

Appendices

This appendix contains additional figures and tables for the thesis. Note that some tables are very small and might be hard to read in the printed version of the thesis. It is recommended to view the tables in the digital version of the thesis, available on GitHub alongside the other resources: <https://github.com/KaiPonel/MRLQuantization>.

Shortening	Task Name
q	Datatype of Embedding (Quantization Technique)
m	Dimensionality of Embedding (Matryoshka size)
l	Inference Layer Number
f32	32-Bit Floating-Point
int	8-Bit Integer
bin	Binary
bits req.	Bits Required to store embedding
CLA-1	Banking77Classification
CLA-2	EmotionClassification
CLA-3	TweetSentimentExtractionClassification
CLA-4	AmazonCounterfactualClassification
CLA-5	MassiveIntentClassification
CLA-6	MassiveScenarioClassification
CLA-7	MTOPDomainClassification
CLA-8	MTOPIntentClassification
CLU-1	ArXivHierarchicalClusteringP2P
CLU-2	ArXivHierarchicalClusteringS2S
CLU-3	BiorxivClusteringP2P.v2
CLU-4	BiorxivClusteringS2S.v2
CLU-5	MedrxivClusteringP2P.v2
CLU-6	MedrxivClusteringS2S.v2
CLU-7	RedditClustering.v2
CLU-8	StackExchangeClustering.v2
CLU-9	StackExchangeClusteringP2P.v2
CLU-10	TwentyNewsgroupsClustering.v2
STS-1	BIOSSES
STS-2	SICK-R
STS-3	STS12
STS-4	STS13
STS-5	STS14
STS-6	STS15
STS-7	STS16
STS-8	STSBenchmark
STS-9	STS17
STS-10	STS22
PCL-1	SprintDuplicateQuestions
PCL-2	TwitterSemEval2015
PCL-3	TwitterURLCorpus
RET-1	ArguAna
RET-2	CQADupstackWebmastersRetrieval
RET-3	NFCorpus
RER-1	AskUbuntuDupQuestions
RER-2	MindSmallReranking
RER-3	StackOverflowDupQuestions
SUM-1	SummEval

Table 2: Shortening and Task Name Mapping

A. Model Selection

Table 3: Top-100 models on the MTEB leaderboard as of 03.08.2024 at 12:00pm. Model size is measured in million parameters. Dimensions refers to the default embedding size of the model.

Rank	Model	Model Size	Memory Usage	Dimensions	Max Tokens	Avg. Score
1	bge-en-icl	7111	26.49	4096	32768	71.67
2	stella_en_1.5B_v5	1543	5.75	8192	131072	71.19
3	SFR-Embedding-2_R	7111	26.49	4096	32768	70.31
4	gte-Qwen2-7B-instruct-Q4_K_M-GGUF					70.24
5	gte-Qwen2-7B-instruct	7613	28.36	3584	131072	70.24
6	stella_en_400M_v5	435	1.62	8192	8192	70.11
7	bge-multilingual-gemma2	9242	34.43	3584	8192	69.88
8	NV-Embed-v1	7851	29.25	4096	32768	69.32
9	voyage-large-2-instruct			1024	16000	68.23
10	Linq-Embed-Mistral	7111	26.49	4096	32768	68.17
11	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56
12	gte-Qwen1.5-7B-instruct	7099	26.45	4096	32768	67.34
13	gte-Qwen2-1.5B-instruct-Q4_0-GGUF					67.16
14	gte-Qwen2-1.5B-instruct	1776	6.62	1536	131072	67.16
15	voyage-lite-02-instruct	1220	4.54	1024	4000	67.13
16	GritLM-7B	7242	26.98	4096	32768	66.76
17	e5-mistral-7b-instruct	7111	26.49	4096	32768	66.63
18	google-gecko.text-embedding-preview-0409	1200	4.47	768	2048	66.31
19	TDTE					65.96
20	GritLM-8x7B	46703	173.98	4096	32768	65.66
21	gte-large-en-v1.5	434	1.62	1024	8192	65.39
22	LLM2Vec-Meta-Llama-3-supervised	7505	27.96	4096	8192	65.01
23	LLM2Vec-Mistral-supervised	7111	26.49	4096	32768	64.8
24	echo-mistral-7b-instruct-lasttoken	7111	26.49	4096	32768	64.68

Continued on next page

Rank	Model	Model Size	Memory Usage	Dimensions	Max Tokens	Avg. Score
25	mxbai-embed-large-v1	335	1.25	1024	512	64.68
26	UAE-Large-V1	335	1.25	1024	512	64.64
27	text-embedding-3-large			3072	8191	64.59
28	voyage-lite-01-instruct			1024	4000	64.49
29	Cohere-embed-english-v3.0			1024	512	64.47
30	multilingual-e5-large-instruct	560	2.09	1024	514	64.41
31	google-gecko-256.text-embedding-preview-0409	1200	4.47	256	2048	64.37
32	GIST-large-Embedding-v0	335	1.25	1024	512	64.34
33	bge-large-en-v1.5	335	1.25	1024	512	64.23
34	b1ade-embed	335	1.25	1024	512	64.21
35	MUG-B-1.6	335	1.25	1024	512	64.2
36	LLM2Vec-Llama-2-supervised	6607	24.61	4096	4096	64.14
37	gte-base-en-v1.5	137	0.51	768	8192	64.11
38	Cohere-embed-multilingual-v3.0			1024	512	64.01
39	GIST-Embedding-v0	109	0.41	768	512	63.71
40	bge-base-en-v1.5	438	1.63	768	512	63.56
41	privacy_embedding_rag_10k_base_final	109	0.41	768	512	63.55
42	privacy_embedding_rag_10k_base_12_final	109	0.41	768	512	63.55
43	privacy_embedding_rag_10k_base_15_final	109	0.41	768	512	63.55
44	privacy_embedding_rag_10k_base_checkpoint_2	109	0.41	768	512	63.55
45	ember-v1	335	1.25	1024	512	63.54
46	sf_model_e5	335	1.25	1024	512	63.34
47	mxbai-embed-2d-large-v1	335	1.25	1024	512	63.25
48	gte-large	335	1.25	1024	512	63.13
49	NoInstruct-small-Embedding-v0	33	0.12	384	512	63.12
50	GIST-small-Embedding-v0	33	0.12	384	512	62.72
51	stella-base-en-v2	55	0.2	768	512	62.61
52	gte-base	109	0.41	768	512	62.39
53	UniVaR-lambda-80	137	0.51	768	8192	62.39

Continued on next page

Rank	Model	Model Size	Memory Usage	Dimensions	Max Tokens	Avg. Score
54	UniVaR-lambda-20	137	0.51	768	8192	62.39
55	UniVaR-lambda-1	137	0.51	768	8192	62.39
56	nomic-embed-text-v1	137	0.51	768	8192	62.39
57	UniVaR-lambda-5	137	0.51	768	8192	62.39
58	nomic-embed-text-v1.5	137	0.51	768	8192	62.28
59	text-embedding-3-small			1536	8191	62.26
60	e5-large-v2	335	1.25	1024	512	62.25
61	bge-small-en-v1.5	33	0.12	384	512	62.17
62	Cohere-embed-english-light-v3.0					62.01
63	text-embedding-3-large-256			256	8191	62.00
64	nomic-embed-text-v1.5-512	138	0.51	512	8192	61.96
65	LLM2Vec-Sheared-Llama-supervised	1280	4.77	2048	4096	61.85
66	instructor-xl	1241	4.62	768	512	61.79
67	instructor-large	335	1.25	768	512	61.59
68	e5-base-v2	109	0.41	768	512	61.50
69	e5-base-4k	112	0.42	768	4096	61.50
70	multilingual-e5-large	560	2.09	1024	514	61.50
71	e5-large	335	1.25	1024	512	61.42
72	nomic-embed-text-v1-ablated	137	0.51	768	8192	61.36
73	gte-small	33	0.12	384	512	61.36
74	nomic-embed-text-v1.5-256	138	0.51	256	8192	61.04
75	text-embedding-ada-002			1536	8191	60.99
76	udever-bloom-7b1	7069	26.33	4096	2048	60.63
77	e5-base	109	0.41	768	512	60.44
78	jina-embeddings-v2-base-en	137	0.51	768	8192	60.38
79	Titan-text-embeddings-v2					60.37
80	snowflake-arctic-embed-m-long	137	0.51	768	8192	60.09
81	Cohere-embed-multilingual-light-v3.0			384	512	60.08
82	e5-small-v2	33	0.12	384	512	59.93

Continued on next page

Rank	Model	Model Size	Memory Usage	Dimensions	Max Tokens	Avg. Score
83	udever-bloom-3b	3003	11.19	2560	2048	59.86
84	nomic-embed-text-v1-unsupervised	137	0.51	768	8192	59.85
85	snowflake-arctic-embed-l	334	1.24	1024	512	59.84
86	snowflake-arctic-embed-m	109	0.41	768	512	59.79
87	instructor-base	110	0.41	768	512	59.54
88	sentence-t5-xxl	4865	18.12	768	512	59.51
89	multilingual-e5-base	278	1.04	768	514	59.45
90	nomic-embed-text-v1.5-128	138	0.51	128	8192	59.34
91	XLM-3B5-embedding					59.29
92	GIST-all-MiniLM-L6-v2	23	0.08	384	512	59.00
93	gtr-t5-xxl	4865	18.12	768	512	58.97
94	snowflake-arctic-embed-s	33	0.12	384	512	58.94
95	SGPT-5.8B-weightedmean-msmarco-specb-bitfit	5874	21.88	4096	2048	58.93
96	e5-small	33	0.12	384	512	58.89
97	gte-tiny	23	0.08	384	512	58.69
98	gtr-t5-xl	1240	4.62	768	512	58.42
99	udever-bloom-1b1			1536	2048	58.29
100	gtr-t5-large	168	0.63	768	512	58.28

55

Table 4: Top-100 models on the MTEB leaderboard as of 03.08.2024 at 12:00pm which do not exceed 1 billion parameters. Model size is measured in million parameters.

Rank	Model	Model Size	Is Opensource?	Supports MRL?	Additional Notes
6	stella_en_400M_v5	435	y	y	
21	gte-large-en-v1.5	434	y	n	
25	mxbai-embed-large-v1	335	y	y	
26	UAE-Large-V1	335	y	n	
30	multilingual-e5-large-instruct	560	y	n	
32	GIST-large-Embedding-v0	335	y	n	
33	bge-large-en-v1.5	335	y	n	
34	b1ade-embed	335	y	n	
35	MUG-B-1.6	335	y	n	
37	gte-base-en-v1.5	137	y	n	
38	Cohere-embed-multilingual-v3.0		n	n	
40	bge-base-en-v1.5	438	y	n	
41	privacy_embedding_rag_10k_base_final	109	y	n	
42	privacy_embedding_rag_10k_base_12_final	109	y	n	
43	privacy_embedding_rag_10k_base_15_final	109	y	n	
44	privacy_embedding_rag_10k_base_checkpoint_2	109	y	n	
45	ember-v1	335	y	n	
46	sf_model_e5	335	y	n	
47	mxbai-embed-2d-large-v1	335	y	y	2D-MRL Model
48	gte-large	335	y	n	
49	NoInstruct-small-Embedding-v0	33	y	n	
50	GIST-small-Embedding-v0	33	y	n	
51	stella-base-en-v2	55	y	n	
52	gte-base	109	y	n	
53	UniVaR-lambda-80	137	y	n	
54	UniVaR-lambda-20	137	y	n	

Continued on next page

Rank	Model	Model Size	Is Opensource?	Supports MRL?	Additional Notes
55	UniVaR-lambda-1	137	y	n	
56	nomic-embed-text-v1	137	y	n	
57	UniVaR-lambda-5	137	y	n	
58	nomic-embed-text-v1.5	137	y	y	
60	e5-large-v2	335	y	n	
61	bge-small-en-v1.5	33	y	n	
64	nomic-embed-text-v1.5-512	138	y	y	Same model as 58.
67	instructor-large	335	y	n	
68	e5-base-v2	109	y	n	
69	e5-base-4k	112	y	n	
70	multilingual-e5-large	560	y	n	
71	e5-large	335	y	n	
72	nomic-embed-text-v1-ablated	137	y	n	
73	gte-small	33	y	n	
74	nomic-embed-text-v1.5-256	138	y	y	
77	e5-base	109	y	n	
78	jina-embeddings-v2-base-en	137	y	n	
80	snowflake-arctic-embed-m-long	137	y	n	
82	e5-small-v2	33	y	n	
84	nomic-embed-text-v1-unsupervised	137	y	n	
85	snowflake-arctic-embed-l	334	y	n	
86	snowflake-arctic-embed-m	109	y	n	
87	instructor-base	110	y	n	
89	multilingual-e5-base	278	y	n	
90	nomic-embed-text-v1.5-128	138	y	y	Same model as 58.
92	GIST-all-MiniLM-L6-v2	23	y	n	
94	snowflake-arctic-embed-s	33	y	n	
96	e5-small	33	y	n	
97	gte-tiny	23	y	n	

Continued on next page

Rank	Model	Model Size	Is Opensource?	Supports MRL?	Additional Notes
100	gtr-t5-large	168	y	n	

B. Additional Figures from Literature

B.1 MRL Figures

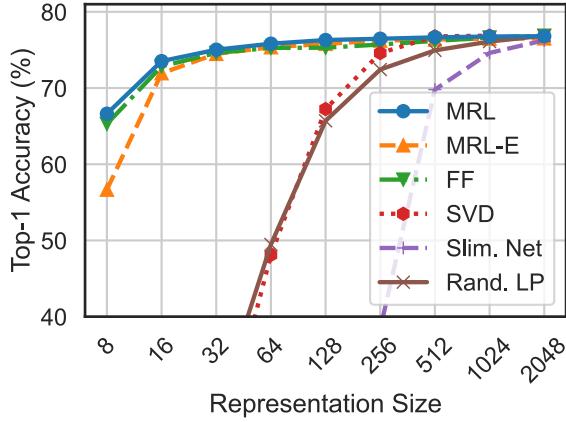


Figure 1: ImageNet-1K linear classification accuracy of ResNet50 models. MRL is as accurate as the independently trained FF models for every representation size. Figure and caption excerpted from the original paper [KBR⁺22].

