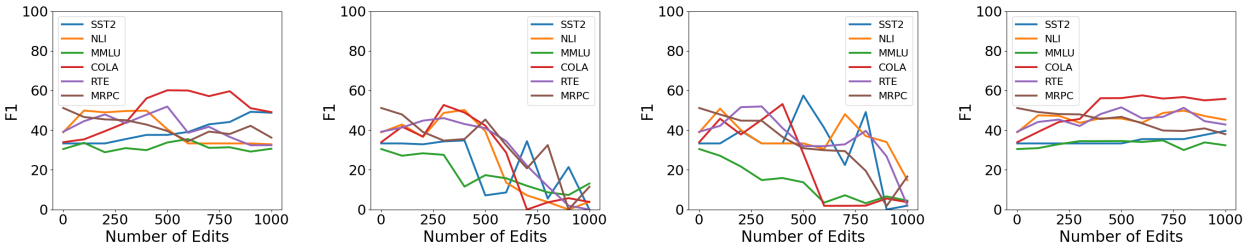
Table 3. Editing performance for GPT2-XL for different method using batch size 10

METHOD	MODEL	EDIT SCORE	PARAPHRASE SCORE	NEIGHBORHOOD SCORE	OVERALL SCORE	GENERATION ENTROPY
MEMIT	GPT2-XL	91.97	77.67	58.05	73.19	503.98
EMMET	GPT2-XL	82.77	69.42	52.59	65.932	574.1
ALPHAEDIT	GPT2-XL	91.24	73.71	56.39	70.988	586.77
ENCORE	GPT2-XL	92.57	78.19	60.36	74.703	510.81

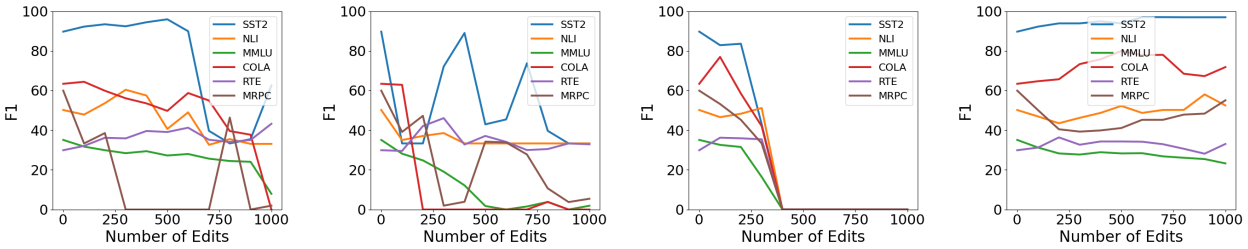


(a) Downstream Performance for GPT2-XL using MEMIT (b) Downstream Performance for GPT2-XL using EMMET (c) Downstream Performance for GPT2-XL using AlphaEdit (d) Downstream Performance for GPT2-XL using ENCORE

Figure 4. Downstream Performance for GPT2-XL for different methods using batch size 10

Table 4. Editing performance for Llama2-7B for different method using batch size 10

METHOD	MODEL	EDIT SCORE	PARAPHRASE SCORE	NEIGHBORHOOD SCORE	OVERALL SCORE	GENERATION ENTROPY
MEMIT	LLAMA2-7B	81.73	65.58	66.22	70.447	522.72
EMMET	LLAMA2-7B	93.11	84.0	54.84	73.386	541.54
ALPHAEDIT	LLAMA2-7B	56.92	51.38	53.73	53.915	494.64
ENCORE	LLAMA2-7B	90.79	79.76	59.7	74.437	555.18

