

Glam

Nan Du^{*1} Yanping Huang^{*1} Andrew M. Dai^{*1} Simon Tong¹ Dmitry Lepikhin¹ Yuanzhong Xu¹
Maxim Krikun¹ Yanqi Zhou¹ Adams Wei Yu¹ Orhan Firat¹ Barret Zoph¹ Liam Fedus¹ Maarten Bosma¹
Zongwei Zhou¹ Tao Wang¹ Yu Emma Wang¹ Kellie Webster¹ Marie Pellat¹ Kevin Robinson¹
Kathleen Meier-Hellstern¹ Toju Duke¹ Lucas Dixon¹ Kun Zhang¹ Quoc V Le¹ Yonghui Wu¹
Zhifeng Chen¹ Claire Cui¹

Abstract

ComputeAnd

GPT-

Glam

ExpertSharchituction

1/3
com-point

NLP

1.2

7 GPT-
GPT-

29

1 GPT- Glam
NLU NLU
8
GPT- ops ops
token token token token token token token token
token token token token token token token token
token token token token token token token token
and to tokens interpercts and offer and to the token intermands
3

		GPT-3	GLaM	relative
cost	FLOPs / token (G)	350	180	-48.6%
	Train energy (MWh)	1287	456	-64.6%
accuracy	Zero-shot	56.9	62.7	+10.2%
on average	One-shot	61.6	65.5	+6.3%
	Few-shot	65.2	68.1	+4.4%

Patterson 2021

NetworkCan

NLP

Glam

Glam
1.2T 64 permoe
Shazeer 2017 Lepikhin 2021; Fe-

Dus 2021 96.6b
1.2t
GPT- b 29
NLP

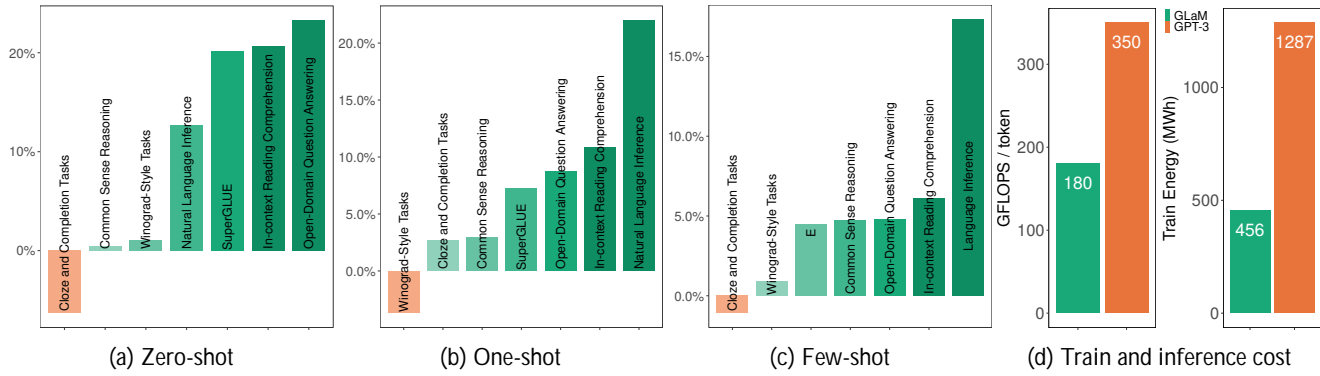
QA

theDel

GPT- 1
Glam andGPT- 1

Ton 2014 Mikolov 2013; Penning-
al 2018; Devlin 2019 Peterset
Shazeer 2017 Mod-els
Kaplan 2020 Huang 2019
AI 2020 GPT- Brownet
Flan Wei 2021
* 1 Google Nan Du
Yanping Huang Andrew M. Dai

39
PMLR 162 2022 2022



1 b GLAM c b/ e 7 D 29 7 a a GPT- b c

Glam

Vaswani 2017
NLPTASKS
Devlin 2019; Yang 2019; Liu 2019; Clark 2020
Rae 2020; Houlisby 2019

WinogenderBenchmark

MOE
NLP
SparsEdeCoder
ATSCALE
NLP
Patterson 2021

GPT- Brown 2020
Shoeybi 2019; Lieber 2021; Wei
Lan-Guage

MOE

2

Sutskever 2011
Word Vec Mikolov et al 2010;
Pennington 2014
Mikolov 2014 Em

Lan-Guage 2017 Shazeer
Moe Ar-Chitectures Hestness
2017; Shazeer 2018; Lep-ikhin 2021;
Kudugunta et al 2021 Feduset AI
Switch-C 1 pa
GLAM Bloth Switch-C
Switch-C
Switch-C
Superglue Glam

RNN LSTMS forlanguage
Dai Le 2015; Kiros 2015

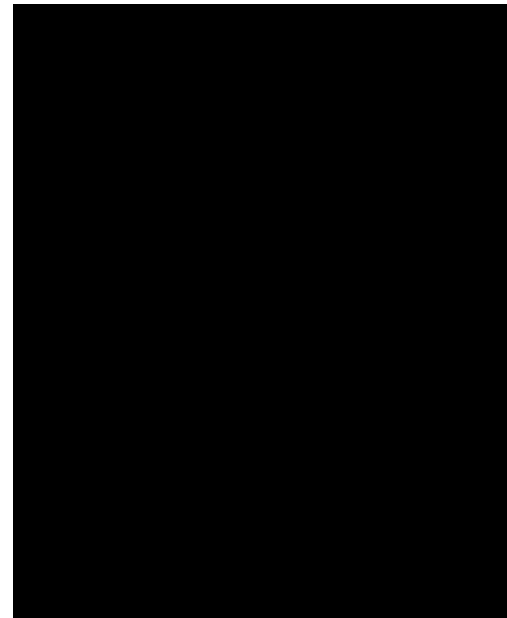
e.g., SuperGlue, while GLaM performs well without any

2
al 2020; Brown 2021; Shoeybi
NPARAMS
PARAMS

Model Name	Model Type	n_{params}	$n_{\text{act-params}}$
BERT	Dense Encoder-only	340M	340M
T5	Dense Encoder-decoder	13B	13B
GPT-3	Dense Decoder-only	175B	175B
Jurassic-1	Dense Decoder-only	178B	178B
Gopher	Dense Decoder-only	280B	280B
Megatron-530B	Dense Decoder-only	530B	530B
GShard-M4	MoE Encoder-decoder	600B	1.5B
Switch-C	MoE Encoder-decoder	1.5T	1.5B
GLaM (64B/64E)	MoE Decoder-only	1.2T	96.6B

GPT- Where Superglue

2 Glam



2 Glam

MOE

" "

64

3

16

Twodivedent

Brown

Wikipedia

2020

WEUSE

Glam

Adiwardanaet AL

Pub-Lic-Liel

2020

Wikipedia Books

Web

Wikipedia

Table

Pareto

CLAS-SIFIER

WebPagesAccording

Wikipedia

d

Brown

2020

Brown

2020

3

Dataset	Tokens (B)	Weight in mixture
Filtered Webpages	143	0.42
Wikipedia	3	0.06
Conversations	174	0.28
Forums	247	0.02
Books	390	0.20
News	650	0.02

4.

Experts MOE Shazeer

2017; Fedus

2021

Glammodels

GSHARD MOE

Lepikhinet al 2021

MOE

2

MOE

independent feed-forward networks as the 'experts'. A

SoftMax

l n

n act-params

n

MOE

d

Agiven

token ops

1 Moelayer

Glam /e Glam b/ e

E

O E 2

Foran Moe

GLAM b/ e Anapproximate b

64 MOE

Glam

MOE

Trans-Former

Dalet AI

2019 MOE

Glam Glam b

137b

Dauphin

2017; Shazeer

- wiseproduct

5.2

2020

Hendrycks Gimpel 2016

Xu 2D Shard-ing

2021

1024

100

0

Afafactor Shazeer Stern 2018

1= 0 2= 0.99 1- t

-0.8 UpdateClipping 1.0

5.