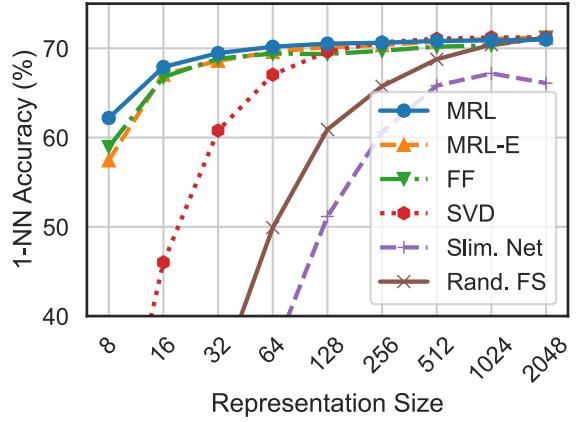


Figure 2: ImageNet-1K 1-NN accuracy of ResNet50 models measuring the representation quality for downstream task. MRL outperforms all the baselines across all representation sizes. Figure and caption excerpted from the original paper [KBR⁺22].

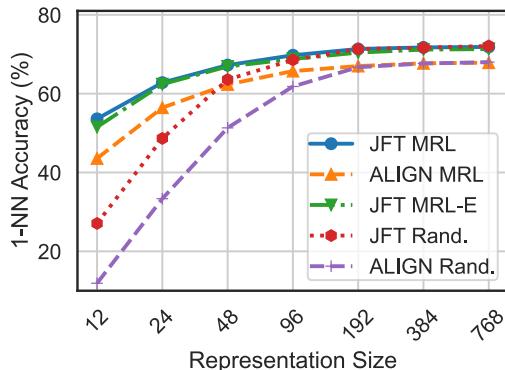


Figure 3: ImageNet-1K 1-NN accuracy for ViT-B/16 models trained on JFT-300M & as part of ALIGN. MRL scales seamlessly to web-scale with minimal training overhead. Figure and caption excerpted from the original paper [KBR⁺22].

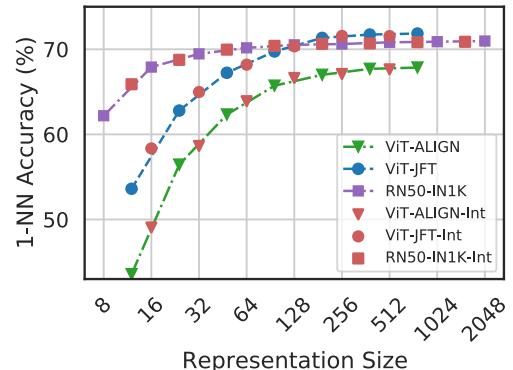


Figure 4: Despite optimizing MRL only for $O(\log(d))$ dimensions for ResNet50 and ViT-B/16 models; the accuracy in the intermediate dimensions shows interpolating behavior. Figure and caption excerpted from the original paper [KBR⁺22].

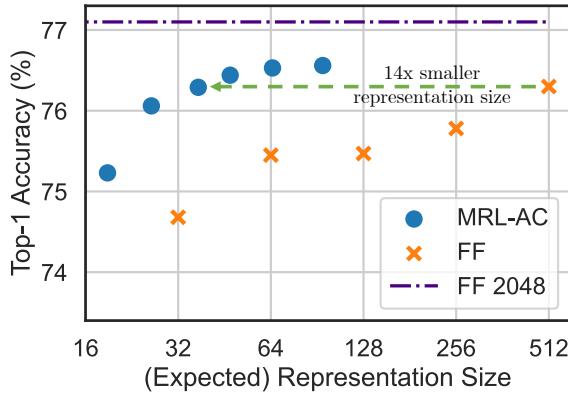


Figure 5: Adaptive classification on MRL ResNet50 using cascades results in $14\times$ smaller representation size for the same level of accuracy on ImageNet-1K (~ 37 vs 512 dims for 76.3%). Figure and caption excerpted from the original paper [KBR⁺²²].

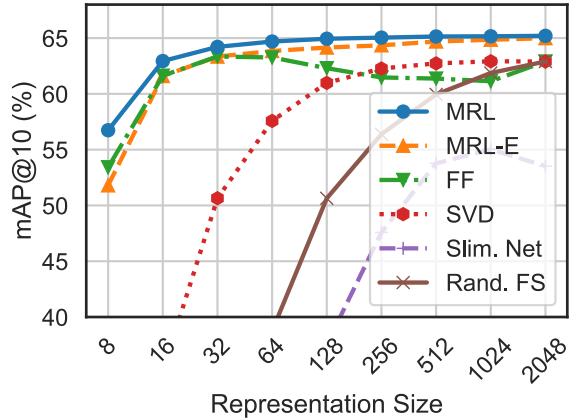
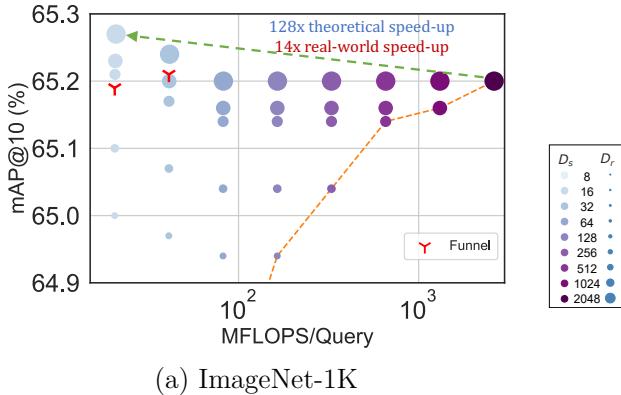
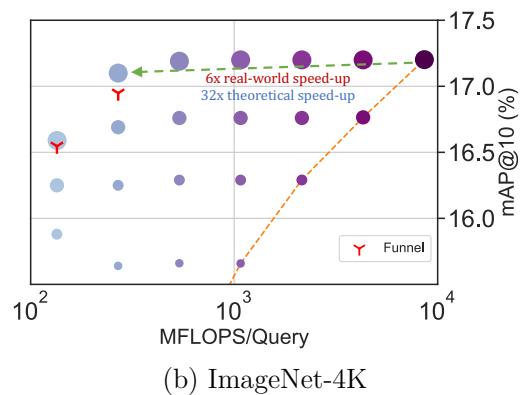


Figure 6: mAP@10 for Image Retrieval on ImageNet-1K with ResNet50. MRL consistently produces better retrieval performance over the baselines across all the representation sizes. Figure and caption excerpted from the original paper [KBR⁺²²].



(a) ImageNet-1K



(b) ImageNet-4K

Figure 7: The trade-off between mAP@10 vs MFLOPS/Query for Adaptive Retrieval (AR) on ImageNet-1K (left) and ImageNet-4K (right). Every combination of D_s & D_r falls above the Pareto line (orange dots) of single-shot retrieval with a fixed representation size while having configurations that are as accurate while being up to $14\times$ faster in real-world deployment. Funnel retrieval is almost as accurate as the baseline while alleviating some of the parameter choices of Adaptive Retrieval. Figure and caption excerpted from the original paper [KBR⁺²²].

B.2 2DMSE Figures

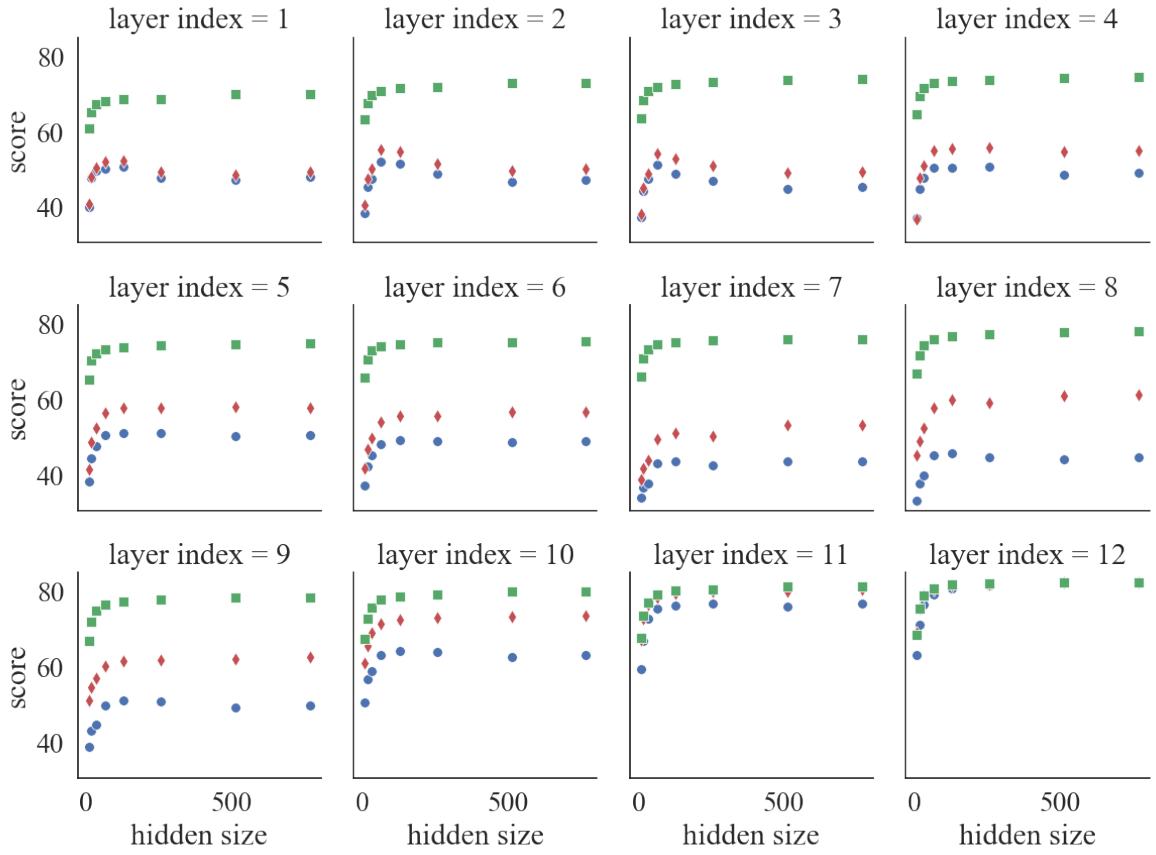
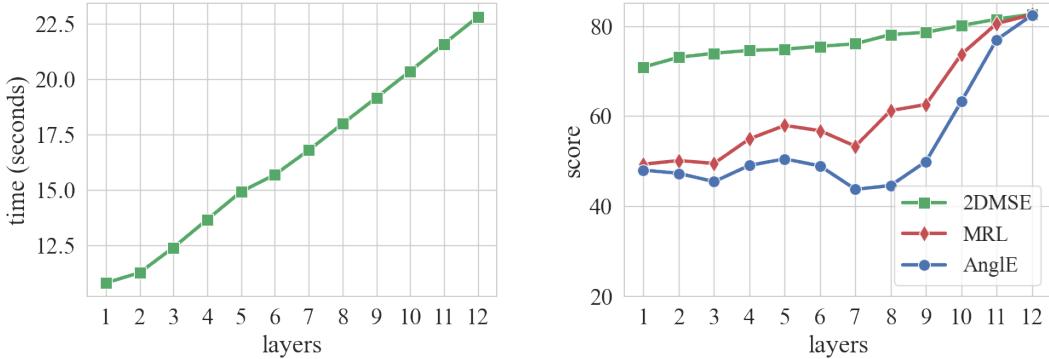


Figure 8: Results of the STS benchmark with a cascade of hidden sizes: $8 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 768$ from $\text{BERT}_{\text{base}}$. The score represents the average Spearman’s correlation. $\text{BERT}_{\text{base}}$ serves as the backbone for all models. The blue \bullet indicates the results of sentence embeddings from AngleE without any scalable sentence embedding learning. The red \blacklozenge represents the results of matryoshka sentence embeddings. The green \blacksquare denotes the results of our proposed 2D Matryoshka Sentence Embeddings (2DMSE). The layer index $= i$ denotes the i -th attention layer. Figure and caption excerpted from the original paper [LLL⁺24].