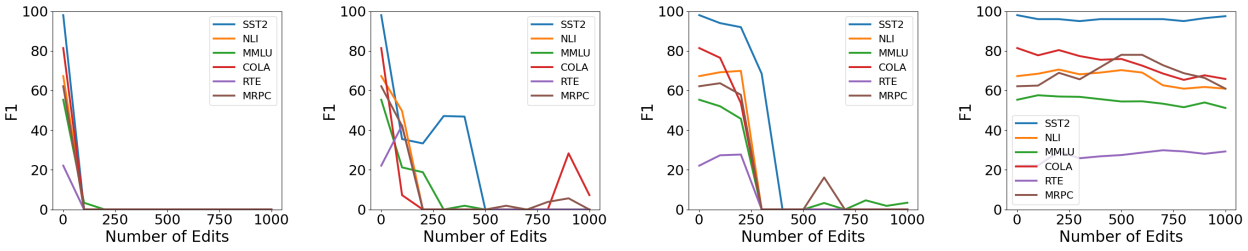


(a) Downstream Performance for Llama2-7B using MEMIT (b) Downstream Performance for Llama2-7B using EMMET (c) Downstream Performance for Llama2-7B using AlphaEdit (d) Downstream Performance for Llama2-7B using ENCORE

Figure 5. Downstream Performance for Llama2-7B for different methods using batch size 10

Table 5. Editing performance for Llama3-8B for different method using batch size 10

METHOD	MODEL	EDIT SCORE	PARAPHRASE SCORE	NEIGHBORHOOD SCORE	OVERALL SCORE	GENERATION ENTROPY
MEMIT	LLAMA3-8B	50.32	49.56	50.59	50.153	281.88
EMMET	LLAMA3-8B	86.16	77.3	47.99	66.108	553.34
ALPHAEDIT	LLAMA3-8B	57.8	56.55	49.7	54.441	430.37
ENCORE	LLAMA3-8B	89.71	80.48	57.63	73.306	539.91

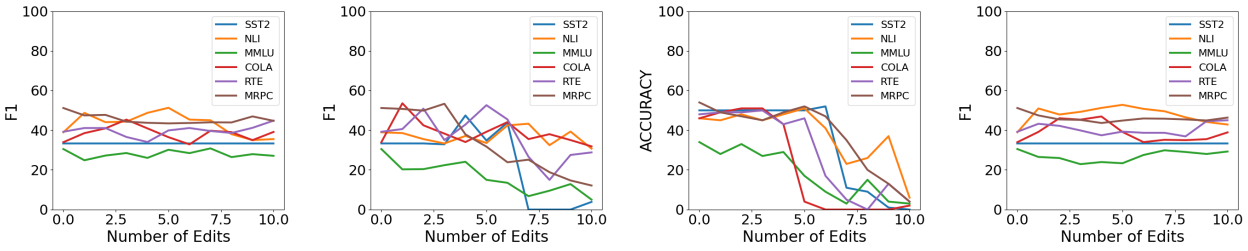


(a) Downstream Performance for Llama3-8B using MEMIT (b) Downstream Performance for Llama3-8B using EMMET (c) Downstream Performance for Llama3-8B using AlphaEdit (d) Downstream Performance for Llama3-8B using ENCORE

Figure 6. Downstream Performance for Llama3-8B for different methods using batch size 10

Table 6. Editing performance for GPT2-XL for different method using batch size 1000

METHOD	MODEL	EDIT SCORE	PARAPHRASE SCORE	NEIGHBORHOOD SCORE	OVERALL SCORE	GENERATION ENTROPY
MEMIT	GPT2-XL	93.55	81.31	59.66	75.472	558.93
EMMET	GPT2-XL	88.04	73.46	53.42	68.664	600.62
ALPHAEDIT	GPT2-XL	92.61	76.78	56.09	72.028	587.19
ENCORE	GPT2-XL	92.57	80.01	61.25	75.71	566.94