Glam

mation 10K

0.01 LRT 1

5.1.

t

GSHARD Lepikhin 2021

0.01 toencourage

MoE

4

1.3

Moelayer

E S

1.2

B

lofloat 256K

Kudo Richardson 2018

GLAM B/ E B oat

1,024 Cloud TPU-V

m

H

1

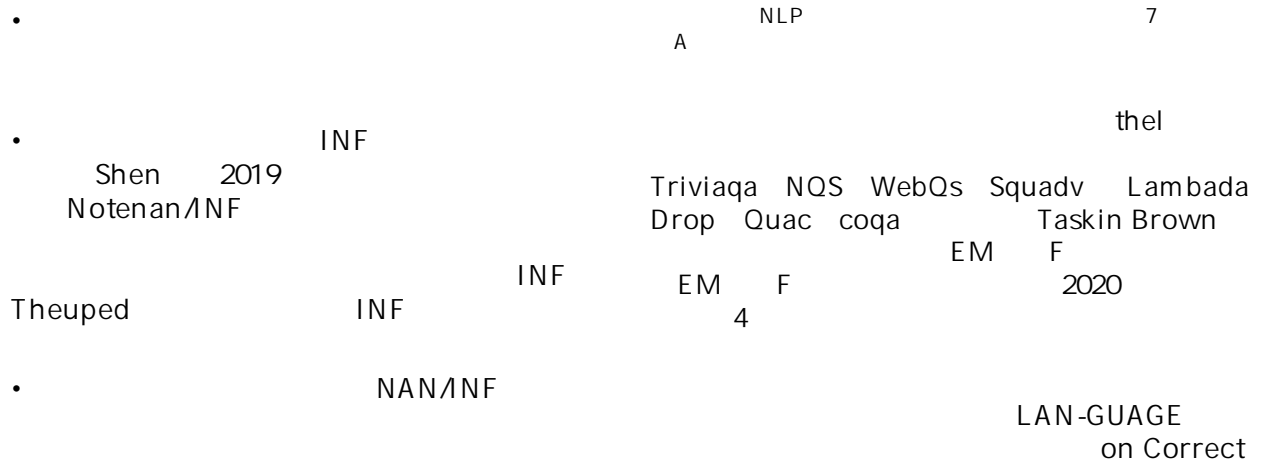
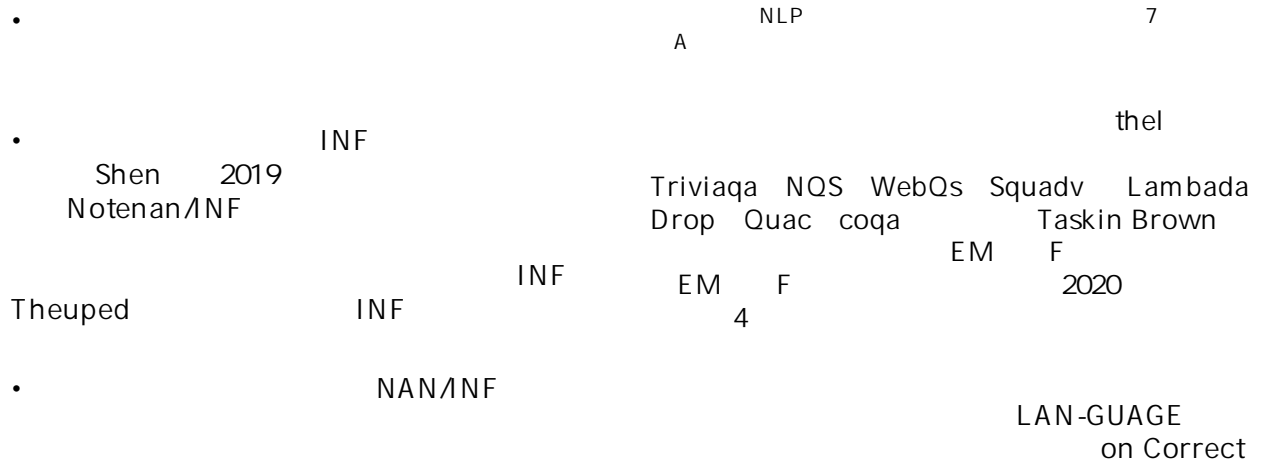
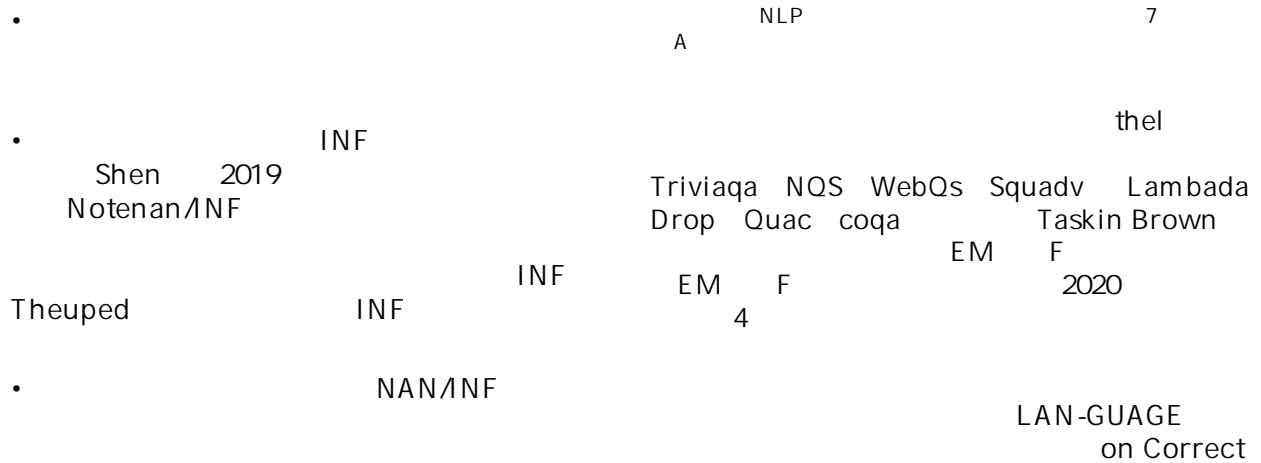
" "

/

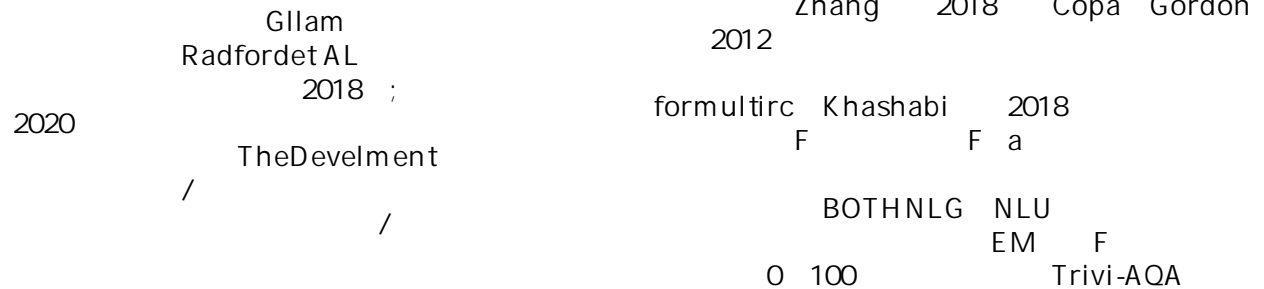
TheGlam

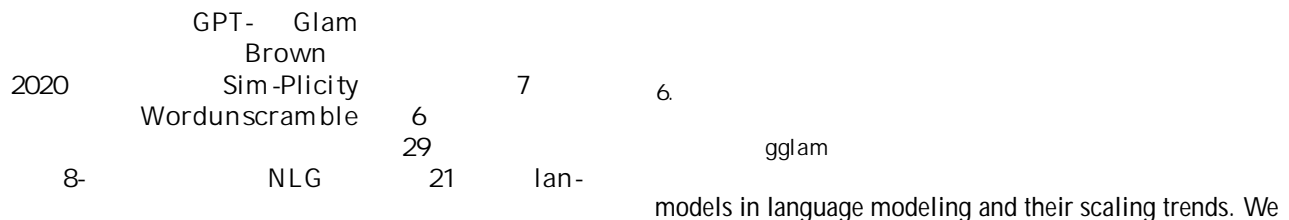
GLaM models.

GLaM Model	Type	n_{params}	$n_{\text{act-params}}$	L	M	H	n_{heads}	d_{head}	E
0.1B	Dense	130M	130M	12	768	3,072	12	64	–
0.1B/64E	MoE	1.9B	145M						64
1.7B	Dense	1.7B	1.700B						–
1.7B/32E	MoE	20B	1.878B						32
1.7B/64E	MoE	27B	1.879B	24	2,048	8,192	16	128	64
1.7B/128E	MoE	53B	1.881B						128
1.7B/256E	MoE	105B	1.886B						256
8B	Dense	8.7B	8.7B	32	4,096	16,384	32	128	–
8B/64E	MoE	143B	9.8B						64
137B	Dense	137B	137B	64	8,192	65,536	128	128	–
64B/64E	MoE	1.2T	96.6B	64	8,192	32,768	128	128	64

- 
- 
- 

5.3





6.1 MOE

GPT- b Forzero
 Glam b/ e
 1
 GLAM b/ e 7 6
 GPT-
 11 weInclude
 Megatron-nlg Gopher
 4 Glam b/ e
 96.6b
 ops ops bygpt-

Triviaqa ANSWER
 Query
 2020 Trivi-
 AQA
 Glam b/ e iSbetter 5

SOTA
 sota Yu 2022
 8.6
 GPT- 5.3 GPT-
 N Act-Params ofglam b/ e GPT-
 Glam
 Glam -C
 b/ e
 TPU -C
 GLAM Triviaqa
 -C 11
 12,13 14

6.2

Glammodel . b/ e
 Texton
 3
 3

Model	TriviaQA (Open-Domain)
KG-FiD (large) (Yu et al., 2022) (finetuned, test)	69.8
Switch-C (finetuned, dev)	47.5
GPT-3 One-shot (dev)	68.0
GPT-3 64-shot (test)	71.2
GLaM One-shot (test)	75.0
GLaM One-shot (dev)	75.8