(a) Inference time *vs* number of layers. (b) Score on STS *vs* number of layers.

Figure 9: Subfigure (a) illustrates the time taken to use embeddings from different layers to encode the entire STS benchmarks. Subfigure (b) displays the average Spearman’s correlation scores of different layers. Both (a) and (b) use an embedding size of 768 and the standard STS benchmark dataset. Figure and caption excerpted from the original paper [LLL⁺24].

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe [RG19]	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
USE [RG19]	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
SBERT [RG19]	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SimCSE [GYC21]	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
AnglE [LL23]	75.09	85.56	80.66	86.44	82.47	85.16	81.23	82.37
MRL ($d = 768$) *	75.72	86.79	81.89	86.91	81.74	85.50	79.44	82.57
2DMSE ($n = 12$, $d = 768$)	75.00	86.69	82.30	86.50	82.09	85.79	80.18	82.65

Table 5: Full-capacity sentence embedding performance on the standard STS benchmark. Results * denote our implementation. BERT_{base} serves as the backbone for all models. Figure and caption excerpted from the original paper [LLL⁺24].

Model	Avg. Spearman's Correlation
$n = 12, d = 768$	
2DMSE	82.65
w/o alignment \mathcal{L}_{align}	82.57 (-0.08)
w/o last layer \mathcal{L}_N^D	81.31 (-1.34)
$n = 8, d = 512$	
2DMSE	78.02
w/o alignment \mathcal{L}_{align}	77.94 (-0.08)
w/o last layer \mathcal{L}_N^D	76.52 (-1.50)
$n = 6, d = 384$	
2DMSE	75.21
w/o alignment \mathcal{L}_{align}	75.08 (-0.13)
w/o last layer \mathcal{L}_N^D	74.98 (-0.23)
$n = 4, d = 256$	
2DMSE	73.93
w/o alignment \mathcal{L}_{align}	73.69 (-0.24)
w/o last layer \mathcal{L}_N^D	73.00 (-0.93)

Table 6: Ablation study results of 2DMSE on the standard STS benchmark using BERT_{base}. n denotes the number of Transformer layers, and d stands for the embedding dimensions. Figure and caption excerpted from the original paper [LLL⁺24].

C. mxbai-embed-large-v1

This section contains additional figures and tables for the MBAI-1D model.

q	m	CLA	CLU	CLU	CLU	CLU	CLU	CLU	CLU	CLU	CLU	CLU	STS	STS	STS	STS	STS	STS	STS	STS	PCL	PCL	PCL	RET	RET	RER	RER	RER	SUM										
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	1	2	3	1	2				
f32	1024	61.98	18.26	27.28	10.67	42.54	26.21	17.41	68.60	14.53	9.32	163.71	18.00	129.11	15.91	233.31	223.57	185.62	11.57	2.31	56.65	24.57	9.96	26.52	21.28	9.19	9.94	1.42	30.34	42.98	55.13	236.84	469.87	741.35	294.39	32.99	294.05	313.21	39.10
f32	512	54.38	17.24	26.87	10.95	37.55	15.32	16.68	56.04	10.06	4.88	161.58	17.82	126.80	15.73	231.72	219.14	177.20	9.98	2.26	56.34	24.60	9.81	26.50	21.35	9.14	9.94	1.41	30.64	39.05	54.40	236.42	458.81	730.16	293.33	33.08	287.05	314.06	39.10
f32	256	50.33	16.98	26.18	10.67	25.08	14.30	15.81	50.18	7.71	4.43	160.39	15.10	124.23	12.48	228.31	215.48	172.08	10.01	2.32	56.33	24.49	9.87	26.41	21.29	9.20	9.91	1.41	30.44	39.09	54.32	236.47	458.58	730.34	292.64	33.15	281.41	314.05	39.04
f32	128	38.79	16.90	26.37	10.77	23.86	13.86	16.04	37.63	6.87	2.61	159.46	15.33	123.31	11.77	227.32	212.13	169.66	9.41	2.23	55.77	24.78	9.84	26.24	21.34	9.25	9.89	1.45	30.33	38.56	54.33	235.67	458.56	730.39	293.15	33.27	275.89	313.79	39.06
f32	64	38.04	16.85	26.45	10.84	23.34	13.55	15.75	36.54	6.60	2.22	159.61	14.24	122.98	12.26	227.95	211.36	166.29	8.33	2.23	55.39	24.82	9.85	26.41	21.28	9.18	9.91	1.41	30.62	36.93	54.37	235.64	465.25	741.75	298.34	33.83	283.23	333.10	40.91
f32	32	37.44	17.07	26.19	10.68	22.87	13.33	15.28	35.39	6.47	2.33	158.95	14.10	123.30	11.69	227.99	211.15	166.12	8.86	2.22	55.65	24.82	9.92	26.31	21.24	9.22	9.96	1.41	30.35	36.96	53.89	237.25	468.68	742.06	298.14	34.09	284.72	331.88	40.97
f32	16	36.47	16.68	25.54	10.63	22.20	13.33	15.16	33.95	6.35	1.94	158.90	13.97	122.90	11.65	226.98	210.39	166.11	8.58	2.26	55.63	24.93	9.94	26.21	21.16	9.27	9.93	1.46	30.26	37.01	54.20	235.43	469.49	741.94	297.75	33.83	280.40	331.17	40.96
f32	8	36.19	16.58	25.45	10.55	21.97	13.43	15.37	33.54	5.73	1.85	158.89	14.00	122.02	10.65	225.98	211.73	166.00	7.93	2.22	55.43	24.88	9.83	26.32	21.33	9.19	9.97	1.41	30.76	36.92	53.76	236.34	468.88	741.17	298.19	33.89	277.67	331.95	40.83
int	1024	64.05	17.85	27.01	10.85	45.47	26.17	16.74	70.43	12.55	8.88	164.43	19.88	130.29	16.84	232.87	223.47	185.13	13.62	2.29	57.22	24.80	9.92	26.72	21.31	9.18	9.97	1.43	30.67	46.75	51.14	238.55	413.14	653.41	264.14	29.59	274.76	289.77	35.82
int	512	55.05	17.54	26.62	10.77	37.80	14.85	16.72	57.25	10.64	5.07	161.29	17.72	127.16	15.50	231.59	218.32	175.31	10.30	2.28	56.58	24.60	9.90	26.53	21.24	9.21	9.93	1.41	30.48	39.80	54.40	236.64	452.53	653.00	264.36	29.73	267.62	289.14	35.77
int	256	50.65	16.86	26.24	11.04	25.21	14.40	16.07	50.75	8.91	4.31	159.81	15.23	123.87	12.17	227.02	212.87	168.82	9.20	2.31	56.20	24.73	9.91	26.43	21.23	9.20	9.92	1.46	30.43	39.39	54.37	236.57	412.79	653.14	264.33	29.59	261.74	289.11	35.71
int	128	38.96	17.24	26.51	10.80	24.15	13.57	15.60	38.22	7.36	3.32	159.80	14.62	123.46	11.93	228.84	212.04	166.97	8.64	2.23	55.51	24.84	9.90	26.21	21.25	9.25	9.96	1.43	30.32	38.60	53.97	236.37	412.67	653.30	264.55	29.70	260.12	288.84	35.68
int	64	37.73	16.86	26.14	10.96	23.32	14.03	16.00	36.03	6.55	2.39	159.27	14.18	122.94	11.28	226.83	211.26	165.12	8.37	2.25	55.59	24.77	9.87	26.22	21.29	9.25	9.93	1.43	30.34	37.31	54.10	235.77	413.02	652.97	264.21	29.64	260.84	289.24	35.62
int	32	37.78	17.01	26.64	11.02	23.01	13.47	15.72	36.03	6.51	1.92	159.27	14.10	122.63	11.16	227.40	210.20	165.07	8.78	2.24	55.33	24.82	9.90	26.20	21.20	9.28	9.97	1.41	30.31	37.78	54.20	235.66	412.67	653.31	264.72	29.61	261.13	289.14	35.81
int	16	37.03	17.03	25.83	10.76	22.97	13.44	15.61	35.13	6.28	2.17	158.77	14.15	122.51	11.01	226.20	211.01	164.93	8.08	2.28	55.31	24.80	9.83	26.18	21.25	9.21	9.91	1.41	30.21	36.72	53.93	236.52	412.53	653.00	264.36	29.56	256.26	288.88	35.62
int	8	36.87	16.54	25.53	10.68	22.73	13.90	16.12	35.01	6.97	1.67	158.65	13.80	122.57	10.79	227.20	209.99	165.20	8.49	2.23	55.68	24.86	9.83	26.33	21.29	9.20	9.94	1.41	30.63	37.60	54.12	235.59	412.78	653.69	264.65	29.59	255.88	288.63	35.72
bin	1024	38.57	17.11	26.07	10.70	24.04	13.82	15.47	37.95	8.78	2.92	159.61	14.70	123.54	12.20	228.88	212.11	166.53	8.80	2.23	55.75	24.82	9.91	26.35	21.14	9.19	9.94	1.42	30.47	38.91	54.33	236.16	412.63	653.17	264.23	29.61	258.72	289.21	35.66
bin	512	37.86	16.88	26.57	10.89	23.29	13.47	15.69	36.45	6.44	2.09	157.91	13.62	123.19	11.34	228.08	211.15	166.16	9.17	2.22	55.46	24.83	9.90	26.40	21.21	9.20	9.96	1.41	30.41	38.23	54.07	236.07	412.73	653.40	264.47	29.60	258.85	288.85	35.64
bin	256	37.18	17.02	26.41	10.74	22.81	13.34	15.40	35.71	7.07	2.09	159.19	14.05	122.83	11.18	227.53	213.91	164.80	8.19	2.22	55.38	24.77	9.86	26.19	21.22	9.24	9.94	1.41	30.31	37.93	54.27	235.63	412.62	653.38	264.56	29.52	258.77	288.83	35.71
bin	128	37.28	16.52	25.95	10.65	22.67	13.32	15.49	35.42	7.55	1.95	158.97	14.01	122.96	11.03	226.70	212.02	164.93	8.13	2.22	55.43	24.79	9.90	26.20	21.29	9.26	9.95	1.41	30.35	37.33	54.21	235.64	412.70	653.61	264.47	29.56	259.46	289.31	35.81
bin	64	36.95	16.61	25.71	10.51	22.62	13.35	15.36	34.76	5.93	1.68	158.74	13.86	122.79	10.91	227.30	210.44	166.02	8.02	2.26	55.33	24.79	9.88	26.23	21.24	9.27	9.91	1.42	30.21	37.71	54.13	235.65	412.59	653.39	264.24	29.65	254.81	289.31	35.77
bin	32	37.14	16.63	25.77	10.56	22.69	12.87	15.04	34.77	5.59	1.77	158.63	13.52	122.54	10.54	225.97	215.97	164.45	7.87	2.25	55.53	24.90	9.87	26.22	21.22	9.24	9.93	1.43	30.31	36.65	53.84	236.22	412.37	652.75	264.47	29.59	253.85	288.50	35.79
bin	16	36.97	16.53	25.56	10.48	22.63	12.81	15.00	34.94	5.91	1.22	158.25	13.49	121.98	10.72	226.34	210.00	166.40	7.78	2.26	55.31	24.90	9.92	26.25	21.29	9.22	9.90	1.46	30.43	36.70	54.17	234.98	412.63	653.82	264.45	29.50	251.74	288.74	35.82
bin	8	36.55	16.22	25.41	10.59	21.99	13.02	15.39	34.65	5.42	1.14	158.86	13.76	122.02	10.35	225.76	211.10	164.72	7.79	2.22	55.30	24.85	9.84	26.28	21.24	9.18	9.96	1.42	30.54	36.30	53.78	235.65	412.13	653.18	264.22	29.55	251.86	288.65	35.03

Table 7: Absolute evaluation time for mxbai-embed-large-v1 on all tasks. Table 2 serves as a legend.

