srcs	bert	t5	gpt 1	gpt2	gpt3	gpt4	palm	llama	llama 2		comparision	Chanllenges & Goto Solutions
Company	Google	Google	OpenAI	OpenAI	OpenAI		Google	Facebook				N/A
Time	2018(after gpt1)	2019(after gpt2)	2018	2019	2020	2023		2023	2023	https://lifearchitect.ai/timeline/		N/A
architecture	encoder	enc-dec	casual-decoder	casual-decoder	casual-decoder	MOE		casual-decoder	casual-decoder			How to choose? Pros and Cons? Generalization? Why decoder has the ability? Triangle full rank? Difficulty in training, achieves better generation ability?
ModelSize	110M, 340M	220M-11B	117M	1.5B	175B	1.76T?	8,62,540B	7-65B	70B		comparision	How to choose model size? 7B？
Objective	MLM, NSP	Text-to-Text	CLM(Casual lang model)	CLM	CLM	Performance, alignment, Auxiliary Objective		CLM	CLM, auto-regressive			
Data	BookCorpus, EnWiki	C4(en)	BookCorpus	WebText(40GB)	Mix (Common Crawl, WebText2, Books1, Books2 and Wikipedia.)			public data	OpenSource		comparision	
Preprocessing	no	Task Prefix	no	task + {q}, {a}	InContextLearning							
Tokenizer	wordpiece	sentencepiece (wordpiece, 32000)	BPE	BPE(50257)	BPE (variant, SentencePiece)		SentencePiece 256k	BPE (variant, SentencePiece)	BPE (variant, SentencePiece, 32k)	tokenizer huggingface introduction	comparision	OOV? Matching? 南京市长江大桥 Vertical Area like medical? WordToVec huffman tree, similar words with quite different probability then they are on far-way leaves of the truee?
PositionalEncoding	absolute positional encoding, segment embedding, learnable	relative positional encoding	absolute positional encoding, learnable positional encoding	absolute positional encoding	absolute positional encoding			ROPE(variant)	ROPE(variant)			Feed into each en-decoder or just one? GPT padding on the right or right to left?
Attention	bidirectional	encoder: fully visible, decoder: casual attention mask. prefix LM					MQA	SparseAttention	GroupQueryAttention			Serving?
FFW+Activation+related	original,gelu	original	GLEU	GLEU, LayerNorm before decoder	PreNorm		SwiGLU	PreNorm, SwiGLU, 2.7x (instead of 4)	PreNorm (RMSNorm), SwiGLU			PreNorm to make sure the embedding is in a reasonalbe space? Otherwise, skip connection may accumurate the positional embedding etc.?
ContextLength	512	512	512	768,1024, 1280, 1600	2048		4,096,819,218,432	2k	4k			
Layers	12,24	12	12	48	96	120	32,64,118	32-80	32-80			
BatchSize	32	128	64	512	32M	16M		4M	4M		comparision	
Evaluation	Glue	cross-entropy loss			zero-one-few shot							
Training technique			pre-train + fine tune	one-train		RLHF						
Optimizer		AdaFactor						AdamW	AdamW			How many extra parameters? 2x, 3x, 12x?
Perpenxity												
Outcome												