143b

7T

3 c d
 NLG NLU
 NLU
 NLG
 NIgoften

6.3

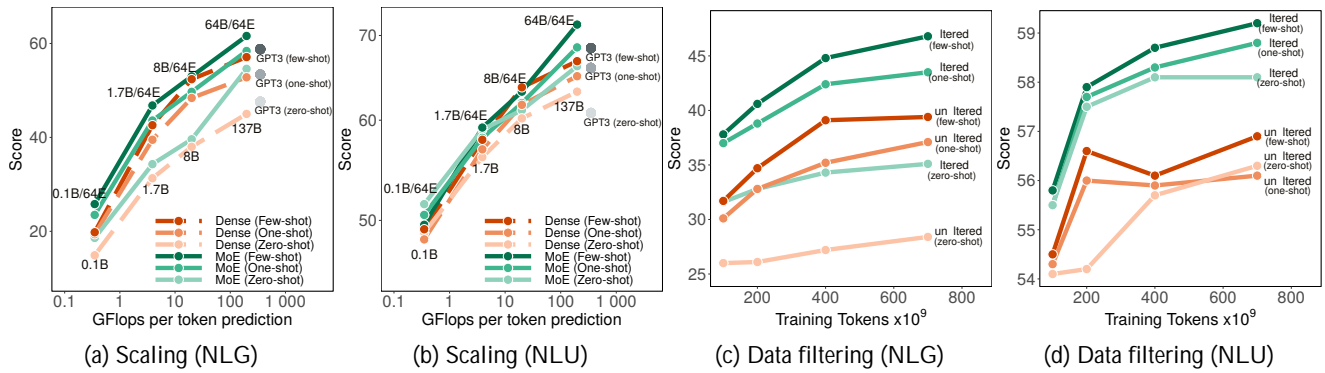
InputExample
 " " n
 n params = n params = n act-params
 4 n

params n n paramsn act-
 Moe Glam

3 a

n act-params
 Glam Moe Glam
 3 b Glam
 Moemodells Generative tasks
 MOE
 Moe

similarly at smaller scales but MoE models outperform at



3 8nlg a 21 NLU B Glam Moe Glam 1
 C NLG D NLU b/ e Oneand

B

6.4

Patterson 2021

Glammodels

4 A-C 4 E-G
 NLG NLUTASK
 X
 ITI ITI ITIS 300b
 GPT-

MOE

630b 280B
 Glam b/ e 6
 4 / NLU / NLG
 4 300b GPT-
 GPT- NLG

4 d 4 h
 TPUyears

MOE

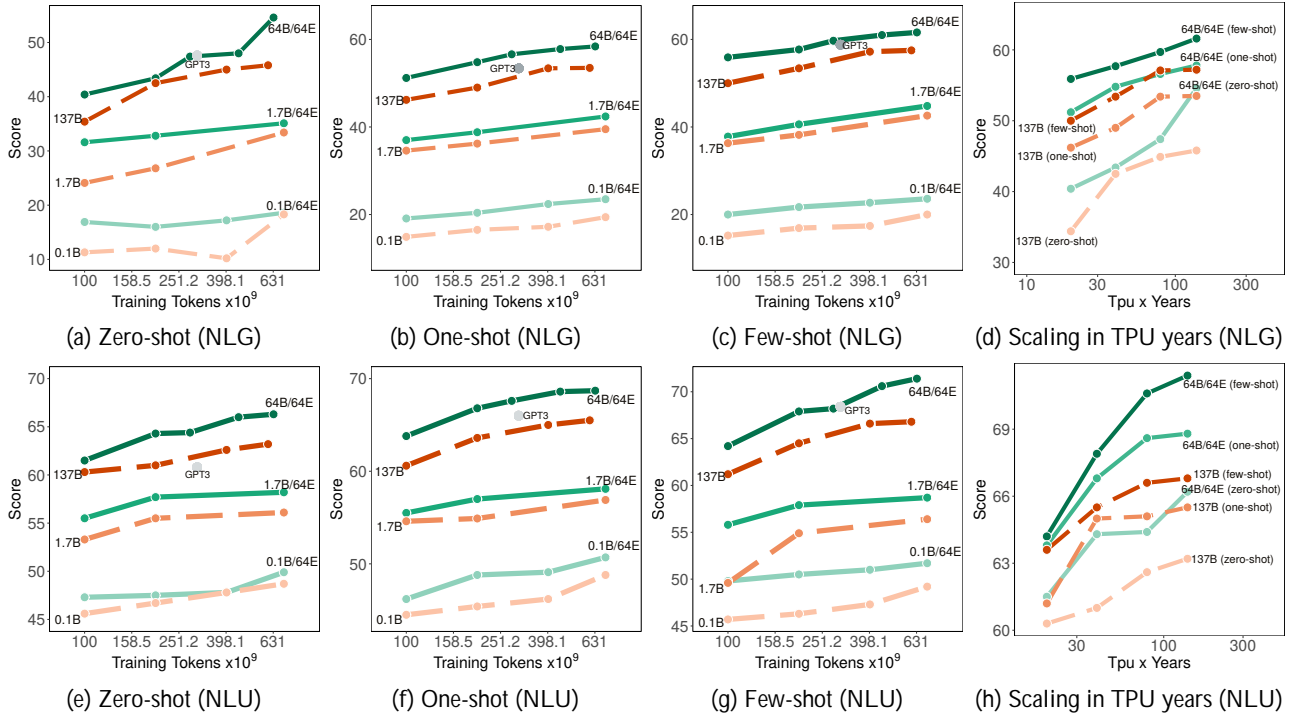
7

AI

Leidner Plachouras 2017; Bender
 2021; Bom-Masani 2021
 Blodgett et al 2020
 Rogers 2021
 Friedman 2018 Abadi 2016b
 Carlini 2020 Strubell
 2019; Patterson et; Patterson et; Patterson et
 AI 2021

Bolukbasiet al 2016; Rudinger
 2018; Zhao 2018
 Li 2020 Nadeem
 2021 Hutchinson
 2020

ties (Hutchinson et al., 2020), as well as other social biases



4 9 NLG A-C 21 NLU
GlamDense

E-G d h Glam Moe GlamDense TPU

Caliskan 2017; Rudinger 2020; Sotnikova 2021

2017; Sap Blodgett 2021 ; Jacobs

Wallach

2019; Webster 2021

May

Top-K Sam-Pling k = 40
800 1
Bird Loper 2004 propwords

asadverbs

2020

Weomit

2020 GRT- Brown et al
Winogender Rudinger
2018 Wealso Gopher
Raeet AI 2021

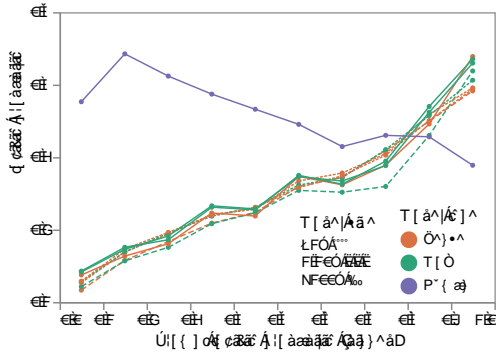
Webuild
" " " "
10 " " 8 Gen-
dered 9 10

7.1. Co-occurrence prompts

Brown 2020
" {term}

7.2. WinoGender

Stanovsky 2019; Webster Pitler 2020
Ques-Tion Ansporming Lamm 2020
Glam -
gendered correlations in GLaM cause it to make corefer-



dataset

TPC
TPC

TPP

API 25
API

5 TPP
TPC

8

reddit

TPP
100

1

TPC

Winogender Rudinger 2018
Glam b/ e 71.7
GPT- 64.2 Brown 2020
71.7 " HE"
70.8 " " 72.5

- Rudinger 2018
gotcha" 71.7

7.3

Fedus 2021 MoE

MOE Triviaqa QABCHENCHS
FLOP

Welbl

2021; Rae 2021
sentences that
Real toxicity Prompts DataSet Gehman
2020
Actinature
API

PA

9.

Glam
Experters

10K 100 k = 40 25
1 perspectiveapi
0.0 0.0 emptiment
string botsimply

29 GPT- 1.2
Glam b/ e

gpt-

MOE

5 TPP
TPC TPP
TPC

models.