q	m	CLA	CLU	STS	PCL	PCL	PCL	RET	RET	RER	RER	RER	SUM		
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8

q	m	Total Time	Average All Tasks	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ
f32	1024	3989.64	104.99	34.12	100.46	19.21	111.65	501.87	213.42	39.10
f32	512	3887.41	102.30	29.38	97.49	19.20	109.95	494.10	211.41	39.10
f32	256	3830.50	100.80	26.19	95.02	19.17	109.96	493.85	209.53	39.04
f32	128	3785.90	99.63	23.03	93.79	19.11	109.52	494.03	207.65	39.06
f32	64	3827.66	100.73	22.67	93.18	19.11	108.98	501.78	216.72	40.91
f32	32	3828.93	100.76	22.28	93.10	19.11	109.37	502.96	216.89	40.97
f32	16	3814.97	100.39	21.75	92.78	19.11	108.88	503.06	215.13	40.96
f32	8	3808.78	100.23	21.63	92.48	19.13	109.01	502.75	214.50	40.83
int	1024	3781.10	99.50	34.82	100.87	19.35	113.24	443.57	198.04	35.82
int	512	3683.75	96.94	29.58	97.29	19.22	110.28	442.99	195.50	35.77
int	256	3623.37	95.35	26.43	94.33	19.18	110.11	443.42	193.48	35.71
int	128	3586.75	94.39	23.13	93.70	19.09	109.65	443.51	192.89	35.68
int	64	3572.91	94.02	22.63	92.82	19.09	109.06	443.40	193.24	35.62
int	32	3573.05	94.03	22.58	92.75	19.07	109.23	443.63	193.29	35.81
int	16	3560.70	93.70	22.23	92.51	19.04	109.06	443.30	191.57	35.62
int	8	3562.39	93.75	22.17	92.53	19.14	109.11	443.71	191.37	35.72
bin	1024	3585.65	94.36	22.97	93.81	19.12	109.80	443.34	192.51	35.66
bin	512	3576.19	94.11	22.64	93.20	19.10	109.52	443.53	192.43	35.64
bin	256	3568.42	93.91	22.33	92.81	19.05	109.28	443.52	192.37	35.71
bin	128	3568.44	93.91	22.16	92.82	19.08	109.06	443.59	192.78	35.81
bin	64	3559.44	93.67	21.98	92.57	19.06	109.16	443.41	191.26	35.77
bin	32	3552.10	93.48	21.93	92.18	19.08	108.91	443.20	190.65	35.79
bin	16	3550.49	93.43	21.86	92.21	19.09	108.62	443.63	189.99	35.82
bin	8	3545.92	93.31	21.73	92.09	19.08	108.58	443.18	190.02	35.03

Table 9: Evaluation time for mxbai-embed-large-v1, averaged per category. Table 2 serves as a legend.

q	m	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ	Average all
f32	1024	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
f32	512	0.884	0.895	0.996	0.965	0.986	0.994	1.000	0.943
f32	256	0.809	0.831	0.998	0.964	0.985	0.988	0.999	0.910
f32	128	0.761	0.792	0.997	0.959	0.986	0.983	0.999	0.888
f32	64	0.753	0.772	0.994	0.947	1.001	1.017	1.046	0.885
f32	32	0.742	0.772	0.994	0.946	1.004	1.020	1.048	0.883
f32	16	0.729	0.763	1.000	0.946	1.004	1.012	1.048	0.879
f32	8	0.727	0.745	0.995	0.944	1.004	1.010	1.044	0.872
int	1024	1.009	1.017	1.004	1.026	0.886	0.919	0.916	0.992
int	512	0.885	0.901	0.998	0.971	0.885	0.911	0.915	0.929
int	256	0.819	0.826	1.002	0.967	0.886	0.903	0.913	0.895
int	128	0.762	0.793	0.996	0.958	0.886	0.902	0.913	0.872
int	64	0.756	0.766	0.996	0.948	0.886	0.903	0.911	0.863
int	32	0.754	0.763	0.994	0.953	0.886	0.903	0.916	0.862
int	16	0.743	0.756	0.994	0.944	0.886	0.897	0.911	0.856
int	8	0.742	0.756	0.995	0.950	0.886	0.896	0.914	0.857
bin	1024	0.756	0.802	0.995	0.963	0.886	0.900	0.912	0.873
bin	512	0.753	0.773	0.994	0.957	0.886	0.900	0.912	0.864
bin	256	0.745	0.763	0.992	0.954	0.886	0.899	0.913	0.859
bin	128	0.738	0.764	0.994	0.949	0.886	0.901	0.916	0.858
bin	64	0.733	0.747	0.995	0.951	0.886	0.896	0.915	0.853
bin	32	0.730	0.740	0.996	0.942	0.885	0.894	0.915	0.850
bin	16	0.727	0.735	0.999	0.943	0.886	0.891	0.916	0.848
bin	8	0.726	0.730	0.994	0.938	0.885	0.891	0.896	0.844

Table 10: Relative evaluation time for mxbai-embed-large-v1, averaged per category. Table 2 serves as a legend.

q	m	req.	CLA bits	CLA 1	CLA 2	CLA 3	CLA 4	CLA 5	CLA 6	CLA 7	CLA 8	CLU 1	CLU 2	CLU 3	CLU 4	CLU 5	CLU 6	CLU 7	CLU 8	CLU 9	CLU 10	STS 1	STS 2	STS 3	STS 4	STS 5	STS 6	STS 7	STS 8	STS 9	STS 10	PCL 1	PCL 2	PCL 3	RET 1	RET 2	RET 3	RER 1	RER 2	RER 3	RER SUM
f32	1024	32768	0.878	0.509	0.597	0.750	0.765	0.799	0.940	0.768	0.601	0.603	0.419	0.400	0.367	0.352	0.561	0.553	0.406	0.508	0.861	0.828	0.791	0.898	0.852	0.893	0.868	0.893	0.892	0.690	0.968	0.785	0.862	0.666	0.403	0.377	0.652	0.325	0.552	0.327	
f32	512	16384	0.875	0.500	0.588	0.725	0.759	0.792	0.932	0.758	0.609	0.591	0.424	0.395	0.372	0.352	0.539	0.545	0.404	0.524	0.856	0.826	0.788	0.899	0.851	0.892	0.865	0.891	0.885	0.695	0.968	0.784	0.860	0.661	0.396	0.370	0.636	0.324	0.549	0.316	
f32	256	8192	0.872	0.483	0.578	0.708	0.743	0.779	0.923	0.741	0.599	0.587	0.410	0.380	0.363	0.343	0.512	0.526	0.405	0.480	0.853	0.823	0.789	0.894	0.848	0.886	0.859	0.889	0.883	0.698	0.965	0.777	0.856	0.632	0.363	0.347	0.640	0.314	0.545	0.308	
f32	128	4096	0.862	0.460	0.565	0.672	0.713	0.745	0.905	0.700	0.583	0.582	0.395	0.365	0.353	0.327	0.481	0.493	0.401	0.451	0.852	0.815	0.777	0.885	0.837	0.879	0.844	0.880	0.868	0.697	0.960	0.767	0.851	0.589	0.321	0.300	0.638	0.303	0.536	0.301	
f32	64	2048	0.834	0.411	0.520	0.661	0.664	0.681	0.861	0.631	0.579	0.569	0.358	0.320	0.340	0.315	0.367	0.423	0.393	0.373	0.815	0.805	0.765	0.865	0.822	0.865	0.829	0.863	0.824	0.686	0.944	0.746	0.836	0.491	0.262	0.233	0.616	0.281	0.516	0.302	
f32	32	1024	0.774	0.374	0.498	0.628	0.576	0.591	0.780	0.502	0.571	0.553	0.312	0.279	0.292	0.298	0.268	0.347	0.391	0.302	0.779	0.787	0.744	0.838	0.792	0.835	0.797	0.793	0.658	0.915	0.708	0.812	0.326	0.180	0.146	0.598	0.265	0.486	0.285		
f32	16	512	0.632	0.332	0.479	0.593	0.431	0.455	0.628	0.344	0.526	0.528	0.253	0.215	0.290	0.270	0.184	0.271	0.376	0.219	0.768	0.751	0.687	0.776	0.728	0.798	0.747	0.787	0.784	0.662	0.823	0.638	0.767	0.151	0.071	0.072	0.557	0.267	0.421	0.278	
f32	8	256	0.396	0.313	0.453	0.588	0.289	0.346	0.480	0.211	0.509	0.511	0.190	0.162	0.245	0.228	0.135	0.224	0.370	0.157	0.661	0.663	0.614	0.672	0.628	0.700	0.689	0.705	0.684	0.647	0.538	0.570	0.702	0.043	0.013	0.032	0.517	0.252	0.356	0.265	
int	1024	8192	0.870	0.492	0.579	0.648	0.734	0.774	0.935	0.713	0.608	0.593	0.424	0.394	0.370	0.350	0.558	0.549	0.404	0.526	0.856	0.829	0.766	0.897	0.843	0.890	0.861	0.889	0.891	0.691	0.970	0.783	0.863	0.661	0.406	0.373	0.640	0.319	0.550	0.321	
int	512	4096	0.867	0.481	0.569	0.631	0.726	0.759	0.926	0.687	0.597	0.585	0.422	0.389	0.369	0.351	0.543	0.540	0.404	0.518	0.852	0.827	0.763	0.895	0.841	0.887	0.858	0.887	0.888	0.688	0.969	0.781	0.862	0.657	0.395	0.361	0.633	0.317	0.546	0.313	
int	256	2048	0.859	0.476	0.556	0.631	0.706	0.741	0.911	0.655	0.597	0.582	0.408	0.382	0.362	0.341	0.521	0.523	0.402	0.478	0.860	0.823	0.758	0.891	0.839	0.882	0.852	0.884	0.881	0.693	0.967	0.774	0.858	0.629	0.371	0.338	0.634	0.312	0.542	0.301	
int	128	1024	0.841	0.441	0.533	0.569	0.668	0.668	0.708	0.887	0.612	0.575	0.572	0.395	0.367	0.352	0.330	0.480	0.490	0.403	0.446	0.851	0.816	0.746	0.882	0.832	0.875	0.843	0.875	0.874	0.697	0.963	0.764	0.852	0.572	0.323	0.301	0.632	0.308	0.533	0.292
int	64	512	0.793	0.378	0.479	0.530	0.604	0.624	0.814	0.509	0.583	0.563	0.342	0.321	0.336	0.315	0.366	0.462	0.491	0.370	0.369	0.807	0.807	0.728	0.865	0.811	0.860	0.824	0.863	0.834	0.679	0.946	0.746	0.838	0.471	0.265	0.232	0.612	0.291	0.506	0.295
int	32	256	0.683	0.330	0.451	0.512	0.472	0.506	0.685	0.340	0.574	0.558	0.302	0.272	0.316	0.301	0.259	0.345	0.392	0.293	0.763	0.792	0.696	0.833	0.781	0.834	0.789	0.823	0.646	0.919	0.708	0.812	0.310	0.168	0.144	0.583	0.277	0.474	0.286		
int	16	128	0.459	0.280	0.470	0.550	0.270	0.309	0.478	0.187	0.520	0.520	0.242	0.212	0.289	0.265	0.179	0.270	0.375	0.207	0.697	0.755	0.646	0.782	0.713	0.779	0.732	0.773	0.764	0.638	0.830	0.627	0.767	0.153	0.057	0.066	0.536	0.266	0.413	0.263	
int	8	64	0.329	0.271	0.438	0.572	0.214	0.271	0.364	0.167	0.510	0.518	0.179	0.157	0.242	0.229	0.134	0.225	0.370	0.153	0.638	0.685	0.573	0.649	0.592	0.669	0.610	0.640	0.630	0.573	0.569	0.701	0.041	0.012	0.029	0.490	0.252	0.346	0.255		
bin	1024	1024	0.764	0.365	0.482	0.599	0.601	0.622	0.800	0.524	0.551	0.540	0.318	0.283	0.315	0.295	0.367	0.404	0.358	0.350	0.714	0.788	0.735	0.839	0.795	0.830	0.849	0.866	0.622	0.936	0.728	0.833	0.395	0.182	0.156	0.592	0.283	0.469	0.311		
bin	512	512	0.634	0.317	0.443	0.559	0.472	0.500	0.648	0.358	0.351	0.387	0.243	0.210	0.286	0.267	0.240	0.314	0.353	0.236	0.657	0.760	0.708	0.792	0.758	0.777	0.802	0.797	0.635	0.889	0.684	0.809	0.232	0.104	0.081	0.557	0.268	0.420	0.293		
bin	256	256	0.446	0.263	0.411	0.532	0.310	0.365	0.495	0.220	0.494	0.507	0.189	0.170	0.260	0.246	0.151	0.252	0.358	0.184	0.569	0.706	0.654	0.714	0.681	0.710	0.704	0.739	0.751	0.589	0.796	0.608	0.760	0.101	0.039	0.048	0.541	0.261	0.358	0.282	
bin	128	128	0.271	0.265	0.427	0.585	0.190	0.224	0.340	0.136	0.492	0.495	0.153	0.140	0.233	0.219	0.120	0.221	0.363	0.146	0.370	0.621	0.545	0.600	0.567	0.592	0.607	0.626	0.608	0.532	0.655	0.521	0.697	0.040	0.015	0.026	0.507	0.261	0.293	0.288	
bin	64	64	0.196	0.251	0.406	0.560	0.126	0.186	0.274	0.070	0.491	0.484	0.142	0.123	0.227	0.213	0.107	0.204	0.360	0.127	0.167	0.526	0.435	0.462	0.436	0.473	0.477	0.493	0.485	0.433	0.611	0.020	0.014	0.023	0.469	0.247	0.257	0.314			
bin	32	32	0.103	0.194	0.385	0.584	0.072	0.147	0.174	0.025	0.484	0.488	0.136	0.118	0.222	0.210	0.101	0.203	0.363	0.120	0.049	0.439	0.386	0.308	0.334	0.371	0.370	0.390	0.428	0.284	0.193	0.371	0.534	0.011	0.006	0.016	0.446	0.244	0.239	0.281	
bin	16	16	0.047	0.181	0.346	0.591	0.041	0.090	0.092	0.024	0.499	0.502	0.118	0.107	0.217	0.205	0.101	0.200	0.344	0.119	0.015	0.204	0.223	0.122	0.171	0.224	0.173	0.182	0.161	0.223	0.070	0.038	0.024	0.294	0.103	0.052	0.195				
bin	8	8	0.030	0.191	0.328	0.528	0.034	0.091	0.087	0.006	0.498	0.502	0.118	0.107	0.217	0.205	0.101	0.200	0.344	0.119	0.015	0.204	0.223	0.122	0.171	0.224	0.173	0.182	0.161	0.223	0.070	0.038	0.024	0.294	0.103	0.052	0.195				

Table 11: Absolute performance for mxbai-embed-large-v1 on all tasks. Table 2 serves as a legend.

