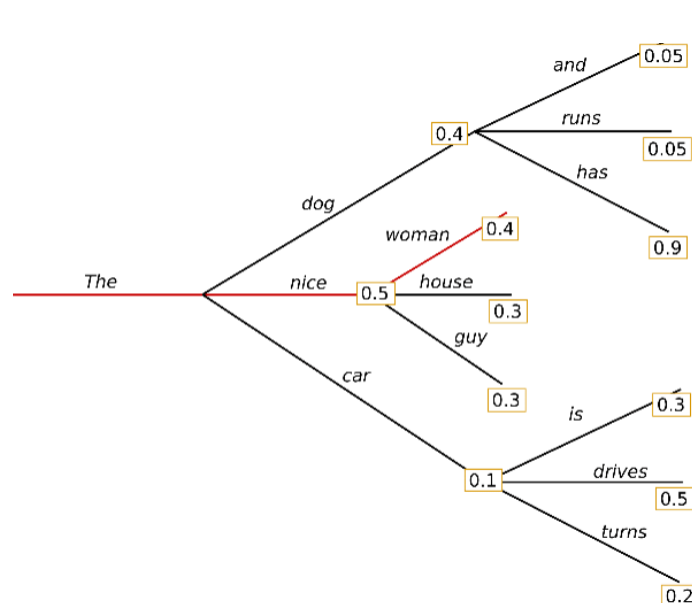
# Generalization vs Memorization

- Bloom filters with 8-grams => estimate overlap
  - Given Datasets A, B.
  - Question: What is the percentage of 8-grams from A that are also in B?
- Interesting: 1BW has overlap of ~13% with its own training set...
- TL;DR – WebText has low or no overlap with the datasets used in the studies

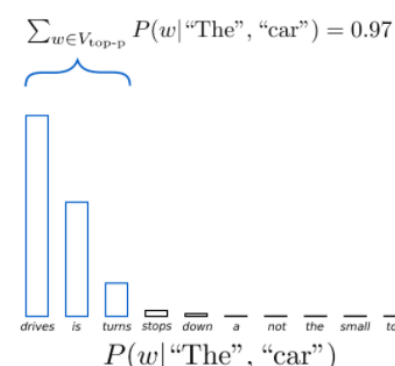
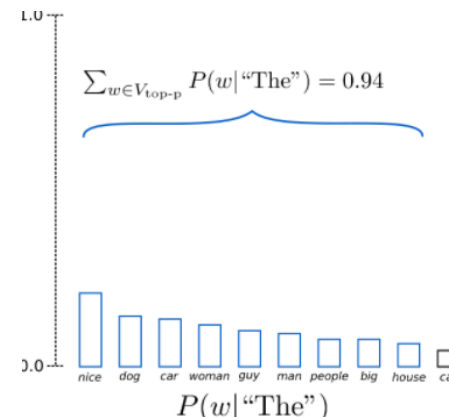
# Text generation from LMs

[Huggingface blog / how to generate text](https://huggingface.co/blog/how-to-generate-text)

- Greedy
- Beam search
- Sampling
  - Top-k sampling
  - Top-p sampling



Top-k



Top-p



**Context (human-written):** In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

**GPT-2:** The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.