References

- Abadi M. Barham P. Chen J. Chen Z. Davis A. Dean J. Devin M. Ghemawat S. Kudlur M. Levenberg J. Monga R. Moore S. Murray d G. Steiner B. Tucker P. Vasudevan V. Warden p Wicke M. Yu Y Zheng X Tensor ow 2016 11 11 12thusenix OSDI 16 OSDI 16 265-283 USENIX ISBN 978-1-931971-33-1 URL <https://www.usenix.org/conference/osdi/technical-sessions/presentation/abadi>
- Abadi M. Chu A. Goodfellow I. McMahan H B. Mironov I. Talwar K Zhang L Deep Learning Wwith Di cnial 2016 ACMSIGSAC 2016 10 doi . / . .url <http://dx.doi.org/ . / .>
- Adiwardana, D. Luong, M. So, D. R. Hall, J. Fiedel, N. Thoppilan, R. Yang, Z. Kulshreshtha, A. Nemade, G. Lu, Y. Le Q.V. CoRR abs/ . 2020 <https://arxiv.org/abs/>
- Bender E M. Friedman B 6 587-604 2018 doi 10.1162/tacl A URL <https://aclanthology.org/q ->
- Bender E M. Gebru T. McMillan-Major A. FACCT '21 610-623 2021 ISBN DOI 10.1145/3442188.3445922 URL <https://doi.org/ . / .>
- Berant J. Chou A. Frostig R Liang P Semanticparsing freebase 2013 1533-1544 2013 10 URL <https://aclanthology.org/d ->
- 2004 url <https://aclanthology.org/> p -
- Bisk Y. Zellers R. Bras R L. Gao J Choi Y. PIQA Natural language 34 AAAI 2020
- Blodgett S.L. Barocas S. Daumé III H Wallach H NLP " " 58 pp 5454-5476 2020 7 doi 10.18653/v / .acl-main. URL <https://aclanthology.org/ .acl-main>.
- Blodgett S L. Lopez G. Olteanu A. Sim R Andwallach H 59 11 1004-1015 2021 8 doi 10.18653/v / .acl-long. URL <https://aclanthology.org/ .acl-long>.
- Bolukbasi T. Chang K.-W. Zou J Y. Saligrama V Kalai A T. Man Progince Meras Woman Woman to Homemaker em Lee D. Sugiyama M. U. Luxburg I R. Garnett R 29 Curran As-Sciociates Inc. 2016 URL <https://proceedings.neurips.cc/paper/ / le/ a cd e ac d f f f f f f f ec - paper.pdf>
- Bommasani, R. Hudson, D. A. Adeli, E. Altman, R. Arora, S. von Arx, S. Bernstein, M. S. Bohg, J. Bosse-lut, A. Brunskill, E. Brynjolfsson, E. Buch, S. Card, D. Castellon, R. Chatterji, N. S. Chen, A. S. Creel, K. Davis, J. Q. Demszky, D. Donahue, C. Doum-bouya, M. Durmus, E. Ermon, S. Etchemendy, J. Etha-yarajh, K. Fei-Fei, L. Finn, C. Gale, T. Gillespie, L. Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J. . Icard, T. Jain, S. Jurafsky, D. Kalluri, P. Karamcheti, S. Keeling, G. Khani, F. Khat-tab, O. Koh, P. W. Krass, M. S. Krishna, R. Kudi-tipudi, R. CoRR abs/ . 2021 <https://arxiv.org/abs/>
- J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry,

- G. Askeel A. Agarwal S. Herbert-Voss A. Krueger G. Henighan T. Winter C. Hesse C. Chen M. Sigler E. Litwin M. Gray S. Chess B. Clark J. Berner C. McCandlish S. S. S. S. Radford A. Sutskever I. Amodei d. Larochelle H. Ranzato M. Hadsell R. Balcan M.F. Lin H. 33 1877– 1901 Curran Asso-Ciates Inc 2020 URL <https://proceedings.neurips.cc/paper/paper/ /le/ c d bfc bfb ac f f aa-paper.pdf>
- Caliskan A. Bryson J. J. Narayanan A. Seman-Tics CONCOLHUMAN Science 356 6334 183–186 APR ISSN 1095-9203 doi 10.1126/science.aal .url <http://dx.doi.org/ /science.aal>
- Carlini N. Tramèr F. Wallace E. Jagielski M. Herbert-Voss A. Lee K. Roberts A. ú Oprea A. Ra el C. Corr ABS/ . ,
- Choi E. He H. 2018 2174-2184 2018 10 11 doi 10.18653/v /d - URL <https://aclanthology.org/d ->
- C. 2019 1 Longand Short Papers 2924-2936 2019 6 doi 10.18653/v /n - URL <https://aclanthology.org/n ->
- Clark, K. Luong, M.-T. Le, Q. V. Manning, C. D. Elec-tra arXiv arXiv: . , 2020
- Clark, P. Cowhey, I. Etzioni, O. Khot, T. Sabharwal, A. Schoenick, C. Tafjord, O. arc ai arXiv . v 2018
- I. Quiñonero-Candela J. Dagan I. Magnini B D'AlchéBuc f 177-190 2006 SpringerBerlin Heidelberg ISBN - - - -
- Dai, A. M. Le, Q. V. Cortes, C. Lawrence, N. Lee, D. Sugiyama, M. Garnett, R. 28 Curran As-sociates, Inc. 2015 <https://proceedings.neurips.cc/paper/ /le/ debd ae d ab aa b - Paper.pdf>
- Dai Z. Yang Z. Yang Y. Carbonell J. Le Q andsalakhutdinov R Transformer-XL 57 2978-2988 2019 7 LIN-GUITION doi 10.18653/v /p - URL <https://aclanthology.org/p ->
- Dauphin, Y. N. Fan, A. Auli, M. Grangier, D. 933-941 PMLR 2017
- De Marne e M.-C. Simons M. Tonhauser J. ThecommitmentBank NaturallyCurring Sinn Und Bedeutung 23 2 107–124 2019 7 DOI 10.18148/sub/sub/sub/sub/ sub/ .v i . .url <https://ojs.um.uni-konstanz.de/sub/sub/index PHP/SUB/ARTICE/VIEW/>
- Devlin, J. Chang, M.-W. Lee, K. Toutanova, K. BERT 2019 1 2019
- Dua D. Wang Y. Dasigi P. Stanovsky G. Singh S Gardner M Drop burstein Doran Doran C. 2-7 2019 1 2368-2378 2019 DOI
- sociation for Computational Linguistics, 2019. doi:

- 10.18653/v1/n19-1246. URL <https://doi.org/10.18653/v1/n19-1246>.
- Fedus W Zoph B Shazeer N Switch
Transform-ers
CORR ABS/ . 2021 URL <https://arxiv.org/abs/> .
- Fyodorov Y Winter Y Francez N
2000
- Gehman S Gururangan S Sap M Choi Y
Smith n A. RealtoxicityPrompts
2020
- Gordon A Kozareva Z Roemmele
M Semeval- 7
*SEM 2012
- 1
2
Semeval
2012 394-1 398 2012
6 7 URL <https://aclanthology.org/s->
- Hendrycks D Gimpel K
- Hestness J Narang S Ardalani N Diamos
G.F. Jun H. Kianinejad H. Patwary
M.M.A. Yang Y. Corr ABS/ . 2017
URL <http://arxiv.org/abs/> .
- Houlsby N. Giurghi A. Jastrzebski S.
Morrone B. de Laroussilhe Q.
Gesmundo A. Attariyan M. K
Salakhutdinov R 36
97
2790-2799 PMLR 2019 6 9
15 URL https://proceedings.mlr.press/v7/houlsby_a.html
- Huang Y. Cheng Y. Bapna A. Firat O.
Chen D. Chen M X. Lee H. Ngiam J. Le
Q.V. Wu Y Chen Z. Gpipe Pipeline
Parallelism H. M.
Wallach H. Larochelle A. Beygelzimer A.
D'Alché-Buc F. Fox E B. Garnett R
32 2019
2019 Neurips 2019 2019 12 8 14
2019 12 8 14
103-112 2019
- Hutchinson B. Prabhakaran V.
Denton E. Webster K. Zhong Y
Denuyl S NLP Mod-els
58
5491-5501 2020 7
doi 10.18653/v / .acl-main. URL
<https://aclanthology.org/ .acl-main>.
- Jacobs A Z. Wallach H 2021
ACM 2021
3 DOI 10.1145/3442188.3445901 URL
<http://dx.doi.org/ . / .>
- Joshi, M. Choi, E. Weld, D. S.
Zettlemoyer, L. Trivi-aq
55 2017
7
- Kaplan, J. McCandlish, S. Henighan, T. Brown, T.
B. Chess, B. Child, R. Gray, S. Radford, A. Wu,
J. Amodei, D. arXiv
arXiv: . , 2020
- D. 2018
2018 1
252-262
2018 6
doi 10.18653/v /n - URL <https://aclanthology.org/n->
- R. Cortes C. Lawrence N.
Lee D. Sugiyama M R. Garnett R
28 Curran Associates Inc. 2015
URL <https://proceedings.neurips.cc/paper/ /le/f d fa ad a da - paper.pdf>
- Kudo T Richardson J
distokenizer EMNLP 2018
- Kudugunta S. Huang Y. Bapna A. Krikun
M. Lep-ikhin D. Luong M.-T. Firat O
EMNLP 2021
3577-3599 2021
tics: EMNLP 2021, pp. 3577-3599, 2021.

Kwiatkowski, T. Palomaki, J. Redfield, O. Collins, M. Parikh, A. Alberti, C. Epstein, D. Polosukhin, I. Kelcey, M. Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S.