Table 12: Relative performance for mxbai-embed-large-v1 on all tasks. Table 2 serves as a legend.

q	m	req. bits	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ	Average all
f32	1024	32768	0.751	0.477	0.847	0.872	0.482	0.510	0.327	0.662
f32	512	16384	0.741	0.475	0.845	0.870	0.476	0.503	0.316	0.658
f32	256	8192	0.728	0.461	0.842	0.866	0.447	0.500	0.308	0.647
f32	128	4096	0.703	0.443	0.834	0.859	0.404	0.492	0.301	0.630
f32	64	2048	0.658	0.404	0.814	0.842	0.329	0.471	0.302	0.596
f32	32	1024	0.590	0.364	0.785	0.811	0.217	0.450	0.285	0.551
f32	16	512	0.487	0.313	0.749	0.743	0.098	0.415	0.278	0.488
f32	8	256	0.384	0.273	0.666	0.604	0.029	0.375	0.265	0.415
int	1024	8192	0.718	0.478	0.841	0.872	0.480	0.503	0.321	0.653
int	512	4096	0.706	0.472	0.839	0.871	0.471	0.498	0.313	0.647
int	256	2048	0.692	0.460	0.836	0.866	0.446	0.496	0.301	0.637
int	128	1024	0.657	0.441	0.829	0.860	0.399	0.491	0.292	0.618
int	64	512	0.591	0.401	0.808	0.843	0.323	0.469	0.295	0.579
int	32	256	0.497	0.361	0.775	0.813	0.207	0.445	0.286	0.527
int	16	128	0.375	0.308	0.728	0.741	0.092	0.405	0.263	0.456
int	8	64	0.328	0.272	0.640	0.614	0.027	0.363	0.255	0.395
bin	1024	1024	0.595	0.378	0.788	0.832	0.245	0.448	0.311	0.561
bin	512	512	0.491	0.315	0.748	0.794	0.139	0.415	0.293	0.497
bin	256	256	0.380	0.281	0.682	0.722	0.062	0.387	0.282	0.433
bin	128	128	0.305	0.258	0.567	0.594	0.027	0.354	0.288	0.366
bin	64	64	0.259	0.248	0.439	0.452	0.019	0.324	0.314	0.306
bin	32	32	0.210	0.245	0.326	0.366	0.011	0.310	0.281	0.256
bin	16	16	0.179	0.244	0.284	0.315	0.011	0.285	0.262	0.232
bin	8	8	0.162	0.241	0.167	0.275	0.001	0.149	0.195	0.180

Table 13: Performance for mxbai-embed-large-v1, averaged per category. Table 2 serves as a legend.

q	m	req. bits	Average Relative classification	Average Relative clustering	Average Relative sts	Average Relative pairclass	Average Relative retrieval	Average Relative rerank	Average Relative summ	Average Relative all
f32	1024	32768	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
f32	512	16384	0.987	0.998	0.998	0.986	0.989	0.968	0.993	
f32	256	8192	0.968	0.967	0.995	0.993	0.923	0.978	0.943	0.974
f32	128	4096	0.933	0.931	0.985	0.985	0.827	0.960	0.919	0.944
f32	64	2048	0.871	0.850	0.962	0.965	0.669	0.914	0.924	0.886
f32	32	1024	0.783	0.770	0.928	0.929	0.442	0.871	0.871	0.812
f32	16	512	0.652	0.662	0.886	0.851	0.199	0.812	0.850	0.714
f32	8	256	0.525	0.574	0.790	0.699	0.060	0.738	0.809	0.609
int	1024	8192	0.955	1.002	0.994	1.000	0.996	0.986	0.982	0.988
int	512	4096	0.938	0.991	0.991	0.998	0.975	0.978	0.958	0.977
int	256	2048	0.920	0.965	0.988	0.993	0.921	0.972	0.921	0.960
int	128	1024	0.872	0.928	0.980	0.985	0.820	0.960	0.893	0.927
int	64	512	0.782	0.844	0.954	0.966	0.660	0.916	0.904	0.862
int	32	256	0.661	0.763	0.916	0.931	0.422	0.869	0.875	0.779
int	16	128	0.510	0.652	0.861	0.848	0.182	0.796	0.805	0.671
int	8	64	0.453	0.570	0.759	0.710	0.056	0.718	0.780	0.583
bin	1024	1024	0.786	0.793	0.931	0.953	0.487	0.876	0.952	0.827
bin	512	512	0.654	0.664	0.884	0.909	0.274	0.814	0.897	0.726
bin	256	256	0.512	0.592	0.806	0.826	0.125	0.760	0.863	0.634
bin	128	128	0.424	0.540	0.672	0.685	0.056	0.704	0.882	0.545
bin	64	64	0.365	0.519	0.521	0.527	0.042	0.648	0.960	0.472
bin	32	32	0.302	0.512	0.387	0.430	0.025	0.623	0.861	0.408
bin	16	16	0.261	0.509	0.336	0.373	0.025	0.579	0.800	0.376
bin	8	8	0.240	0.501	0.200	0.328	0.002	0.287	0.597	0.299

Table 14: Relative performance for mxbai-embed-large-v1, averaged per category. Table 2 serves as a legend.

D. stella_en_400M_v5

This section contains additional figures and tables for the STELLA model.

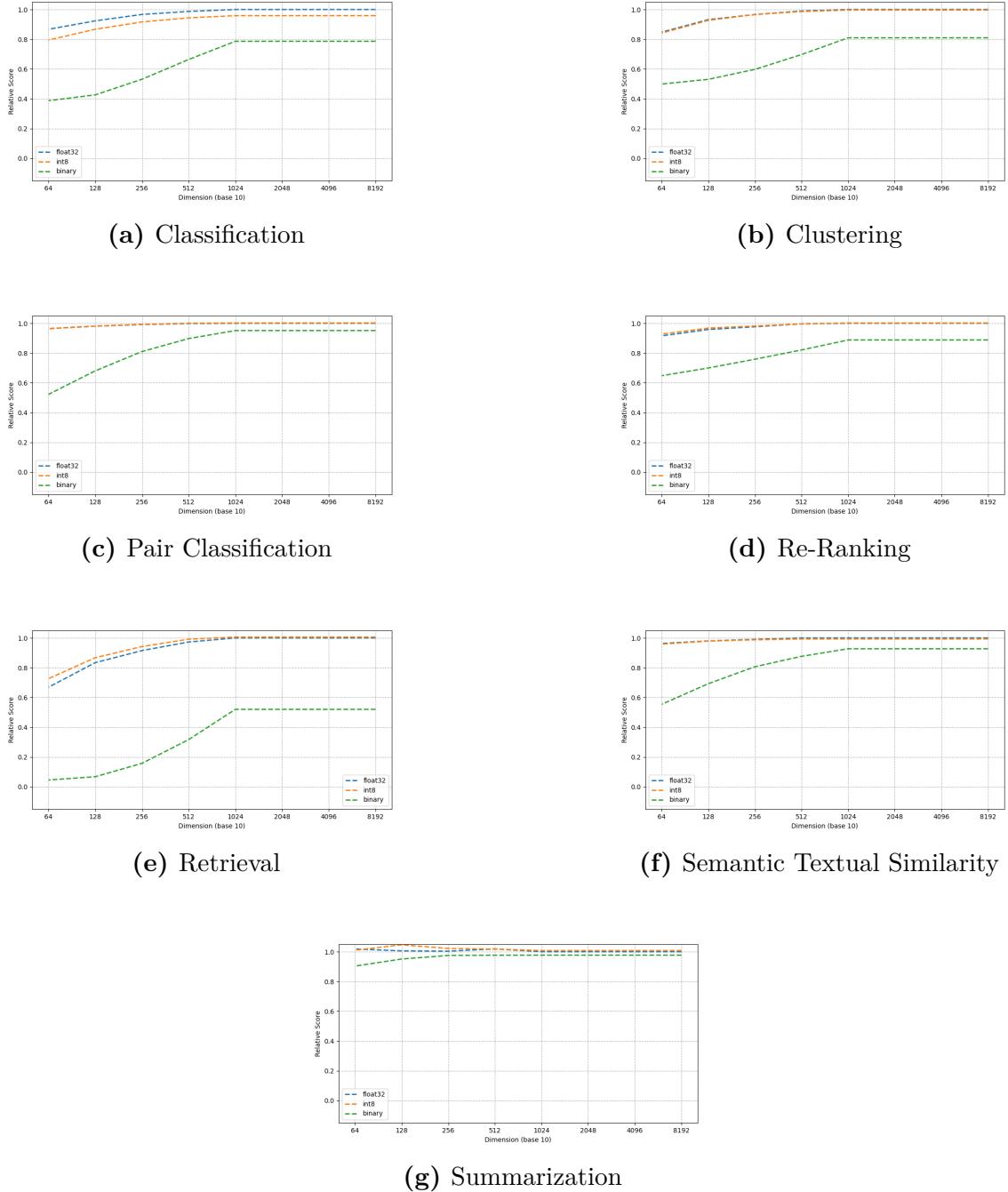


Figure 10: Relative performance of the STELLA model for each task category in the MTEB. The x-axis of each subfigure represents the embedding size while the y-axis represents the relative performance. For each task the reference point is the performance of the float32 embeddings at the full embedding size. Blue: float32, Orange: int8, Green: binary.

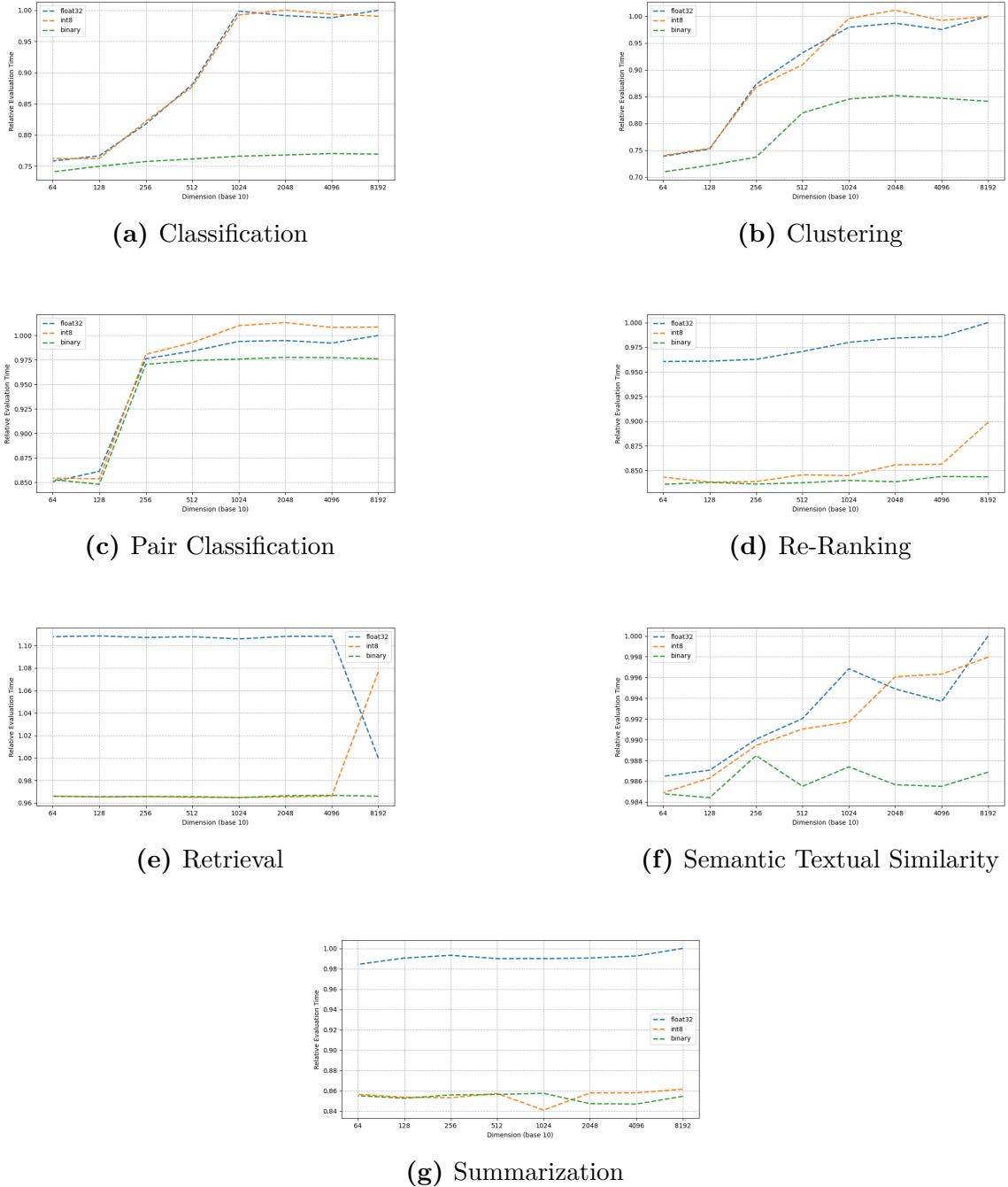


Figure 11: Relative evaluation time of the STELLA model for each task category in the MTEB. The x-axis of each subfigure represents the embedding size while the y-axis represents the relative evaluation time. For each task the reference point is the evaluation time of the float32 embeddings at the full embedding size. Blue: float32, Orange: int8, Green: binary.

