

# GPT-2 initial baselines

## === TREC Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	5452.000000	5452.000000	5452.000000	5452.000000	5452.000000
mean	50.632062	67.119222	50.448092	45.423331	4.155984
std	20.743208	16.499787	19.287510	20.399941	1.508103
min	13.000000	33.000000	11.375000	0.000000	0.000000
25%	36.000000	55.000000	37.125000	32.000000	3.076923
50%	48.000000	66.000000	48.375000	40.000000	4.000000
75%	61.000000	77.000000	60.750000	56.000000	5.142857
max	196.000000	155.000000	164.250000	192.000000	11.000000

## === MedQuAD Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	16407.000000	16407.000000	16407.000000	16407.000000	16407.000000
mean	50.684952	67.827513	51.158339	48.875480	4.078711
std	16.926465	13.449138	16.027120	17.097808	1.500853
min	16.000000	36.000000	14.000000	8.000000	0.400000
25%	38.000000	58.000000	40.500000	36.000000	3.076923
50%	48.000000	67.000000	49.500000	48.000000	4.000000
75%	61.000000	77.000000	60.750000	60.000000	4.800000
max	191.000000	158.000000	166.500000	156.000000	11.428571

## === CaseHold Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	30612.000000	30612.000000	30612.000000	30612.000000	30612.000000
mean	129.843558	109.467006	113.725002	86.782308	4.060649
std	38.623308	20.343574	28.676497	32.082202	1.349260
min	12.000000	32.000000	10.500000	12.000000	0.400000
25%	101.000000	94.000000	92.250000	64.000000	3.111111
50%	131.000000	111.000000	114.750000	84.000000	4.000000
75%	161.000000	126.000000	136.125000	104.000000	4.857143
max	199.000000	160.000000	181.125000	292.000000	12.000000

# GPT-2 fine-tuned

== MedQuAD Fine-Tuned Dataset Results ==

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	3282.000000	3282.000000	3282.000000	3282.000000	3282.000000
mean	50.623400	67.723949	51.045094	22.828763	1.953329
std	17.375205	13.719419	16.440331	9.856418	0.878121
min	16.000000	36.000000	14.000000	12.000000	0.387097
25%	38.000000	57.000000	39.375000	20.000000	1.333333
50%	48.000000	67.000000	49.500000	20.000000	1.818182
75%	60.000000	77.000000	60.750000	28.000000	2.461538
max	154.000000	137.000000	140.625000	120.000000	7.000000

== CaseHold Fine-Tuned Dataset Results ==

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	6123.000000	6123.000000	6123.000000	6123.000000	6123.000000
mean	129.665033	109.366324	113.658031	61.814797	2.897218
std	38.502812	20.220943	28.608095	24.993884	1.090732
min	14.000000	34.000000	14.625000	12.000000	0.461538
25%	100.000000	94.000000	92.250000	44.000000	2.133333
50%	131.000000	111.000000	115.875000	60.000000	2.769231
75%	160.000000	125.000000	136.125000	76.000000	3.500000
max	199.000000	158.000000	175.500000	176.000000	10.000000



=== GPT-2 MedQuAD Fine-Tuned Performance on TREC Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	5452.000000	5452.000000	5452.000000	5452.000000	5452.000000
mean	50.632062	67.119222	50.448092	68.121790	6.089574
std	20.743208	16.499787	19.287510	27.814779	1.439281
min	13.000000	33.000000	11.375000	12.000000	1.500000
25%	36.000000	55.000000	37.125000	48.000000	5.090909
50%	48.000000	66.000000	48.375000	64.000000	6.000000
75%	61.000000	77.000000	60.750000	84.000000	7.000000
max	196.000000	155.000000	164.250000	248.000000	12.000000

=== GPT-2 MedQuAD Fine-Tuned Performance on Casehold Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	6123.000000	6123.000000	6123.000000	6123.000000	6123.000000
mean	129.665033	109.366324	113.658031	146.858729	6.770882
std	38.502812	20.220943	28.608095	43.295556	1.230562
min	14.000000	34.000000	14.625000	36.000000	2.838710
25%	100.000000	94.000000	92.250000	116.000000	5.894737
50%	131.000000	111.000000	115.875000	144.000000	6.714286
75%	160.000000	125.000000	136.125000	176.000000	7.578947
max	199.000000	158.000000	175.500000	356.000000	12.000000

=== GPT-2 Casehold Fine-Tuned Performance on TREC Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	5452.000000	5452.000000	5452.000000	5452.000000	5452.000000
mean	50.632062	67.119222	50.448092	70.949376	6.584099
std	20.743208	16.499787	19.287510	22.133490	1.672130
min	13.000000	33.000000	11.375000	32.000000	2.105263
25%	36.000000	55.000000	37.125000	56.000000	5.411765
50%	48.000000	66.000000	48.375000	68.000000	6.400000
75%	61.000000	77.000000	60.750000	84.000000	7.555556
max	196.000000	155.000000	164.250000	216.000000	12.000000

