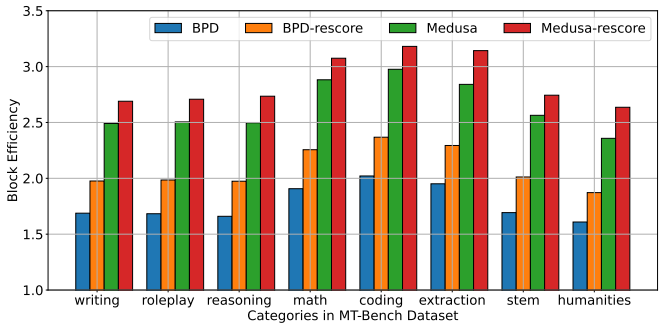
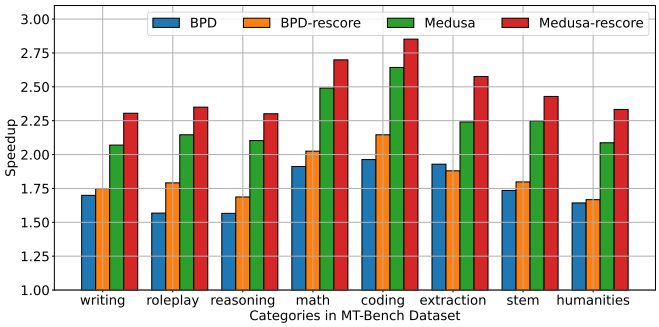
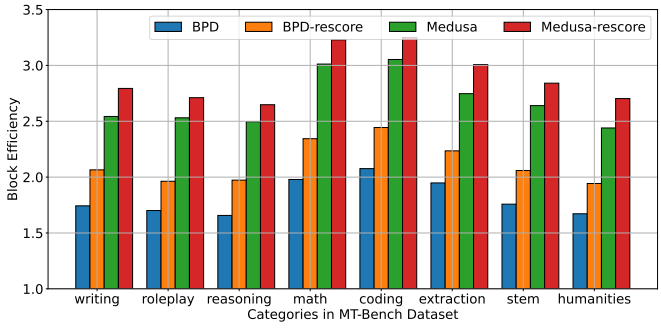


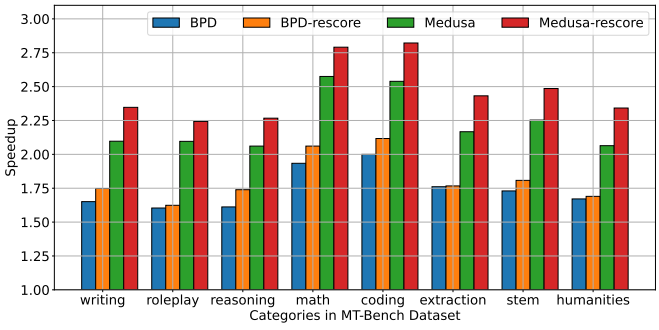
(a) Vicuna 7B - Block efficiency



(b) Vicuna 7B - Speed up



(c) Vicuna 13B - Block efficiency



(d) Vicuna 13B - Speed up

Figure A. Block efficiency (i.e. hardware-agnostic metric) and speedup ratio (i.e., latency) relative to the standard autoregressive decoding on MT-Bench dataset [4] having sub-categories as writing, roleplay, reasoning, math, coding, extraction, stem, and humanities. Speedup ratio (i.e., latency) relative to the standard autoregressive decoding is evaluated on a single **NVIDIA A100 GPU**. The local rescoring drafter is small decoder-only LM 68M.

Table A. **Block efficiency** (i.e. hardware-agnostic metric) on various additional datasets: MT-bench [4], summarization (CNN/Daily), question & anwering (QA) [3], GSM8K, and RAG [2]. Target model is Vicuna 7B v1.3 and the local rescoring drafter is small decoder-only LM 68M.

	Temperature=0.0					Temperature=0.7					Temperature=1.0				
Method	MT-bench	Sum	QA	GSM8K	RAG	MT-bench	Sum	QA	GSM8K	RAG	MT-bench	Sum	QA	GSM8K	RAG
BPD	1.768	1.496	1.501	1.752	1.502	1.879	1.539	1.539	1.834	1.574	1.918	1.569	1.612	1.879	1.620
+ Local rescoring	**2.073** ●	**1.715** ●	**1.744** ●	**2.018** ●	**1.759** ●	**2.289** ●	**1.777** ●	**1.850** ●	**2.204** ●	**1.918** ●	**2.410** ●	**1.813** ●	**1.962** ●	**2.331** ●	**1.985** ●
Medusa [1]	2.639	2.013	2.049	2.484	2.086	2.738	2.094	2.221	2.663	1.998	2.793	2.164	2.341	2.753	2.285
+ Local rescoring	**2.864** ●	**2.415** ●	**2.484** ●	**2.818** ●	**2.528** ●	**3.040** ●	**2.553** ●	**2.587** ●	**3.027** ●	**2.660** ●	**3.104** ●	**2.621** ●	**2.745** ●	**3.096** ●	**2.730** ●

Table B. **Speedup ratio** (i.e., latency) relative to the standard augtoregressive decoding on various additional datasets: MT-bench [4], summarization (CNN/Daily), question & anwering (QA) [3], GSM8K, and RAG [2]. Target model is Vicuna 7B v1.3 and the rescoring drafter is small decoder-only LM 68M. Results are evaluated on a single **NVIDIA A100 GPU** with a batch size 1.