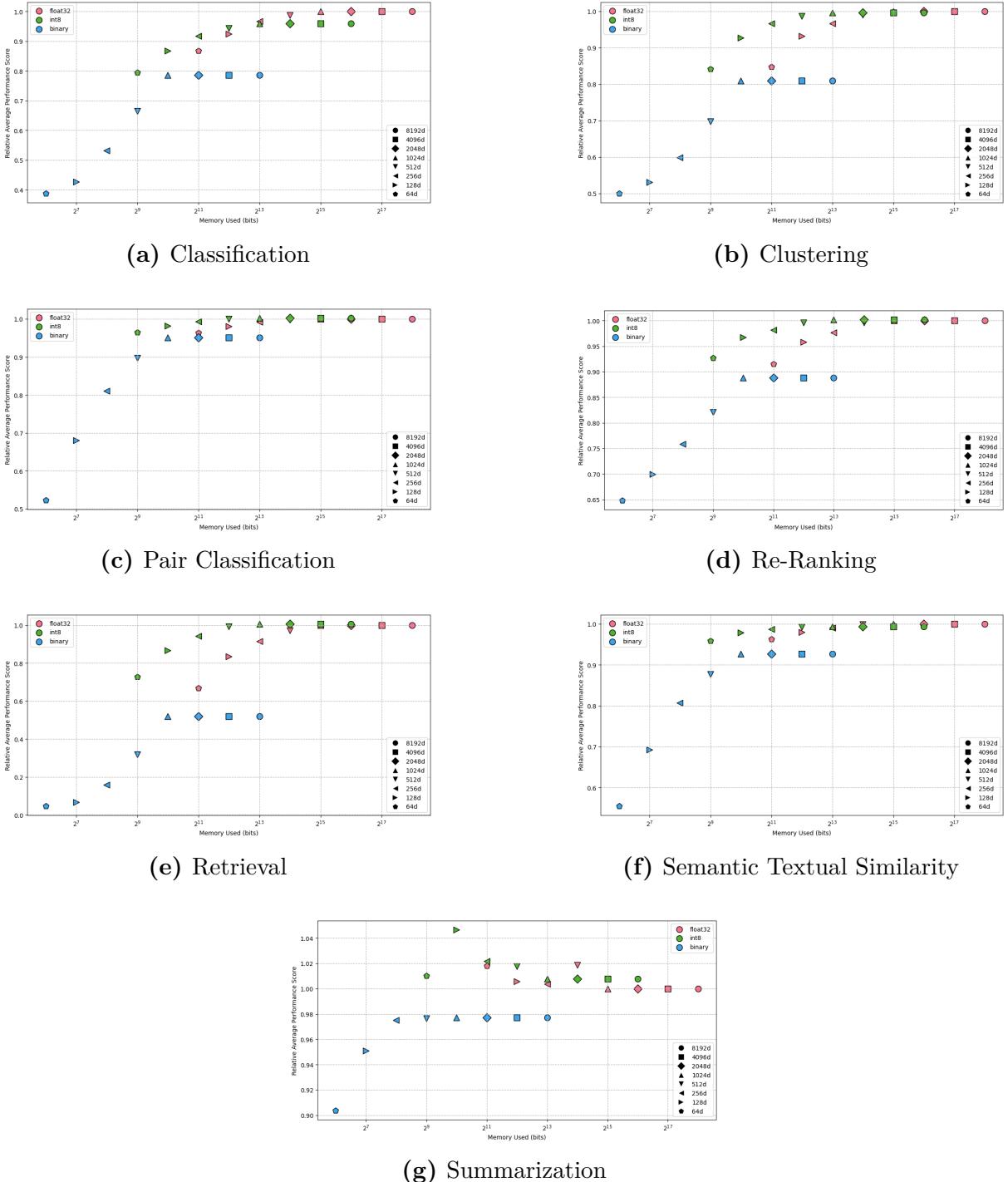


Figure 12: Relative accuracy compute trade-off of the STELLA model, averaged over all selected MTEB tasks within each category. The x-axis represents memory consumption (log scale) and the y-axis represents relative performance (averaged over tasks within the category). Reference point: float32 embeddings at full size.

q	m	CLA 1	CLA 2	CLA 3	CLA 4	CLA 5	CLA 6	CLA 7	CLA 8	CLU 1	CLU 2	CLU 3	CLU 4	CLU 5	CLU 6	CLU 7	CLU 8	CLU 9	CLU 10	STS 1	STS 2	STS 3	STS 4	STS 5	STS 6	STS 7	STS 8	STS 9	STS 10	PCL 1	PCL 2	PCL 3	RET 1	RET 2	RER 1	RER 2	RER 3	RER 4	SUM 1		
f32	8192	68.87	19.61	26.83	10.00	45.49	26.83	19.20	73.21	13.70	9.45	203.77	19.37	163.82	16.57	178.64	171.79	219.06	11.83	2.13	70.07	27.63	11.22	29.00	24.27	10.62	11.36	1.79	38.50	49.42	72.46	296.81	529.52	731.93	312.82	35.43	339.23	350.92	52.36		
f32	4096	66.67	19.92	26.79	9.69	43.97	26.94	18.93	72.99	12.40	7.54	203.48	19.04	166.01	16.48	188.83	175.55	217.73	11.67	2.17	70.09	27.56	11.27	29.12	24.01	10.51	11.23	1.71	38.07	48.92	71.51	296.74	586.85	816.29	344.50	34.79	331.49	350.35	51.97		
f32	2048	66.89	19.83	26.77	9.72	44.14	26.98	19.45	72.56	11.78	7.55	203.81	19.47	167.52	17.52	189.60	176.49	221.25	12.14	2.19	70.13	27.57	11.25	29.09	23.94	10.49	11.19	1.73	38.06	49.41	71.57	295.93	587.52	816.11	344.02	34.84	329.05	350.70	51.87		
f32	1024	68.45	19.89	26.79	9.76	44.95	27.02	19.12	74.36	11.91	7.60	203.35	19.43	167.14	16.88	189.22	175.62	219.14	11.83	2.21	70.14	27.59	11.26	28.97	23.93	10.54	11.29	1.72	38.35	49.38	71.41	295.88	586.12	814.61	343.47	34.84	324.32	351.03	51.84		
f32	512	60.00	19.01	26.68	9.85	38.83	16.10	18.34	60.24	10.46	6.57	203.08	18.28	165.17	15.47	187.93	173.07	213.95	11.13	2.17	70.03	27.43	11.24	28.88	23.78	10.44	11.30	1.72	38.22	48.41	71.13	294.55	587.22	816.29	343.89	34.94	313.19	351.92	51.84		
f32	256	55.69	18.79	26.42	9.82	26.74	15.12	17.54	54.37	8.74	7.36	201.91	17.88	163.87	14.98	185.89	170.20	210.63	10.48	2.15	69.65	27.44	11.18	29.01	23.97	10.49	11.27	1.69	38.28	47.05	71.17	295.17	587.46	815.04	343.64	34.86	306.45	351.22	50.21		
f32	128	44.21	18.53	26.42	9.76	25.42	14.73	17.45	41.76	6.67	2.66	183.28	14.80	149.70	12.37	163.87	149.24	189.31	9.24	2.11	69.20	27.56	11.15	28.84	23.94	10.55	11.30	1.69	38.10	42.30	61.98	258.96	587.22	816.22	344.42	34.95	303.29	351.63	51.86		
f32	64	43.37	18.27	26.11	9.96	24.67	14.68	17.21	40.37	7.08	2.16	182.62	14.61	149.90	12.73	161.90	147.39	187.57	8.13	2.07	69.03	27.59	11.08	28.97	24.13	10.53	11.27	1.70	38.21	40.74	62.04	258.66	587.93	816.07	343.61	34.96	304.09	350.32	51.54		
int	8192	70.14	19.04	26.74	9.94	46.23	26.53	17.42	75.61	13.40	8.94	204.93	19.17	165.54	16.47	187.74	173.88	214.92	12.32	2.18	71.14	27.57	11.32	29.17	24.00	10.51	11.26	1.73	38.14	50.77	72.22	297.34	566.43	794.73	336.03	34.69	290.25	302.42	45.12		
int	4096	68.68	19.26	26.80	9.81	46.28	26.84	17.94	75.57	13.29	8.49	203.59	19.11	165.99	16.21	188.49	173.50	214.11	12.22	2.20	70.88	27.55	11.29	29.29	24.05	10.49	11.26	1.69	38.38	50.59	72.32	297.5	511.30	710.30	345.30	300.87	30.28	289.15	302.42	44.93	
int	2048	71.36	19.59	26.90	9.70	46.41	26.83	17.97	76.04	13.61	9.02	204.04	19.94	166.15	17.24	188.31	173.71	216.49	12.15	2.18	70.69	27.57	11.29	29.05	24.03	10.53	11.28	1.71	38.23	51.43	72.38	296.75	510.33	710.13	301.00	30.13	289.56	302.80	44.92		
int	1024	69.98	19.43	26.40	9.70	45.75	26.85	18.13	75.31	13.28	8.63	203.18	19.42	166.12	16.42	188.59	173.95	215.05	12.03	2.16	70.38	27.50	11.19	29.00	24.00	10.53	11.24	1.69	38.18	51.02	72.23	297.11	509.71	710.16	300.89	30.12	286.58	294.46	44.04		
int	512	60.75	18.79	26.31	9.88	39.17	16.07	17.61	60.99	9.13	5.81	202.90	18.37	164.81	15.12	187.49	171.82	210.78	11.10	2.16	70.04	27.41	11.07	28.94	23.99	10.46	11.24	1.73	37.98	49.84	71.90	266.73	511.02	709.68	300.51	30.17	279.57	302.15	44.89		
int	256	56.08	18.49	26.25	10.12	27.36	15.32	17.42	54.61	7.45	4.53	201.90	17.63	163.63	15.43	184.59	168.39	208.35	10.79	2.13	69.72	27.53	11.28	29.08	23.88	10.48	11.21	1.69	38.13	47.78	70.98	295.48	511.58	710.03	300.58	30.21	272.72	301.61	44.67		
int	128	44.23	18.34	25.96	9.81	25.36	14.72	17.23	41.71	7.77	2.75	182.84	15.72	148.83	12.08	162.21	147.50	188.06	8.39	2.09	69.43	27.50	11.10	29.01	24.07	10.55	11.23	1.68	38.17	41.13	56.31	81.25	258.48	511.47	709.45	65.300	47.30	32.12	272.48	302.15	44.70
int	64	43.45	18.27	26.19	10.02	24.85	14.92	17.35	40.37	7.28	2.22	183.14	14.62	148.45	11.62	162.38	148.71	187.64	8.79	2.08	69.08	27.46	11.12	28.79	24.06	10.53	11.18	1.71	38.02	41.09	62.25	258.94	511.55	709.91	301.06	30.17	277.02	302.24	44.85		
bin	8192	45.49	18.35	26.00	9.95	25.19	15.10	17.38	41.96	7.70	2.80	201.38	17.29	163.27	13.93	185.42	168.54	207.13	10.33	2.10	69.35	27.63	11.12	28.83	24.06	10.53	11.16	1.71	38.06	47.11	71.06	295.18	511.19	710.28	300.98	30.36	276.14	301.55	44.75		
bin	4096	44.41	18.38	26.04	10.13	25.26	15.15	17.39	41.84	8.02	2.81	201.88	17.36	163.02	14.16	185.78	169.54	207.45	10.37	2.07	69.31	27.50	11.13	28.90	23.98	10.45	11.16	1.72	38.12	47.03	71.01	290.26	511.67	710.78	301.21	30.22	277.29	302.15	44.35		
bin	2048	44.61	18.35	26.10	9.94	25.24	14.88	17.54	41.85	7.98	3.14	201.45	17.48	163.01	14.43	185.82	168.44	207.54	10.43	2.08	69.51	27.45	11.10	28.90	23.97	10.43	11.23	1.71	38.26	47.51	70.81	295.20	511.99	709.70	301.15	30.16	276.48	297.98	44.37		
bin	1024	44.16	18.27	26.06	9.91	25.34	15.13	17.28	41.88	8.18	2.80	201.62	17.29	163.39	14.07	184.98	168.17	207.37	10.31	2.09	69.34	27.48	11.10	28.90	23.93	10.45	11.20	1.74	38.28	47.13	71.01	295.08	510.44	709.11	30.14	274.27	301.86	44.90			
bin	512	43.21	18.48	26.36	9.87	24.77	14.59	17.57	40.36	6.06	2.10	201.96	17.26	162.69	13.74	184.97	168.38	207.11	10.25	2.09	69.13	27.34	11.25	28.94	23.93	10.45	11.17	1.71	38.18	46.94	71.94	264.22	511.28	710.18	302.00	30.21	271.81	301.24	44.85		
bin	256	42.97	18.39	26.10	9.93	24.50	14.56	17.54	39.50	5.72	1.99	201.02	17.06	149.02	16.11	161.15	146.93	186.74	7.90	2.14	69.00	27.46	11.12	28.80	24.04	10.48	11.22	1.71	38.24	46.57	70.88	294.14	511.52	709.92	300.70	30.11	270.58	302.13	44.82		
bin	128	42.50	17.96	25.56	10.09	24.14	14.45	17.24	39.36	6.66	1.97	182.34	14.49	148.03	11.23	161.86	146.45	185.74	8.20	2.08	69.03	27.42	11.16	28.86	24.02	10.49	11.17	1.69	38.33	40.44	61.80	258.54	511.51	709.85	300.77	30.17	271.70	302.19	44.65		
bin	64	44.16	18.13	25.69	9.43	24.00	14.05	17.42	38.70	5.98	1.76	181.73	14.21	147.65	11.17	160.92	147.39	187.23	7.81	2.09	69.32	27.49	11.12	28.54	23.91	10.56	11.26	1.69	38.13	41.13	61.88	258.88	511.73	709.66	300.87	30.12	270.03	302.34	44.78		

