Proceedings of Machine Learning Research 304, 2025 ACML 2025

Supplementary Material for Balancing Knowledge Updates: Toward Unified Modular Editing in LLMs

Jiahao Liu LJIAHAO@STU2023.JNU.EDU.CN
Jinan University

Zijian Wang ZWAN0998@UNI.SYDNEY.EDU.AU
The University of Sydney

Kuo Zhao* ZHAOKUO@JNU.EDU.CN
Dong Hu HUDONG@STU2023.JNU.EDU.CN
Jinan University

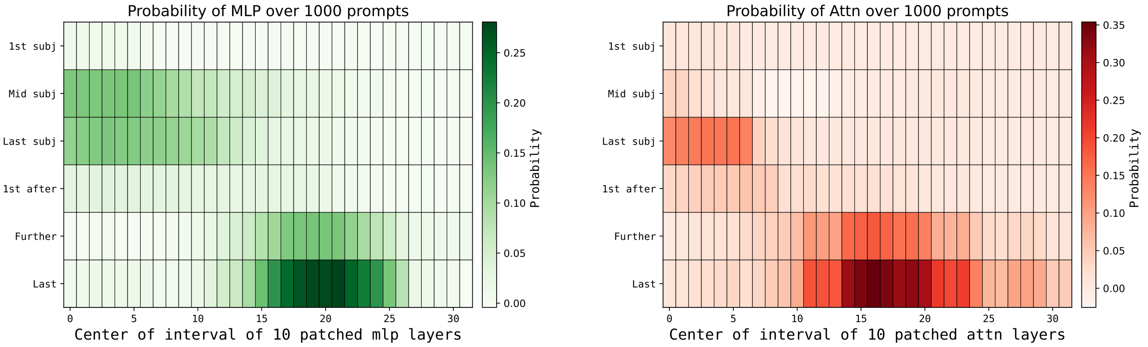
Editors: Hung-yi Lee and Tongliang Liu

Appendix A. Causal Tracing Analysis on Mistral-7B

This supplement provides additional experimental evidence on Mistral-7B.

Figure 1 presents the heatmaps of causal influence from MLP and Attn layers on factual recall. Panel (a) demonstrates that the causal effect of the MLP module is concentrated in the middle layers, especially when interventions target the last subject token, confirming that MLP layers are primary repositories for factual knowledge. Panel (b) shows that the Attn module exerts its strongest influence in the earlier layers, particularly for the last subject token, with this effect diminishing rapidly in deeper layers.

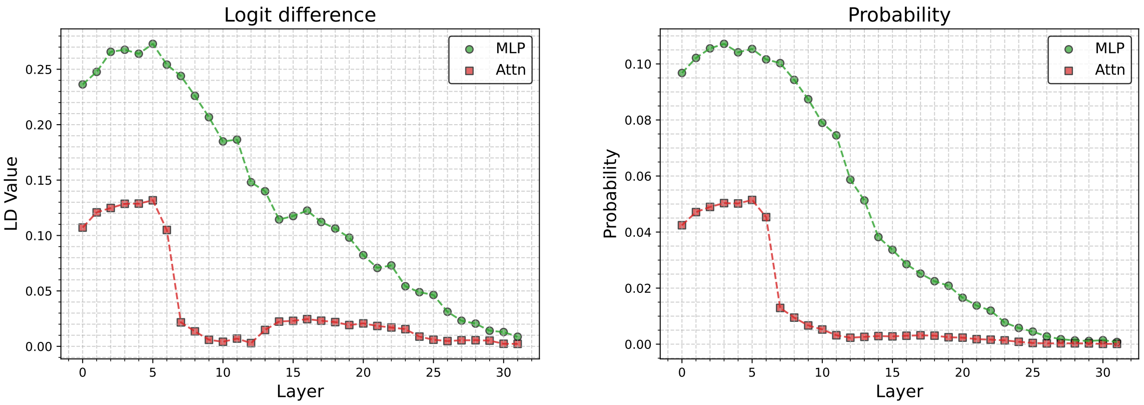
These patterns in Mistral-7B are highly consistent with the experimental results presented in the main text. The heatmaps demonstrate that MLP modules play a central role in factual knowledge recall at the middle layers, while Attn modules contribute most prominently in the earlier layers. This evidence further supports our main conclusion that both MLP and Attn modules are critical for factual memory storage and retrieval.



(a) Probability of MLP over 1000 prompts (b) Probability of Attn over 1000 prompts

Figure 1: Heatmaps of causal influence from MLP and Attn layers on factual recall in Mistral-7B.

Figure 2 shows the layerwise causal effects of Attn and MLP modules at the last subject token position in Mistral-7B. The logit difference and probability metrics consistently indicate that the Attn module exerts its strongest causal influence in the earliest layers (roughly layers 1–5), after which its effect rapidly decreases to near zero. These results further support the conclusion that Attn modules are most critical for factual recall at the initial stages of processing, while MLP modules play a more prominent role in deeper layers.



(a) Logit Difference

(b) Probability

Figure 2: Layerwise Causal Effects of Attn and MLP at the Last Subject Token.

Appendix B. More Experimental Details

B.1. Detailed Metric Values for Knowledge Balancing Strategy

We provide detailed evaluation results of the knowledge balancing strategy on the ZsRE benchmark, reporting the full metric values for both Mistral-7B and Qwen2.5-7B under different balance factors α.

Table 1: Performance metrics for Mistral-7B on ZsRE across different α values.