	Temperature=0.0					Temperature=0.7					Temperature=1.0				
Method	MT-bench	Sum	QA	GSM8K	RAG	MT-bench	Sum	QA	GSM8K	RAG	MT-bench	Sum	QA	GSM8K	RAG
BPD	1.752	1.509	1.489	1.696	1.409	1.781	1.523	1.512	1.790	1.496	1.858	1.528	1.613	1.890	1.525
+ Local rescoring	**1.843** ●	**1.534** ●	**1.555** ●	**1.780** ●	**1.501** ●	**1.903** ●	**1.561** ●	**1.666** ●	**1.842** ●	**1.544** ●	**1.998** ●	**1.579** ●	**1.656** ●	**1.954** ●	**1.622** ●
Medusa [1]	2.254	2.002	2.045	2.317	1.833	2.425	2.094	2.121	2.563	1.998	2.511	2.096	2.334	2.650	2.010
+ Local rescoring	**2.482** ●	**2.076** ●	**2.114** ●	**2.357** ●	**2.000** ●	**2.597** ●	**2.228** ●	**2.279** ●	**2.630** ●	**2.139** ●	**2.657** ●	**2.281** ●	**2.386** ●	**2.655** ●	**2.173** ●

Table C. **Block efficiency** (i.e. hardware-agnostic metric) on various additional datasets: MT-bench [4], summarization (CNN/Daily), question & anwering (QA) [3], GSM8K, and RAG [2]. Target model is Vicuna 13B v1.3 and the rescoring drafter is small decoder-only LM 68M.

	Temperature=0.0					Temperature=0.7					Temperature=1.0				
Method	MT-bench	Sum	QA	GSM8K	RAG	MT-bench	Sum	QA	GSM8K	RAG	MT-bench	Sum	QA	GSM8K	RAG
BPD	1.817	1.531	1.489	1.796	1.531	1.914	1.596	1.602	1.907	1.620	1.978	1.618	1.623	1.941	1.669
+ Local rescoring	**2.124** ●	**1.767** ●	**1.733** ●	**2.098** ●	**1.799** ●	**2.287** ●	**1.853** ●	**1.982** ●	**2.249** ●	**1.901** ●	**2.375** ●	**1.889** ●	**2.022** ●	**2.302** ●	**1.977** ●
Medusa [1]	2.683	2.089	2.099	2.585	2.118	2.731	2.203	2.335	2.710	2.296	2.821	2.258	2.496	2.790	2.335
+ Local rescoring	**2.903** ●	**2.486** ●	**2.458** ●	**2.918** ●	**2.578** ●	**3.108** ●	**2.589** ●	**2.689** ●	**3.037** ●	**2.776** ●	**3.176** ●	**2.664** ●	**2.758** ●	**3.145** ●	**2.921** ●

Table D. **Speedup ratio** (i.e. latency) relative to the standard augtoregressive decoding on various additional datasets: MT-bench [4], summarization (CNN/Daily), question & anwering (QA) [3], GSM8K, and RAG [2]. Target model is Vicuna 13B v1.3 and the rescoring drafter is Vicuna 68M. Results are evaluated on a single **NVIDIA A100 GPU** with a batch size 1.

	Temperature=0.0					Temperature=0.7					Temperature=1.0				
Method	MT-bench	Sum	QA	GSM8K	RAG	MT-bench	Sum	QA	GSM8K	RAG	MT-bench	Sum	QA	GSM8K	RAG
BPD	1.745	**1.530**	1.488	1.794	1.483	1.881	1.555	1.559	1.875	1.558	2.043	1.684	1.675	2.078	1.664
+ Local rescoring	**1.819** ●	1.522 ●	**1.519** ●	**1.819** ●	**1.501** ●	**1.990** ●	**1.643** ●	**1.761** ●	**1.998** ●	**1.679** ●	**2.188** ●	**1.731** ●	**1.805** ●	**2.206** ●	**1.779** ●
Medusa [1]	2.232	2.000	1.986	2.507	1.945	2.559	2.080	2.272	2.700	2.069	2.844	2.278	2.501	2.942	2.256
+ Local rescoring	**2.467** ●	**2.136** ●	**2.154** ●	**2.519** ●	**2.068** ●	**2.738** ●	**2.211** ●	**2.368** ●	**2.700** ●	**2.293** ●	**2.981** ●	**2.392** ●	**2.574** ●	**3.010** ●	**2.447** ●