Table 15: Absolute evaluation time for stella_en_400M_v5 on all tasks. Table 2 serves as a legend.

q	m	CLA	CLU	STS	PCL	PCL	PCL	RET	RET	RER	RER	RER	SUM																											
		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	1							
f32	8192	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000							
f32	4096	0.968	1.015	0.998	0.969	0.967	1.004	0.986	0.997	0.904	0.798	0.999	0.983	1.013	0.995	1.057	1.022	0.994	0.987	1.020	1.000	0.997	1.004	1.004	0.989	0.990	0.988	0.955	0.989	0.990	0.987	1.000	1.108	1.115	1.101	0.982	0.977	0.998	0.993	
f32	2048	0.971	1.011	0.998	0.972	0.970	1.006	1.013	0.991	0.860	0.799	1.000	1.005	1.023	1.057	1.061	1.027	1.027	1.001	0.998	1.003	1.003	0.986	0.987	0.985	0.970	0.989	1.000	0.988	0.997	1.109	1.115	1.100	0.983	0.970	0.999	0.991			
f32	1024	0.994	1.014	0.999	0.976	0.988	1.007	0.996	1.016	0.869	0.804	0.998	1.003	1.020	1.018	1.059	1.022	1.000	1.000	1.037	1.001	0.998	1.003	0.999	0.986	0.993	0.994	0.962	0.996	0.999	0.985	0.997	1.107	1.113	1.098	0.984	0.956	1.000	0.990	
f32	512	0.871	0.969	0.995	0.985	0.854	0.600	0.953	0.823	0.763	0.695	0.997	0.944	1.008	0.933	1.052	0.977	0.941	0.107	0.999	0.993	1.001	0.996	0.980	0.983	0.994	0.965	0.993	0.974	0.983	0.996	1.109	1.115	1.099	0.984	0.923	1.003	0.990		
f32	256	0.809	0.958	0.985	0.982	0.588	0.564	0.914	0.743	0.638	0.398	0.991	0.923	1.000	0.904	1.041	0.991	0.962	0.886	1.008	0.994	0.993	0.996	1.001	0.988	0.987	0.992	0.949	0.994	0.952	0.982	0.995	1.109	1.114	1.099	0.984	0.903	1.001	0.993	
f32	128	0.642	0.945	0.985	0.976	0.559	0.549	0.909	0.570	0.487	0.281	0.899	0.764	0.914	0.747	0.917	0.869	0.864	0.781	0.989	0.988	0.997	0.994	0.995	0.986	0.993	0.994	0.946	0.990	0.956	0.855	0.873	1.109	1.116	1.101	0.987	0.894	1.002	0.991	
f32	64	0.630	0.931	0.973	0.996	0.542	0.547	0.896	0.551	0.517	0.228	0.896	0.754	0.915	0.768	0.906	0.858	0.856	0.687	0.972	0.985	0.998	0.988	0.999	0.994	0.992	0.953	0.993	0.824	0.856	0.872	1.110	1.115	1.098	0.987	0.886	0.998	0.984		
int	8192	1.018	0.971	0.997	0.994	1.016	0.989	0.907	1.033	0.978	0.946	1.006	0.990	1.010	0.982	1.051	1.012	0.981	1.041	1.021	1.015	0.998	1.008	1.006	0.989	0.989	0.991	0.971	0.991	1.027	0.997	1.002	1.070	1.086	1.074	0.979	0.856	0.862	0.862	
int	4096	0.997	0.982	0.999	0.982	1.017	1.000	0.937	1.035	0.970	0.898	0.999	0.986	1.013	0.978	1.055	1.010	0.977	1.033	1.033	1.012	0.997	1.006	1.003	0.991	0.987	0.991	0.947	0.997	1.024	0.998	1.003	0.966	0.971	0.962	0.855	0.852	0.862	0.858	
int	2048	1.036	0.999	1.003	0.971	1.020	1.000	0.936	1.036	0.993	0.954	1.001	0.209	1.014	1.040	1.054	1.011	0.988	1.028	1.024	1.009	0.998	1.006	1.002	0.990	0.991	0.992	0.957	0.993	1.041	0.999	1.000	0.964	0.970	0.962	0.850	0.854	0.863	0.858	
int	1024	1.016	0.991	0.984	0.970	1.006	1.001	0.944	0.229	0.960	0.915	0.997	1.003	1.014	0.991	1.056	1.013	0.982	1.017	1.014	1.004	0.995	0.997	1.000	0.989	0.991	0.989	0.946	0.992	1.032	0.997	1.001	0.963	0.970	0.962	0.850	0.845	0.839	0.841	
int	512	0.882	0.958	0.981	0.988	0.861	0.599	0.917	0.833	0.666	0.615	0.996	0.948	1.006	0.912	1.050	1.004	0.962	0.939	0.914	1.000	0.992	0.986	0.998	0.985	0.989	0.972	0.987	0.990	0.998	1.000	0.965	0.970	0.961	0.851	0.824	0.861	0.857		
int	256	0.814	0.943	0.978	1.012	0.602	0.571	0.911	0.746	0.544	0.479	0.991	0.910	0.998	0.870	1.038	0.984	0.951	0.912	1.000	0.995	0.996	1.005	1.003	0.984	0.987	0.987	0.948	0.991	0.967	0.980	0.996	0.966	0.970	0.961	0.853	0.804	0.860	0.853	
int	128	0.642	0.935	0.968	0.981	0.557	0.548	0.897	0.570	0.567	0.291	0.897	0.812	0.908	0.729	0.908	0.859	0.859	0.709	0.982	0.991	0.995	0.989	1.000	0.992	0.993	0.989	0.941	0.992	0.837	0.853	0.871	0.966	0.970	0.961	0.850	0.803	0.861	0.854	
int	64	0.631	0.931	0.976	1.002	0.546	0.556	0.904	0.551	0.531	0.235	0.899	0.755	0.906	0.702	0.909	0.866	0.865	0.743	0.973	0.986	0.994	0.991	0.993	0.991	0.991	0.991	0.984	0.957	0.988	0.831	0.859	0.872	0.966	0.970	0.962	0.851	0.817	0.861	0.857
bin	8192	0.660	0.935	0.969	0.995	0.554	0.563	0.905	0.573	0.562	0.296	0.898	0.892	0.997	0.841	1.038	0.981	0.946	0.872	0.983	0.990	1.000	0.991	0.994	0.991	0.982	0.958	0.989	0.953	0.981	0.995	0.965	0.970	0.962	0.857	0.814	0.859	0.855		
bin	4096	0.645	0.937	0.971	1.013	0.555	0.565	0.906	0.572	0.572	0.297	0.991	0.896	0.995	0.855	1.040	0.987	0.947	0.877	0.972	0.989	0.995	0.992	0.997	0.988	0.983	0.982	0.966	0.990	0.952	0.983	0.998	0.966	0.971	0.963	0.851	0.817	0.861	0.847	
bin	2048	0.648	0.935	0.973	0.994	0.555	0.554	0.913	0.572	0.582	0.332	0.989	0.902	0.995	0.871	1.040	0.981	0.947	0.881	0.977	0.992	0.993	0.989	0.997	0.987	0.982	0.988	0.957	0.994	0.961	0.977	0.995	0.967	0.970	0.963	0.851	0.815	0.849	0.847	
bin	1024	0.641	0.932	0.972	0.991	0.557	0.564	0.900	0.572	0.597	0.297	0.899	0.892	0.997	0.849	1.036	0.979	0.947	0.871	0.978	0.990	0.995	0.989	0.997	0.986	0.984	0.986	0.975	0.994	0.954	0.980	0.994	0.964	0.969	0.961	0.851	0.808	0.860	0.858	
bin	512	0.627	0.942	0.982	0.987	0.544	0.544	0.915	0.551	0.442	0.222	0.991	0.891	0.993	0.829	1.036	0.980	0.946	0.867	0.979	0.987	0.997	0.990	0.998	0.986	0.984	0.983	0.960	0.992	0.950	0.982	0.991	0.965	0.970	0.961	0.853	0.801	0.858	0.857	
bin	256	0.624	0.938	0.973	0.993	0.539	0.542	0.913	0.540	0.418	0.210	0.987	0.881	0.910	0.688	0.902	0.855	0.853	0.668	1.003	0.985	0.994	0.991	0.997	0.990	0.986	0.988	0.958	0.994	0.942	0.978	0.991	0.966	0.970	0.961	0.850	0.798	0.861	0.856	
bin	128	0.617	0.915	0.953	1.010	0.531	0.538	0.898	0.538	0.480	0.209	0.895	0.748	0.904	0.675	0.906	0.853	0.848	0.693	0.976	0.985	0.992	0.994	0.995	0.989	0.987	0.983	0.946	0.996	0.818	0.855	0.871	0.965	0.970	0.962	0.852	0.801	0.861	0.853	
bin	64	0.617	0.924	0.958	0.943	0.527	0.523	0.907	0.529	0.436	0.186	0.892	0.733	0.901	0.674	0.901	0.858	0.855	0.660	0.981	0.989	0.995	0.991	0.947	0.991	0.832	0.854	0.872	0.966	0.970	0.962	0.850	0.796	0.862	0.855					

Table 16: Relative evaluation time for stella_en_400M_v5 on all tasks. Table 2 serves as a legend.