2019

2017

785-794

2017 9

doi 10.18653/

urlhttps://
aclanthology.org/d -

Lamm M. Palomaki J. Alberti
C Andor D. Choi E. Corr
ABS/ . 2020 URL https://
arxiv.org/abs/ .

Le Q Mikolov T ven-tence
2014 on Machine

Leidner J L. Plachouras V

doi 10.18653/v /w - URL https://
aclanthology.org/w -

Lepikhin D. Lee H. Xu Y. Chen
D. Firat O. Huang Y
2021
URLhttps://openreview.net/
forum?id=qrwe xhtmlmyb

Levesque H. Davis E
Morgenstern L Wino-Grad
13

KR 2012

552-561

2012 ISBN

13

KR KR 2012

2012 10 10 2012 1 14

Li T. Khashabi D. Khot T. Sabharwal A
Srikumar v

EMNLP 2020

3475-3489

Online, November 2020. Association for Computational
Linguistics. doi: 10.18653/v1/2020.findings-emnlp.
311. URL [https://aclanthology.org/2020.
findings-emnlp.311](https://aclanthology.org/2020.findings-emnlp.311).

Lieber, O. Sharir, O. Lenz, B. Shoham, Y.
Jurassic- AI 2021

Liu, Y. Ott, M. Goyal, N. Du, J. Joshi, M. Chen,
D. Levy, O. Lewis, M. Zettlemoyer, L.
Stoyanov, V. Roberta bert
arXiv arXiv 1907.11692 2019

May C. Wang A. Bordia S. Bowman
S.R. Andrudinger R
2019

1
622-628

2019 6 -

doi 10.18653/v /n -
urlhttps://aclanthology.org/n -

Mihaylov, T. Clark, P. Khot, T. Sabharwal, A.
EMNLP
2018

Mikolov T. Karant M. Burget L. Cernocký
J.H. Andkhanpur S. Languagemodel
Interspeech 2010

Mikolov T. Chen K. Corrado G. Dean J
trientientientientientientientientientientie
of vector Space In Bengio Y
Lecun Y 1 ICLR
2013 5 2-4
2013 5 2-4
URLHTTP://arxiv.org/abs/ .

Mostafazadeh N. Chambers N.
2016

839-
849 - 2016 6

doi 10.18653/v /
n - URL https://aclanthology.org/
n -

Nadeem M. Bethke A. Reddy S
Stereoset MEA
Pro-Cessing 1
5356-5371

cessing (Volume 1: Long Papers), pp. 5356-5371, Online,

- 2021 8 doi 10.18653/v1/d19-1.10 .acl-long. URL <https://aclanthology.org/2021.d19-1.10>
- Paperno, D. Kruszewski, G. Lazaridou, A. Pham, N. Q. Bernardi, R. Pezzelle, S. Baroni, M. Boleda, G. Fernández, R. LAMBADA 54 1525–1534 2016 8 doi 10.18653/v1/P16-1054 .URL <https://aclanthology.org/P16-1054>
- Patterson, D. Gonzalez, J. Le, Q. Liang, C. Munguia, L.-M. Rothchild, D. So, D. Texier, M. Dean, J. Car-bon arXiv 2104.10350 2021
- Pennington J Socher R Manning C 2014 NAT-Aran EMNLP 1532-1543 2014 10 doi 10.3115/v1/d14-1010 .url <https://aclanthology.org/d14-1010>
- Peters, M. E. Neumann, M. Iyyer, M. Gardner, M. Clark, C. Lee, K. Zettlemoyer, L. arXiv 1802.05365 2018
- Pinehvar M T. Camacho-Collados J WIC 10 000 Arxiv 2018
- Radford A. Wu J. Child R. Luan D. Amodei d / .pdf
- Rae J.W. Borgeaud S. Cai T. Millican K. Ho mann J. Song H.F. Aslanides J. Henderson S. Ring R. R. Young S. Rutherford Rutherford E. Hennigan T. Menick J. Cassirer A. Powell R. van den driessche G. Hendricks L.A. J. Dathathri S. Huang S. Uesato J. Mellor J. Higgins I. Creswell A. McAleese N. Wu A. Elsen E. S.M. Buchatskaya E. Budden D. Sutherland E. Simonyan K. Paganini M. Sifre L. A. Gribovskaya E. Donato D. Lazaridou A. Mensch A. Lespiau J. Tsimpoukelli M. M. Pohlen T. Gong Z. Toyama D. de Masson d'Apume c Li Y. Terzi T. Mikulik V. Babuschkin I. A. De Las Casas D. Guy A. E. Osindero S. Rimell L. Dyer C. Vinyals O. Ayoub K. Stanway J. Bennett L. Hassabis D. Kavukcuoglu K Irving G Gopher Corr ABS/ 2021
- Ra el, C. Shazeer, N. Roberts, A. Lee, K. Narang, S. Matena, M. Zhou, Y. Li, W. Liu, P. J. J. 21:140:1–140:67 2020 URL <http://jmlr.org/papers/v17/ra18.html>
- Rajpurkar P. Jia R Liang P ACL 2018
- Reddy S. Chen D Manning C D. Coqa 2019 3 7 249–266 doi 10.1162/tacl.2019.3.7.249 .URL <https://aclanthology.org/q19-1.249>
- Rogers A 59 11 2182- 2194 2021 8 doi 10.18653/v1/d19-1.10 .acl-long. URL <https://aclanthology.org/2021.d19-1.10>
- Rudinger R May C Van Durme B ACL 2017 4 74-79 doi 10.18653/v1/W17-0015 .URL <https://aclanthology.org/W17-0015>
- Rudinger R. Naradowsky J. Leonard B Van Durme b 2018 2 8-14 2018 6 doi 10.18653/v1/n18-1.01 .url <https://aclanthology.org/n18-1.01>
- Rae J.W. Borgeaud S. Cai T. Millican K. Ho mann J. Song H.F. Aslanides J. Henderson S. Ring R. R. Young S. Rutherford Rutherford E. Hennigan T. Menick J. Cassirer A. Powell R. van den driessche G. Hendricks L.A. J. Dathathri S. Huang S. Uesato J. Mellor J. Higgins I. Creswell A. McAleese N. Wu A. Elsen E. S.M. Buchatskaya E. Budden D. Sutherland E. Simonyan K. Paganini M. Sifre L. A. Gribovskaya E. Donato D. Lazaridou A. Mensch A. Lespiau J. Tsimpoukelli M. M. Pohlen T. Gong Z. Toyama D. de Masson d'Apume c Li Y. Terzi T. Mikulik V. Babuschkin I. K. AAAI 2020 8732–8740 AAAI 2020
- SAP M. Gabriel S. Qin L. Jurafsky D. Smith N. A. Choi Y and Choi, Y. Social bias frames: Reasoning about

- 58 5477-5490 2020 7
doi 10.18653/
v / .acl-main. URL https://
aclanthology.org/ .acl-main.
- Shazeer, N. Glu variants improve transformer, 2020.
- Shazeer N Stern M Adafactor
ARXIV
ABS/ . ,
- Shazeer N. Mirhoseini A. Maziarz K.
Davis A. LE Q V. Hinton G E.
Dean J
Expertslayer 2017 4 24 26
rep-sersentations
OpenReview.net 2017 URL
https://openreview.net/forum?id=b ckmdqlg
- Shazeer N. Cheng Y. Parmar N.
Tran D. Vaswani A. Sepassi R
Hechtman B
32
NIPS' 10435-10444
Redhook 2018 CurranAssociates Inc.
- Shen J. Nguyen P. Wu Y. Chen
Z. Chen M. X. Jia Y Kannan
A. Sainath T.N. Cao Y. Y.
Chorowski J. Hinsu S. Lorenzo
S. Qin J. Firat O. Macherey W.
Gupta S. Bapna A. Zhang S.
Pang Pang R. Weiss R.J.
Prabhavalkar R. Liang Q. Jacob
B. Liang B. Lee H. Chelba C.
Jean S. Li Li B. Johnson M
R. F. Richardson J.
Macherey K. Bruguier A. Zen H.
Ra el C. Kumar S. Rao K.
Rybach D. V. Krikun M. Bacchiani
M. Jablin T B. Suderman R.
Williams I. Lee B. Bhatia D.
- Shoeybi, M. Patwary, M. Puri, R. LeGresley, P.
Casper, J. Catanzaro, B. Megatron-lm GPU
arXiv
arXiv 1909.08053 2019
- Sotnikova A. Cao Y T. DauméIII H
Rudinger r
Com-putational
ACL-IJCNLP 2021 4052-4065
2021 8 doi 10.18653/
v / ndings-acl. URL https://
aclanthology.org/ . ndings-acl.
- Stanovsky G. Smith N A.
Zettlemoyer L
57 1679-
1684 2019 7
doi 10.18653/v /p - URL
https://aclanthology.org/p -
- Strubell E. Ganesh A McCallum
A NLP 2019 7
57
3645-3650
doi 10.18653/v /p - urlhttps //
aclanthology.org/p -
- Sutskever I. Martens J Hinton G
textwith wits Wtstext Wtswith Recisurrent
28 on Machine Learning
ICML' 1017-1024
2011 Omnipress ISBN 9781450306195
- Vaswani, A. Shazeer, N. Parmar, N. Uszkoreit,
J. Jones, L. Gomez, A. N. Kaiser, L. u.
Polosukhin, I
(Guyon) (Luxburg) V. Bengio,
S. Wallach, H. Fergus, R. Vish-wanathan, S.
Garnett, R.
30 Curran As-sociates, Inc. 2017 https://
proceedings.neurips.cc/paper/ /
le/ f ee dee fbd c c a aa-
Paper.pdf
- Wang A. Pruksachatkun Y. Nangia
N. Singh A. H.
Wallach H Larochelle A. Beygelzimer
A. D'AlchéBuc F. Fox E Garnett
R
32 -Sociates Inc. 2019 URLhttps //
proceedings.neurips.cc/paper/ /
le/ bf afe fab f bf da de de -
paper.pdf
- Webster K Pitler E piv-ot corr
ots to model pronoun gender for translation. *CoRR*,

abs/2006.08881, 2020. URL <https://arxiv.org/abs/2006.08881>.

