Additional file 9
Table S9: Summary of the Large Language Models

Model	Lab	Training Corpus	Vocabulary size	Model size (parameters) used/available	Layers (L), Hidden size (H), Heads (A)	Maximum sequence length	Date	Reference
GPT-2	OpenAI	WebText: 40 GB of text, 8M documents, from 45M webpages upvoted on Reddit (all Wikipedia pages removed)	50,257	124M/1.5B	L = 24, H = 1024, A = 16	1024	Feb 2019	[5]
GPT-3	OpenAI	570 GB plaintext, 0.4 trillion tokens. (410B CommonCrawl, 19B WebText2, 3B English Wikipedia, and two books corpora (12B Books1 and 55B Books2))		175B	L = 96, H = 12288, A = 96	2048	June 2020	[6]
GPT-neo	EleutherAI	The Pile - a large scale curated dataset created by EleutherAI 825GB, 380 billion tokens	50,257	1.3B/2.7B	L = 24, H = 2048, A = 16	2048	March 2021	[31]]
Bloom	BigScience	45 natural languages 12 programming languages In 1.5TB of pre-processed text, converted into 350B unique tokens	250,680	560M-3B/176B	L = 24, H = 1024, A = 16	2048	July 2022	[33]]
BioGPT	Microsoft	15M PubMed abstracts (trained from scratch)	42,384	347M/347M	L = 24, H = 1024, A = 16	1024	Oct 2022	[34]]
BioGPT-large	Microsoft	15M PubMed abstracts (trained from scratch)	42,384	1.5B/1.5B	L = 48, H = 1600, A = 25	1024	Oct 2022	[34]]
Galactica	Meta AI	106B tokens; Documents: 48M papers, 2M code, 8M reference material, 2M knowledge bases, 0.9M filtered commoncrawl, 1.3M prompts, 0.02M other	50,000	1.3B/120B	L = 12, H = 12, A = 64	2048	Nov 2022	[35]
ChatGPT	OpenAI			20B (175B)		3000 words (4000 sub-word tokens)	Nov 2022	
GPT-4	OpenAI						Feb 2023	

The time for each model is based on its first arXiv version (if exists) or estimated submission time.