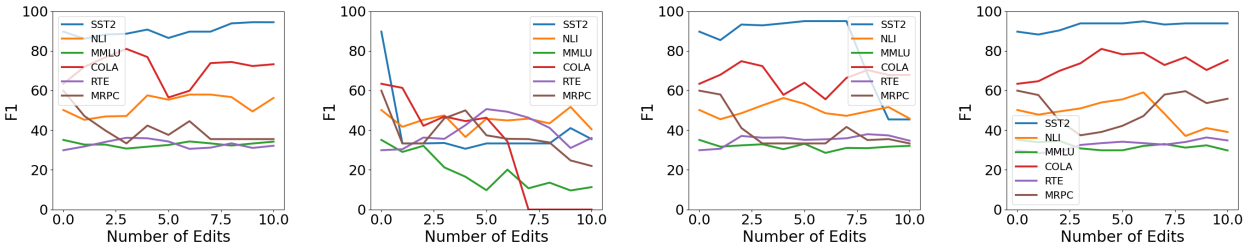


(a) Downstream Performance for GPT2-XL using MEMIT (b) Downstream Performance for GPT2-XL using EMMET (c) Downstream Performance for GPT2-XL using AlphaEdit (d) Downstream Performance for GPT2-XL using ENCORE

Figure 7. Downstream Performance for GPT2-XL for different methods using batch size 1000

Table 7. Editing performance for Llama2-7B for different method using batch size 1000

METHOD	MODEL	EDIT SCORE	PARAPHRASE SCORE	NEIGHBORHOOD SCORE	OVERALL SCORE	GENERATION ENTROPY
MEMIT	LLAMA2-7B	93.94	76.83	67.64	78.034	577.69
EMMET	LLAMA2-7B	92.37	80.65	56.89	73.524	576.26
ALPHAEDIT	LLAMA2-7B	96.11	87.27	61.63	78.762	588.09
ENCORE	LLAMA2-7B	92.71	79.86	67.58	78.729	591.8

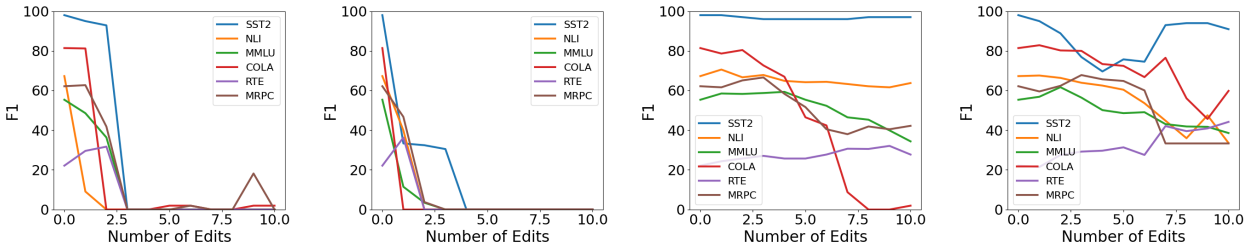


(a) Downstream Performance for Llama2-7B using MEMIT (b) Downstream Performance for Llama2-7B using EMMET (c) Downstream Performance for Llama2-7B using AlphaEdit (d) Downstream Performance for Llama2-7B using ENCORE

Figure 8. Downstream Performance for Llama2-7B for different methods using batch size 1000

Table 8. Editing performance for Llama3-8B for different method using batch size 1000

METHOD	MODEL	EDIT SCORE	PARAPHRASE SCORE	NEIGHBORHOOD SCORE	OVERALL SCORE	GENERATION ENTROPY
MEMIT	LLAMA3-8B	74.45	61.25	57.33	63.558	457.97
EMMET	LLAMA3-8B	79.67	68.27	49.62	63.354	537.03
ALPHAEDIT	LLAMA3-8B	91.29	76.88	68.95	77.994	593.35
ENCORE	LLAMA3-8B	93.21	81.12	71.53	81.002	579.51



(a) Downstream Performance for Llama3-8B using MEMIT (b) Downstream Performance for Llama3-8B using EMMET (c) Downstream Performance for Llama3-8B using AlphaEdit (d) Downstream Performance for Llama3-8B using ENCORE

Figure 9. Downstream Performance for Llama3-8B for different methods using batch size 1000