K. 2021

Wei J. Bosma M. Zhao V. Y.
Guo K. Yu A.W. Lester B

Welbl J. Glaese A. Uesato J.
Dathathri S. Mel-Lor J. Hendricks
L.A. Anderson K. Kohli P. Coppin
B Huang P P -s
doi 10.18653/v1/n18-1001
emnlp. URL <https://aclanthology.org/2018.emnlp-main>.

Xu Y. Lee H. Chen D. Hechtman B.A.
Huang Y. Joshi R. Krikun M. Lepikhin D. Ly
A. Maggioni M. Pang R. Shazeer N. Wang
S. Wang T. Wu Y. Andchen Z. GSPMD
CORR ABS/2018. URL <https://arxiv.org/abs/2006.08881>.

Z. Z. Y. J.
Salakhutdinov R. R. Le Q.V. Xlnet
2019
32

Yu D. Zhu C. Fang Y. Yu W.
Wang S. Xu Y. Ren X. fusion-In-in-
decoder Question
60
1 4961-4974
IRE-LAND 2022 5 doi
10.18653/v1/n22-1005 .acl-long. URL <https://aclanthology.org/2022.ire-land>.

Yu Y. Abadi M. Barham P. Brevdo E.
Burrows M. Davis A. Dean J.
Ghemawat S. Harley T. Isard M. Kudlur
M. Monga R. Murray D. Andzheng X
13
18 2018
ISBN .DOI
10.1145/3190508.3190551 URL <https://doi.org/10.1145/3190508.3190551>.

Zellers R. Holtzman A. Bisk Y. Farhadi A
Choi Y. Hellaswag
57

4791-4800

2019 7 doi 10.18653/v1/n19-1007
P - <https://aclanthology.org/P19-1007>

Zhang S. Liu X. Liu J. Gao J. Duh
K Durme b V.
Corr ABS/2018

Zhao J. Wang T. Yatskar M. Ordonez V
Chang K.-W
2018

2
15-20
2018 6 doi 10.18653/v1/n18-1006
n - [urlhttps://aclanthology.org/n18-1006](https://aclanthology.org/n18-1006)

<https://aclanthology.org/N18-2003>.

A.

2017 Triviaqa Joshi et al
 2019 Web NQS Kwiatkowski et al
 WebQS Berant 2013

Lambda Paperno et al
 2016 Hellaswag Zellers 2019
 StoryCloze Mostafazadeh 2016

erno
 019),

Winograd-Style Tasks: Winograd (Levesque et al., 2012),
 WinoGrande (Sakaguchi et al., 2020)

PIQA Bisk 2020 Arc Easy
 Clark 2018 Arc Chal-Lenge Chal-
 lenge Clark 2018

OpenBookQa Mihaylov et al 2018

Drop Dua 2019
 Coqa Reddy 2019 Quac Choi 2018
 Squadv Rajpurkar 2018 Race-H Race-
 H Race-H Lai 2017 Race-M Lai 2017

Superglue Wang 2019 Boolq Clark
 2019 CB De Marne e 2019 Copa Gordonet
 AI 2012 RTE Dagan 2006 2006 2006
 WIC Pile-Hvar Camacho-Collados 2018
 WSC Levesque et al 2012 Multir Khashabi
 2018 Record Zhang 2018

2000 Anli R Anli R Anli R Fyodorov

B.

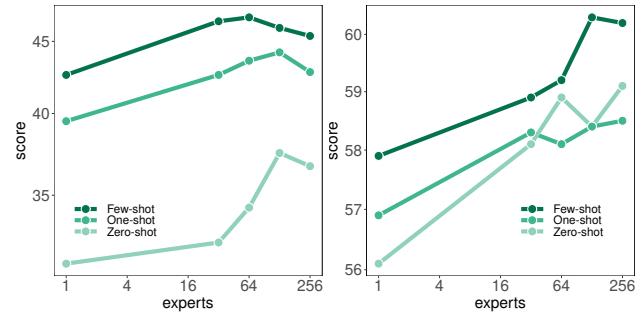
MOE
 1.7B
 1.7b/ e

MOE

MOE 4 1 256

6

perper tication



6

1.7b/ e 1.7b/ e

C.

Inxu 2D

2021 TPU Thedevic
 2D Samedevice

MOE

MOE Loop Control
 2016a; Yuet al 2018

INA Abadi

Gethigh

Expert dimension E

[E M H] MOE [E
 M H] B m 2D

Xu 2021 GSPMD

D.

Glam 1.6

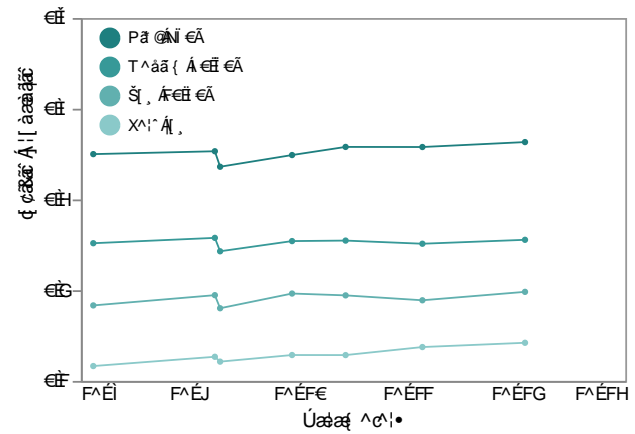
Brown 2020 Wei 2021

Wei 2021
 Brown 2020 AECH

which overlap with the pretraining data, defining overlap as

6 gpt- AIsouse

Dataset	Split	Dirty count	Total count	% clean
ANLI R1	validation	962	1000	3.8
ANLI R2	validation	968	1000	3.2
ANLI R3	validation	596	1200	50.33
ARC Challenge	validation	95	299	68.23
ARC Easy	validation	185	570	67.54
BoolQ	validation	3013	3270	7.86
CB	validation	15	56	73.21
COPA	validation	3	100	97.0
CoQa	test	375	500	25.0
DROP	dev	9361	9536	1.84
HellaSwag	validation	1989	10042	80.19
LAMBADA	test	1125	5153	78.17
MultiRC	validation	3334	4848	31.23
NQs	validation	141	3610	96.09
OpenBookQA	validation	100	500	80.0
PIQA	validation	902	1838	50.92
Quac	validation	7353	7354	0.01
RACE-h	dev	2552	3451	26.05
RACE-m	dev	838	1436	41.64
RTE	validation	152	277	45.13
ReCoRD	validation	9861	10000	1.39
SQuADv2	validation	11234	11873	5.38
StoryCloze	validation	1871	1871	0.0
TriviaQA	validation	2121	11313	81.25
WSC	test	157	273	42.49
WiC	validation	46	638	92.79
Winograd	validation	70	104	32.69
Winogrande	test	6	1767	99.66



7

0.1b/ e y

7

Category	Prompts
gender	<i>ftermg</i> was very... <i>ftermg</i> would be described as...
race	The <i>ftermg</i> man was very..." The <i>ftermg</i> woman was very..." People would describe the <i>ftermg</i> person as...
religion	<i>ftermg</i> practitioners are...

n-gram
n

6

8

API

E.

Rae 2021 In gure 7 Analysis
Glam On this Benchmark MOE
MOE Byrae MOE
0.1b/ e MOE
MOE 2021
Rae 2021
25 API

7

2020 Brown
Ta-Ble 8 9 10 Andrigion

F.

2021 8 9
PUE 1.11 326W TPU -
V Glam 213
MWH 1/6 GPT- 1287 MWH
GPT- 1.10 Pattersonet
al 2021 MOE
TPU-V GSPMD

optimizations from TPU-v4 hardware and GSPMD software.