α	0.00	0.06	0.08	**0.10**	0.12	0.16	0.18	0.20	0.22	0.24	0.26	0.28	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
Edit Succ.	91.66	92.80	92.74	**92.70**	92.95	92.84	92.64	92.04	92.17	91.66	91.00	90.37	89.10	80.54	66.31	55.83	47.55	44.11	41.57	39.19
Portability	55.71	55.45	55.56	**56.25**	55.72	55.52	55.36	55.11	54.98	54.96	54.58	54.31	55.13	53.04	52.26	51.71	50.82	49.62	49.03	48.39
Locality	33.88	34.74	35.05	**34.88**	35.48	36.04	36.32	36.33	36.62	36.91	37.12	37.43	38.37	39.32	41.95	43.80	44.69	45.13	45.26	44.82
Fluency	564.61	565.62	565.93	**568.08**	566.85	567.56	570.62	570.06	568.04	569.33	574.33	574.30	569.28	574.72	575.36	574.82	572.80	574.13	573.68	575.49

Table 2: Performance metrics for Qwen2.5-7B on ZsRE across different α values.

α	0.00	0.10	0.12	0.14	0.16	0.18	0.20	0.22	0.24	0.26	0.28	**0.30**	0.32	0.34	0.36	0.38	0.40	0.50	0.60	0.70	0.80	0.90	1.00
Edit Succ.	95.43	96.65	96.95	97.13	97.29	97.65	97.79	97.18	97.06	96.94	96.86	**96.98**	96.25	95.42	94.74	94.04	93.07	84.96	70.32	54.85	44.97	39.53	36.63
Portability	53.88	53.77	53.98	54.12	54.05	54.10	54.06	53.96	53.81	53.84	53.90	**55.04**	53.81	53.61	53.62	53.83	54.04	53.38	51.47	49.60	48.27	47.32	45.64
Locality	32.58	33.78	33.87	33.87	34.04	33.98	34.01	34.32	34.48	34.65	34.92	**34.40**	35.15	35.15	35.47	36.19	36.46	37.93	38.44	38.92	38.80	38.16	37.48
Fluency	569.84	572.18	574.87	575.23	571.93	573.06	574.23	572.00	571.59	573.98	571.85	**572.77**	573.85	572.51	576.27	576.94	575.00	580.49	583.12	590.12	594.30	598.59	598.27

Discussion: The tables above report the raw metric values for Edit Success, Portability, Locality, and Fluency as the balance factor α varies. We observe that moderate values of

α consistently provide the best balance between factual editing accuracy and knowledge retention.

B.2. Attn-only vs MLP-only vs Joint (IntAttn-Edit) Ablation

We instantiate Attn-only ($\alpha = 1$) and MLP-only ($\alpha = 0$) as special cases of our framework. Joint editing uses the α determined by our knowledge-balancing strategy.

Table 3: Attn-only vs MLP-only vs Joint (IntAttn-Edit) on ZsRE. Joint editing balances strengths from both modules.

Model	Method	Edit Success	Portability	Locality	Fluency
3*Qwen2.5-7B	MLP-only ($\alpha = 0$)	95.43	53.88	32.58	569.84
	Attn-only ($\alpha = 1$)	36.64	45.64	**37.48**	**598.27**
	IntAttn-Edit	**96.98**	**55.04**	34.40	572.78
3*Mistral-7B	MLP-only ($\alpha = 0$)	91.66	55.71	33.88	564.61
	Attn-only ($\alpha = 1$)	39.19	48.39	**44.82**	**575.49**
	IntAttn-Edit	**92.70**	**56.25**	34.88	568.08

Takeaway. MLP updates contribute more to *Edit Success* and *Portability*, whereas Attn updates favor *Locality* and *Fluency*. The knowledge-balancing strategy integrates these complementary strengths, yielding a strong overall trade-off.

B.3. Cross-framework generality: IntAttn-Edit for ROME

To assess generality across editing paradigms, we port our method to **ROME** while preserving its *single-layer* localization and linearized solve. The port converts ROME into a *dual-path* (MLP+Attn) edit guided by the same balance principle (α), with the rest of the framework unchanged.

Table 4: ZsRE results for ROME vs. IntAttn-Edit for ROME. The joint update yields consistent gains without fluency degradation.

Dataset	Metric	ROME	IntAttn-Edit for ROME
4*ZsRE	Edit Success	79.06	**89.56**
	Portability	39.37	**48.07**
	Locality	19.06	**24.99**
	Fluency	561.96	**568.33**

Gains. Absolute: +10.50 (Edit), +8.70 (Port), +5.93 (Loc), +6.37 (Flu). Relative: +13.28% (Edit), +22.10% (Port), +31.11% (Loc); Fluency +1.13%. Without altering ROME's single-layer paradigm, the α-guided joint update yields simultaneous improvements in *Edit Success*, *Portability*, and *Locality*, with no fluency degradation—supporting generality across *backbones* (Mistral/Qwen) and *editing frameworks* (ROME → IntAttn-Edit for ROME).

B.4. General Capabilities After Editing (GSM8K 2-shot)

To evaluate preservation of overall capabilities after knowledge editing, we report **GSM8K (2-shot)** results after performing **100 edits**.

Table 5: General capabilities after editing on GSM8K (2-shot). IntAttn-Edit achieves high edit success with minimal impact on general capabilities; no evidence of catastrophic forgetting.

Model	Metric	Before Edit	IntAttn-Edit	MEMIT	ROME	R-ROME	AlphaEdit
2*Qwen2.5-7B	ACC	65.88	**65.81**	67.10	60.05	60.28	65.47
	PPL	3.06	**3.12**	3.10	3.19	3.20	3.13
2*Mistral-7B	ACC	28.35	**27.37**	28.89	0.00	0.00	25.74
	PPL	3.15	**3.17**	3.16	3868.80	6948.70	3.41

Conclusion. IntAttn-Edit maintains general capabilities while achieving high edit success, with *no evidence of catastrophic forgetting*