=== GPT-2 Casehold Fine-Tuned Performance on MedQuAD Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	3282.000000	3282.000000	3282.000000	3282.000000	3282.000000
mean	50.623400	67.723949	51.045094	76.978672	6.415512
std	17.375205	13.719419	16.440331	19.925192	1.597604
min	16.000000	36.000000	14.000000	32.000000	2.666667
25%	38.000000	57.000000	39.375000	64.000000	5.263158
50%	48.000000	67.000000	49.500000	76.000000	6.181818
75%	60.000000	77.000000	60.750000	88.000000	7.384615
max	154.000000	137.000000	140.625000	180.000000	12.000000

# OPT-1.3B initial baselines

=== OPT-1.3B on TREC Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	5452.000000	5452.000000	5452.000000	5452.000000	5452.000000
mean	50.632062	67.119222	50.448092	46.421864	3.887541
std	20.743208	16.499787	19.287510	18.402169	1.270400
min	13.000000	33.000000	11.375000	12.000000	0.750000
25%	36.000000	55.000000	37.125000	32.000000	3.000000
50%	48.000000	66.000000	48.375000	44.000000	3.750000
75%	61.000000	77.000000	60.750000	56.000000	4.666667
max	196.000000	155.000000	164.250000	180.000000	10.666667

=== OPT-1.3B on MedQuAD Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	3282.000000	3282.000000	3282.000000	3282.000000	3282.000000
mean	50.623400	67.723949	51.045094	49.907374	3.875231
std	17.375205	13.719419	16.440331	14.619296	1.265917
min	16.000000	36.000000	14.000000	12.000000	0.705882
25%	38.000000	57.000000	39.375000	40.000000	2.947368
50%	48.000000	67.000000	49.500000	48.000000	3.714286
75%	60.000000	77.000000	60.750000	60.000000	4.666667
max	154.000000	137.000000	140.625000	128.000000	9.142857

=== OPT-1.3B on CaseHold Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	6123.000000	6123.000000	6123.000000	6123.000000	6123.000000
mean	129.665033	109.366324	113.658031	96.179650	4.352379
std	38.502812	20.220943	28.608095	27.128195	1.255293
min	14.000000	34.000000	14.625000	24.000000	1.263158
25%	100.000000	94.000000	92.250000	76.000000	3.454545
50%	131.000000	111.000000	115.875000	92.000000	4.190476
75%	160.000000	125.000000	136.125000	112.000000	5.076923
max	199.000000	158.000000	175.500000	224.000000	12.000000

# OPT-1.3 Fine-tuned

=== OPT-1.3B MedQuAD Fine-Tuned on TREC Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	5452.000000	5452.000000	5452.000000	5452.000000	5452.000000
mean	50.632062	67.119222	50.448092	64.799707	5.263152
std	20.743208	16.499787	19.287510	27.094030	1.290636
min	13.000000	33.000000	11.375000	16.000000	1.333333
25%	36.000000	55.000000	37.125000	44.000000	4.363636
50%	48.000000	66.000000	48.375000	60.000000	5.200000
75%	61.000000	77.000000	60.750000	80.000000	6.000000
max	196.000000	155.000000	164.250000	228.000000	10.857143

=== OPT-1.3B MedQuAD Fine-Tuned on MedQuAD Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	3282.000000	3282.000000	3282.000000	3282.000000	3282.000000
mean	50.623400	67.723949	51.045094	16.319317	1.287250
std	17.375205	13.719419	16.440331	8.905501	0.680635
min	16.000000	36.000000	14.000000	8.000000	0.250000
25%	38.000000	57.000000	39.375000	8.000000	0.727273
50%	48.000000	67.000000	49.500000	16.000000	1.230769
75%	60.000000	77.000000	60.750000	20.000000	1.635294
max	154.000000	137.000000	140.625000	112.000000	5.555556

=== OPT-1.3B MedQuAD Fine-Tuned on CaseHold Dataset Results ===

	original_size_bytes	gzip_size_bytes	lzw_size_bytes	llm_bits_total	llm_bits_per_token
count	6123.000000	6123.000000	6123.000000	6123.000000	6123.000000
mean	129.665033	109.366324	113.658031	130.476564	5.760113
std	38.502812	20.220943	28.608095	41.153440	1.309593
min	14.000000	34.000000	14.625000	32.000000	2.000000
25%	100.000000	94.000000	92.250000	100.000000	4.800000
50%	131.000000	111.000000	115.875000	128.000000	5.714286
75%	160.000000	125.000000	136.125000	156.000000	6.608696
max	199.000000	158.000000	175.500000	380.000000	11.000000