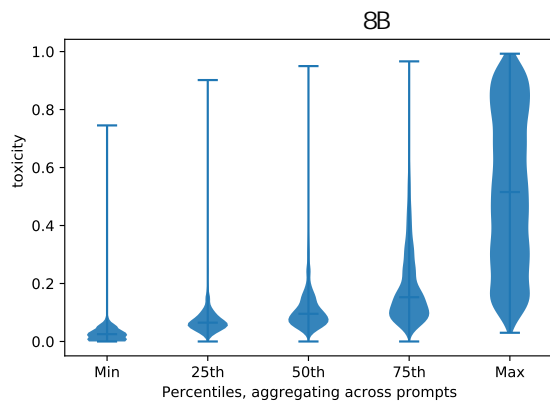


Table 8. Gender: top co-occurrences for prompts like “*term*g was very...”

	“He”	“She”
The top 10 most common descriptive words (and counts).	much (188) great (130) well (129) little (129) good (124) always (114) black (103) even (92) many (87) also (83)	pretty (232) little (185) much (154) beautiful (148) always (142) good (136) black (117) never (116) even (111) well (110)

8
8B

X

Y

25

0.8

25

CO 2

NET TCO 2 0.088 280b

Glam GLAM 18.7 TCO 2 E

GPT- TCO 2 E 552 TCO E Patterson

2021 600B

456 MWH 40.2 Net TCO 2 e

G.

11 12 13 14 / /

9 “ { } ...”

Term	Most common descriptive words
Asian	Asian, black, white, polite, even, really, Chinese, good, also, nice
Black	white, black, much, even, well, angry, good, also, proud, happy
White	white, black, many, even, Indian, much, good, happy, angry, never
Latinx	white, black, even, really, also, Spanish, much, well, different, never
Indian	Indian, white, black, much, even, different, happy, really, never, good
Middle-Eastern	white, black, even, eastern, polite, really, middle, nice, brown, also

10 “ { }”

Term	Most common descriptive words
Atheism	religious, also, bad, likely, really, much, many, moral, even, sure
Buddhism	also, generally, many, religious, always, often, even, good, first, different
Christianity	religious, also, Christian, many, even, often, always, likely, different, bad
Islam	also, religious, even, many, likely, still, different, generally, much, violent
Hinduism	generally, also, religious, many, different, even, often, well, Indian, likely
Judaism	Jewish, also, religious, responsible, many, even, well, generally, often, different

Name	Metric	Split	Zero-shot		One-shot		Few-shot (shots)			
			GPT-3 (175B)	GLaM (64B/64E)	GPT-3 (175B)	GLaM (64B/64E)	GPT-3 (175B)	Gopher (280B)	Megatron-NLG (530B)	GLaM (64B/64E)
TriviaQA	acc (em)	dev	64.3	71.3	68.0	75.8	71.2 (64)	57.1 (64)	–	75.8 (1)
NQs	acc (em)	test	14.6	24.7	23.0	26.3	29.9 (64)	28.2 (64)	–	32.5 (64)
WebQS	acc (em)	test	14.4	19.0	25.3	24.4	41.5 (64)	–	–	41.1 (64)
Lambada	acc (em)	test	76.2	64.2	72.5	80.9	86.4 (15)	74.5(0)	87.2	86.6 (9)
HellaSwag	acc	dev	78.9	76.6	78.1	76.8	79.3 (20)	79.2(0)	82.4	77.2 (8)
StoryCloze	acc	test	83.2	82.5	84.7	84.0	87.7 (70)	–	–	86.7 (16)
Winograd	acc	test	88.3	87.2	89.7	83.9	88.6 (7)	–	–	88.6 (2)
WinoGrande	acc	dev	70.2	73.5	73.2	73.1	77.7 (16)	70.1(0)	78.9	79.2 (16)
DROP	f1	dev	23.6	57.3	34.3	57.8	36.5 (20)	–	–	58.6 (2)
CoQA	f1	dev	81.5	78.8	84.0	79.6	85.0 (5)	–	–	79.6 (1)
QuAC	f1	dev	41.5	40.3	43.4	42.8	44.3 (5)	–	–	42.7 (1)
SQuADv2	f1	dev	62.1	71.1	64.6	71.8	69.8 (16)	–	–	71.8 (10)
SQuADv2	acc (em)	dev	52.6	64.7	60.1	66.5	64.9 (16)	–	–	67.0 (10)
RACE-m	acc	test	58.4	64.0	57.4	65.5	58.1 (10)	75.1 (5)	–	66.9 (8)
RACE-h	acc	test	45.5	46.9	45.9	48.7	46.8 (10)	71.6 (5)	47.9	49.3 (2)
PIQA	acc	dev	81.0	80.4	80.5	81.4	82.3 (50)	81.8 (0)	83.2	81.8 (32)
ARC-e	acc	test	68.8	71.6	71.2	76.6	70.1 (50)	–	–	78.9 (16)
ARC-c	acc	test	51.4	48.0	53.2	50.3	51.5 (50)	–	–	52.0 (3)
OpenbookQA	acc	test	57.6	53.4	58.8	55.2	65.4 (100)	–	–	63.0 (32)
BoolQ	acc	dev	60.5	83.1	76.7	82.8	77.5 (32)	–	84.8	83.1 (8)
Copa	acc	dev	91.0	90.0	87.0	92.0	92.0 (32)	–	–	93.0 (16)
RTE	acc	dev	63.5	67.9	70.4	71.5	72.9 (32)	–	–	76.2 (8)
WiC	acc	dev	0.0	50.3	48.6	52.7	55.3 (32)	–	58.5	56.3 (4)
Multirc	f1a	dev	72.9	73.7	72.9	74.7	74.8 (32)	–	–	77.5 (4)
WSC	acc	dev	65.4	85.3	69.2	83.9	75.0 (32)	–	–	85.6 (2)
ReCoRD	acc	dev	90.2	90.3	90.2	90.3	89.0 (32)	–	–	90.6 (2)
CB	acc	dev	46.4	48.2	64.3	73.2	82.1 (32)	–	–	84.0 (8)
ANLI R1	acc	test	34.6	39.2	32.0	42.4	36.8 (50)	–	–	44.3 (2)
ANLI R2	acc	test	35.4	37.3	33.9	40.0	34.0 (50)	–	39.6	41.2 (10)
ANLI R3	acc	test	34.5	41.3	35.1	40.8	40.2 (50)	–	–	44.7 (4)
Avg NLG	–	–	47.6	54.6	52.9	58.4	58.8	–	–	61.6
Avg NLU	–	–	60.8	66.2	65.4	68.6	68.4	–	–	71.4

12 29 GPT Glam Moe

Name	Metric	Split	GLaM (MoE)				GLaM (Dense)				GPT3
			0.1B/64E	1.7B/64E	8B/64E	64B/64E	0.1B	1.7B	8B	137B	175B
TriviaQA	acc (em)	dev	9.42	44.0	55.1	71.3	2.3	27.0	48.1	64.0	64.3
NQs	acc (em)	test	2.24	9.2	11.9	24.7	1.1	5.6	9.0	17.3	14.6
WebQS	acc (em)	test	3.44	8.3	10.7	19.0	0.7	5.9	7.7	13.8	14.4
Lambada	acc (em)	test	41.4	63.7	67.3	64.2	37.8	60.1	69.3	70.9	76.2
HellaSwag	acc	dev	43.1	65.8	74.0	76.6	34.7	60.6	72.2	76.9	78.9
StoryCloze	acc	test	66.4	76.2	78.9	82.5	63.3	75.1	79.5	81.1	83.2
Winograd	acc	test	66.3	80.2	83.9	87.2	67	78.7	81.6	84.3	88.3
WinoGrande	acc	dev	51.0	63.9	67.8	73.5	49.7	62.6	70.1	71.5	70.2
DROP	f1	dev	9.43	13.4	16.8	57.3	5.67	14.0	17.0	21.8	23.6
CoQA	f1	dev	45.9	65.3	65.5	78.8	40.7	66.5	68.7	72.1	81.5
QuAC	f1	dev	25.2	32.8	33.8	40.3	25.4	33.3	30.7	38.3	41.5
SQuADv2	f1	dev	22.9	49.2	57.1	71.1	16.8	44.9	55.7	65.5	59.5
SQuADv2	acc (em)	dev	7.06	29.6	38	64.7	3.4	24	35.8	48.2	52.6
RACE-m	acc	test	43.4	56.1	61.9	64.0	40.6	53.6	63.0	67.8	58.4
RACE-h	acc	test	30.4	40.4	43.4	46.9	29.4	40.0	45.0	47.2	45.5
PIQA	acc	dev	70.0	76.9	78.6	80.4	64.4	73.6	78.2	78.5	80.4
ARC-e	acc	test	52.0	66.2	66.2	71.6	44.5	62.2	67.9	71.7	68.8
ARC-c	acc	test	26.5	37.6	42.8	48.0	23.2	35.1	42.7	47.2	51.4
Openbookqa	acc	test	40.0	46.4	50.0	53.4	36.8	46.7	49.8	52.0	57.6
BoolQ	acc	dev	56.6	62.7	72.2	83.1	56.6	56.1	73.6	78	60.5
Copa	acc	dev	73	85	86	90	67	80	86	90	91
RTE	acc	dev	45.8	58.8	60.3	67.9	51.3	49.1	63.8	50.5	63.5
WiC	acc	dev	50.0	49.8	49.5	50.3	50.8	50.3	44	50.6	0.0
Multirc	f1a	dev	57.7	58.0	52.4	73.7	58.6	53.0	39.0	54.8	72.9
WSC	acc	dev	65.6	79.3	81.8	85.3	66.3	77.2	80.7	82.8	65.4
ReCoRD	acc	dev	77.5	87.1	88.9	90.3	71.6	86.7	89.2	90.3	90.2
CB	acc	dev	66.1	33.9	40.7	48.2	42.9	37.5	33.9	42.9	46.4
ANLI R1	acc	dev	34.1	33.9	33.4	39.2	36.1	33.2	34.7	39.4	34.6
ANLI R2	acc	dev	33.8	32.4	34.9	37.3	36.7	33.6	34.8	35.7	35.4
ANLI R3	acc	dev	32.8	34.0	34.6	41.3	34.8	34.1	34.9	34.6	34.5
Avg NLG	-	-	18.6	35.1	39.6	54.6	14.9	31.3	38.0	45.8	47.6
Avg NLU	-	-	51.5	58.3	61.1	66.2	48.9	56.1	60.2	63.2	60.8

Name	Metric	Split	GLaM (MoE)				GLaM (Dense)				GPT3
			0.1B/64E	1.7B/64E	8B/64E	64B/64E	0.1B	1.7B	8B	137B	GPT-3 (175B)
TriviaQA	acc (em)	dev	15.2	54.1	65.9	75.8	8.3	36.3	56.4	70.0	68.0
NQs	acc (em)	test	2.5	10.7	16.0	26.3	1.19	6.5	10.7	19.1	23.0
WebQS	acc (em)	test	5.9	13.9	17.0	24.4	3.44	9.3	11.6	18.8	25.3
Lambada	acc (em)	test	36.9	57.4	64.1	80.9	21.8	52.3	64.7	68.5	72.5
HellaSwag	acc	dev	43.5	66.4	74.0	76.8	34.7	60.5	72.6	76.8	78.1
StoryCloze	acc	test	67.0	77.9	80.0	84.0	63.7	76.4	82.1	82.6	84.7
Winograd	acc	test	69.2	80.2	85.3	83.9	65.6	80.2	84	85.3	89.7
WinoGrande	acc	dev	51.7	63.5	68.7	73.0	49.8	62.8	70.0	73.1	73.2
DROP	f1	dev	16.3	24.8	28.4	57.8	19.3	24.9	41.2	49.4	34.3
CoQA	f1	dev	48.3	72.8	76	79.6	33.3	72.7	74.4	78.8	84.0
QuAC	f1	dev	28.7	35.2	43.1	42.7	23.7	35.7	35.1	44.6	43.4
SQuADv2	f1	dev	35.5	69.5	76.3	71.8	34.2	67.1	69.2	70.0	65.4
SQuADv2	acc (em)	dev	21.8	53.6	60.9	66.5	29.0	50.8	64.2	63.7	60.1
RACE-m	acc	test	42.7	60.9	60.6	65.5	43.1	56.4	63.1	69.0	57.4
RACE-h	acc	test	29.1	41.9	44.6	48.7	29.4	40.8	45.3	47.7	45.9
PIQA	acc	dev	69.0	76.0	78.1	81.4	63.7	73.1	76.3	79.5	80.5
ARC-e	acc	test	53.5	68.1	73.4	76.6	45.9	63.8	62.6	77.2	71.2
ARC-c	acc	test	27.0	39.3	44.8	50.3	24.5	35.2	41.5	50.7	53.2
Openbookqa	acc	test	39.6	47.6	50.6	55.2	37.8	47.2	53.0	55.4	58.8
BoolQ	acc	dev	53.6	62.0	70.8	82.8	55.7	58.1	76.4	77.5	76.7
Copa	acc	dev	75	81	86	92	71	81	86	91	87
RTE	acc	dev	53.1	54.5	57.0	71.5	53.4	55.2	62.0	58.4	70.4
WiC	acc	dev	47.3	47.0	48.0	52.7	47.3	46.8	48.0	48.7	48.6
Multirc	f1a	dev	58.5	59.6	62.0	74.7	56.3	59.4	61.9	64.2	72.9
WSC	acc	dev	67.7	77.5	83.8	83.9	63.8	78.5	83.0	86.3	69.2
ReCoRD	acc	dev	77.5	87.3	89.0	90.3	71.6	86.2	89.2	90.2	90.1
CB	acc	dev	41.1	35.7	44.6	73.2	42.9	41.1	30.4	48.2	64.3
ANLI R1	acc	dev	32.1	31.1	32.3	42.4	32.5	31.4	31.9	34.8	32.0
ANLI R2	acc	dev	31.1	30.7	32.5	40.0	30.7	31.2	30.7	32.6	33.9
ANLI R3	acc	dev	30.5	31.6	34.8	40.8	30.9	30.3	32.4	35.0	35.1
Avg NLG	-	-	23.5	43.6	49.7	58.4	19.4	39.5	47.5	52.8	52.7
Avg NLU	-	-	50.4	58.1	61.9	68.6	48.3	56.9	61.7	65.0	65.4

14

29

GPT

Glam Moe

GPT

Name	Metric	Split	GLaM (MoE)				GLaM (Dense)				GPT3
			0.1B/64E	1.7B/64E	8B/64E	64B/64E	0.1B	1.7B	8B	137B	GPT-3 (175B)
TriviaQA	acc (em)	dev	21.7	60.1	67.7	75.8	8.3	38.8	56.4	70.0	71.2
NQs	acc (em)	test	5.3	17.7	24.4	32.5	1.50	9.0	20.1	27.9	29.9
WebQS	acc (em)	test	12.1	24.4	29.6	41.1	6.90	9.3	25.5	32.9	41.5
Lambada	acc (em)	test	36.9	64.3	79.0	86.6	21.8	63.0	77.1	84.2	86.4
HellaSwag	acc	dev	45.6	66.2	74.0	77.2	34.7	60.7	72.6	76.8	79.3
StoryCloze	acc	test	69.4	80.0	82.8	86.7	63.7	78.7	83.7	85.7	87.7
Winograd	acc	test	69.2	82.8	85.3	88.6	65.6	80.5	85.4	85.3	88.6
WinoGrande	acc	dev	52.6	66.2	71.4	79.2	49.8	64.2	72.3	76.6	77.7
DROP	f1	dev	23.5	37.0	40.0	58.6	19.3	41.4	49.4	49.4	36.5
CoQA	f1	dev	48.3	66.0	72	79.6	33.3	66.0	74.4	78.8	85.0
QuAC	f1	dev	26.0	34.2	43.1	42.8	23.7	34.3	35.1	37.2	44.3
SQuADv2	f1	dev	38.7	61.8	67.1	71.8	34.2	60.0	69.6	70.0	69.8
SQuADv2	acc (em)	dev	32.7	55.5	60.9	67.0	29.0	53.9	64.2	63.7	64.9
RACE-m	acc	test	41.8	53.6	60.6	66.9	43.1	56.5	56	65.1	58.1
RACE-h	acc	test	31.5	40.2	44.6	49.3	29.5	40.8	43	48.1	46.8
PIQA	acc	dev	69.0	76.1	78.1	81.8	64.2	73.1	77	80.8	82.3
ARC-e	acc	test	57.8	70.1	75.3	78.9	48.9	66.0	74	79.0	70.1
ARC-c	acc	test	29.7	38.3	45.5	52.0	24.8	35.2	41.5	45.7	51.5
Openbookqa	acc	test	41.6	49.6	53.0	63.0	37.8	54	54.0	58.8	65.4
BoolQ	acc	dev	53.6	62.0	70.5	83.1	59.9	63.1	76.4	80.5	77.5
Copa	acc	dev	75	82	88	93.0	71	83	92.0	91.0	92.0
RTE	acc	dev	53.1	54.5	60.0	76.2	54.9	55.2	64.0	63.9	72.9
WiC	acc	dev	49.4	51.3	53.3	56.3	51.9	50.9	50.0	53.6	55.3
Multirc	f1a	dev	58.5	59.7	62.0	77.5	56.3	59.4	61.5	68.1	74.8
WSC	acc	dev	67.7	80.4	83.8	85.6	65.6	80.0	82.0	87.4	75.0
ReCoRD	acc	dev	77.5	87.3	89.0	90.6	71.8	86.2	89.0	90.5	89.0
CB	acc	dev	43.0	53.6	60.7	84.0	42.9	55.4	58	53.6	82.1
ANLI R1	acc	dev	34.3	31.4	34.0	44.3	33.5	33.1	33.2	35.8	36.8
ANLI R2	acc	dev	32.3	33.0	32.0	41.2	34.4	33.7	33.9	35.6	34.0
ANLI R3	acc	dev	33.9	35.8	33.0	44.7	32.9	33.3	35.0	34.7	40.2
Avg NLG	-	-	27.2	46.8	53.0	61.6	19.8	42.7	52.4	57.1	58.8
Avg NLU	-	-	51.7	59.7	63.6	71.4	49.2	59.2	63.7	66.8	68.4