q	m	Total Time	Average All Tasks	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ
f32	8192	4295.55	113.04	36.26	100.80	22.66	139.56	524.76	241.86	52.36
f32	4096	4463.78	117.47	35.74	101.87	22.57	139.06	582.55	238.88	51.97
f32	2048	4470.14	117.64	35.79	102.71	22.56	138.97	582.55	238.20	51.87
f32	1024	4461.36	117.40	36.29	102.21	22.60	138.89	581.40	236.73	51.84
f32	512	4393.43	115.62	31.13	100.51	22.52	138.27	582.46	233.35	51.84
f32	256	4342.03	114.26	28.06	98.83	22.51	137.80	582.05	230.84	52.01
f32	128	4156.98	109.39	24.78	88.11	22.44	121.08	582.73	229.96	51.86
f32	64	4143.28	109.03	24.33	87.41	22.46	120.48	582.54	229.79	51.54
int	8192	4325.80	113.84	36.46	101.71	22.70	140.11	565.73	209.12	45.12
int	4096	4143.34	109.04	36.43	101.50	22.69	140.22	507.54	207.28	44.93
int	2048	4151.41	109.25	36.85	102.07	22.65	140.19	507.15	207.49	44.92
int	1024	4130.42	108.70	36.44	101.67	22.59	140.12	506.92	203.72	44.04
int	512	4067.18	107.03	31.19	99.73	22.50	139.09	507.07	203.96	44.89
int	256	4019.45	105.78	28.21	98.30	22.51	138.08	507.40	201.51	44.67
int	128	3831.03	100.82	24.67	87.61	22.48	120.55	507.20	201.58	44.70
int	64	3833.36	100.88	24.43	87.48	22.40	120.76	507.51	203.14	44.85
bin	8192	3990.36	105.01	24.93	97.78	22.45	137.78	507.48	202.69	44.75
bin	4096	3995.47	105.14	24.83	98.04	22.44	138.15	507.88	203.22	44.35
bin	2048	3988.21	104.95	24.81	97.97	22.46	137.84	507.61	201.54	44.37
bin	1024	4024.40	104.88	24.76	97.82	22.45	137.74	506.75	202.09	44.90
bin	512	3976.34	104.64	24.40	97.45	22.43	137.44	507.31	201.09	44.85
bin	256	3888.11	102.32	24.19	88.89	22.43	137.20	507.38	200.94	44.82
bin	128	3813.89	100.37	23.91	86.70	22.42	120.31	507.25	201.36	44.65
bin	64	3811.24	100.30	23.73	86.58	22.41	120.63	507.42	200.83	44.78

Table 17: Evaluation time for stella_en_400M_v5, averaged per category. Table 2 serves as a legend.

q	m	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ	Average all
f32	8192	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
f32	4096	0.988	0.975	0.994	0.992	1.108	0.986	0.993	0.996
f32	2048	0.992	0.987	0.995	0.995	1.108	0.984	0.991	1.000
f32	1024	0.999	0.980	0.997	0.994	1.106	0.980	0.990	1.000
f32	512	0.881	0.932	0.992	0.984	1.108	0.971	0.990	0.960
f32	256	0.818	0.873	0.990	0.976	1.107	0.963	0.993	0.929
f32	128	0.767	0.752	0.987	0.861	1.109	0.961	0.991	0.877
f32	64	0.758	0.739	0.987	0.851	1.108	0.961	0.984	0.870
int	8192	0.991	1.000	0.998	1.009	1.077	0.899	0.862	0.993
int	4096	0.994	0.992	0.996	1.008	0.966	0.856	0.858	0.979
int	2048	1.000	1.011	0.996	1.013	0.965	0.856	0.858	0.985
int	1024	0.993	0.996	0.992	1.010	0.965	0.845	0.841	0.977
int	512	0.877	0.909	0.991	0.993	0.965	0.846	0.857	0.929
int	256	0.822	0.868	0.989	0.981	0.966	0.839	0.853	0.904
int	128	0.762	0.754	0.986	0.854	0.965	0.838	0.854	0.851
int	64	0.762	0.740	0.985	0.854	0.966	0.843	0.857	0.847
bin	8192	0.769	0.841	0.987	0.976	0.966	0.844	0.855	0.885
bin	4096	0.770	0.847	0.986	0.977	0.967	0.844	0.847	0.887
bin	2048	0.768	0.852	0.986	0.978	0.966	0.839	0.847	0.887
bin	1024	0.766	0.845	0.987	0.976	0.965	0.840	0.858	0.886
bin	512	0.762	0.820	0.986	0.974	0.966	0.838	0.857	0.877
bin	256	0.758	0.737	0.989	0.971	0.966	0.836	0.856	0.855
bin	128	0.750	0.722	0.984	0.848	0.966	0.838	0.853	0.839
bin	64	0.741	0.710	0.985	0.853	0.966	0.836	0.855	0.834

Table 18: Relative evaluation time for stella_en_400M_v5, averaged per category. Table 2 serves as a legend.

Table 19: Absolute performance for stella_en_400M_v5 on all tasks. Table 2 serves as a legend.

q	m	req.	CLA	CLU	STS	PCL	PCL	PCL	RET	RET	RER	RER	RER	SUM																											
		bits	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	1							
f32	8192	262144	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000								
f32	4096	131072	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000								
f32	2048	65536	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000								
f32	1024	32768	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000								
f32	512	16384	0.999	0.967	0.978	0.983	0.995	0.991	0.996	0.987	0.995	0.995	0.990	0.987	0.983	0.998	0.966	0.985	0.914	0.983	0.100	0.998	0.995	0.002	0.997	0.997	0.998	0.998	0.002	0.996	0.999	0.001	0.955	0.963	0.999	0.993	0.996	0.019			
f32	256	8192	0.992	0.923	0.945	0.961	0.979	0.979	0.990	0.966	0.973	0.964	0.973	0.983	0.920	0.967	1.006	0.969	0.989	0.992	0.980	0.990	0.992	0.990	0.994	0.994	0.996	0.988	0.991	0.995	0.963	0.903	0.879	0.994	0.958	0.978	1.004				
f32	128	4096	0.977	0.840	0.888	0.906	0.943	0.947	0.970	0.919	0.987	0.975	0.920	0.929	0.960	0.961	0.831	0.889	0.993	0.876	0.963	0.984	0.974	0.981	0.972	0.989	0.985	0.979	0.993	0.977	0.977	0.988	0.909	0.832	0.762	0.967	0.953	0.955	1.006		
f32	64	2048	0.950	0.775	0.831	0.852	0.880	0.882	0.929	0.838	0.956	0.948	0.821	0.819	0.918	0.917	0.667	0.756	0.973	0.722	0.963	0.966	0.958	0.961	0.957	0.959	0.968	0.960	0.955	0.985	0.959	0.956	0.977	0.791	0.655	0.561	0.948	0.873	0.926	0.018	
int	8192	65536	0.989	0.977	0.975	0.975	0.866	0.961	0.963	1.000	0.934	0.976	0.994	0.982	1.003	0.991	0.985	0.996	0.998	1.012	1.022	0.973	0.998	0.976	1.012	1.005	1.000	0.985	0.994	1.002	0.988	1.004	1.001	1.001	1.030	0.991	0.997	1.000	1.009	1.008	
int	4096	32768	0.989	0.977	0.975	0.866	0.961	0.968	1.000	0.934	0.976	0.994	0.982	1.003	0.991	0.985	0.996	0.998	1.012	1.022	0.973	0.998	0.976	1.012	1.005	1.000	0.985	0.994	1.002	0.988	1.004	1.001	1.001	1.030	0.991	0.997	1.000	1.009	1.008		
int	2048	16384	0.989	0.977	0.975	0.866	0.961	0.968	1.000	0.934	0.976	0.994	0.982	1.003	0.991	0.985	0.996	0.998	1.012	1.022	0.973	0.998	0.976	1.012	1.005	1.000	0.985	0.994	1.002	0.988	1.004	1.001	1.001	1.030	0.991	0.997	1.000	1.009	1.008		
int	1024	8192	0.989	0.977	0.975	0.866	0.961	0.963	1.000	0.934	0.976	0.994	0.982	1.003	0.991	0.985	0.996	0.998	1.012	1.022	0.973	0.998	0.976	1.012	1.005	1.000	0.985	0.994	1.002	0.988	1.004	1.001	1.001	1.030	0.991	0.997	1.000	1.009	1.008		
int	512	4096	0.987	0.941	0.945	0.861	0.952	0.958	0.996	0.911	0.983	0.970	0.983	0.991	0.997	0.973	0.986	1.017	0.982	0.979	0.996	0.976	1.014	1.001	0.996	0.982	0.994	1.002	0.997	1.001	0.999	1.000	1.034	0.960	0.982	0.984	0.997	1.007	1.018		
int	256	2048	0.979	0.874	0.912	0.826	0.933	0.948	0.987	0.874	0.979	0.979	0.969	0.946	0.967	0.967	0.983	0.926	0.961	1.004	0.964	0.972	0.992	0.967	1.006	0.991	0.992	0.977	0.986	0.997	0.996	0.991	0.997	1.006	0.921	0.899	0.976	0.976	0.993	1.022	
int	128	1024	0.954	0.751	0.846	0.843	0.882	0.898	0.956	0.808	0.976	0.966	0.922	0.932	0.950	0.967	0.840	0.885	0.979	0.855	0.966	0.985	0.947	0.995	0.984	0.983	0.973	0.980	0.989	0.988	0.977	0.977	0.990	0.955	0.820	0.827	0.972	0.967	0.964	1.046	
int	64	512	0.912	0.665	0.777	0.815	0.799	0.815	0.886	0.691	0.927	0.929	0.832	0.817	0.909	0.907	0.665	0.761	0.956	0.716	0.936	0.969	0.936	0.974	0.962	0.964	0.949	0.949	0.959	0.988	0.960	0.956	0.978	0.820	0.659	0.700	0.943	0.908	0.931	1.010	
bin	8192	8192	0.885	0.619	0.760	0.714	0.793	0.798	0.912	0.712	0.924	0.918	0.778	0.788	0.870	0.870	0.655	0.678	0.884	0.724	0.850	0.949	0.904	0.944	0.927	0.938	0.954	0.940	0.924	0.944	0.949	0.959	0.929	0.963	0.654	0.457	0.450	0.904	0.900	0.858	0.977
bin	4096	4096	0.885	0.619	0.760	0.814	0.793	0.798	0.912	0.712	0.924	0.918	0.778	0.788	0.870	0.870	0.655	0.678	0.884	0.724	0.850	0.949	0.904	0.944	0.927	0.938	0.954	0.940	0.924	0.944	0.949	0.959	0.929	0.963	0.654	0.457	0.450	0.906	0.900	0.858	0.977
bin	2048	2048	0.885	0.619	0.760	0.814	0.793	0.798	0.912	0.712	0.924	0.918	0.778	0.788	0.870	0.870	0.655	0.678	0.884	0.724	0.850	0.949	0.904	0.944	0.927	0.938	0.954	0.940	0.924	0.944	0.949	0.959	0.929	0.963	0.654	0.457	0.450	0.906	0.900	0.858	0.977
bin	1024	1024	0.885	0.619	0.760	0.814	0.793	0.798	0.912	0.712	0.924	0.918	0.778	0.788	0.870	0.870	0.655	0.678	0.884	0.724	0.850	0.949	0.904	0.944	0.927	0.938	0.954	0.940	0.924	0.944	0.949	0.959	0.929	0.963	0.654	0.457	0.450	0.906	0.900	0.858	0.977
bin	512	512	0.751	0.505	0.681	0.774	0.638	0.651	0.776	0.538	0.888	0.893	0.641	0.607	0.789	0.775	0.432	0.516	0.886	0.758	0.802	0.908	0.857	0.884	0.868	0.892	0.888	0.884	0.914	0.898	0.858	0.937	0.450	0.256	0.245	0.870	0.829	0.764	0.977		
bin	256	256	0.537	0.448	0.662	0.738	0.429	0.470	0.607	0.362	0.834	0.859	0.484	0.456	0.732	0.696	0.277	0.434	0.847	0.362	0.774	0.840	0.763	0.829	0.784	0.808	0.823	0.801	0.774	0.873	0.770	0.772	0.888	0.225	0.094	0.155	0.832	0.789	0.656	0.975	
bin	128	128	0.303	0.445	0.689	0.730	0.261	0.302	0.432	0.254	0.816	0.824	0.355	0.334	0.629	0.627	0.204	0.387	0.846	0.288	0.505	0.721	0.646	0.728	0.667	0.718	0.727	0.711	0.717	0.789	0.554	0.673	0.815	0.089	0.048	0.066	0.774	0.761	0.565	0.951	
bin	64	64	0.249	0.243	0.679	0.715	0.200	0.284	0.357	0.163	0.806	0.814	0.270	0.257	0.618	0.615	0.185	0.367	0.819	0.244	0.422	0.548	0.521	0.529	0.507	0.610	0.593	0.561	0.541	0.711	0.304	0.560	0.704	0.062	0.018	0.058	0.732	0.725	0.487	0.904	

Table 20: Relative performance for stella_en_400M_v5 on all tasks. Table 2 serves as a legend.

q	m	req. bits	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ	Average all
f32	8192	262144	0.778	0.500	0.826	0.859	0.477	0.485	0.306	0.664
f32	4096	131072	0.778	0.500	0.826	0.859	0.477	0.485	0.306	0.664
f32	2048	65536	0.778	0.500	0.826	0.859	0.477	0.485	0.306	0.664
f32	1024	32768	0.778	0.500	0.826	0.859	0.477	0.485	0.306	0.664
f32	512	16384	0.769	0.495	0.826	0.857	0.465	0.483	0.312	0.660
f32	256	8192	0.755	0.482	0.818	0.852	0.439	0.475	0.307	0.649
f32	128	4096	0.724	0.465	0.809	0.843	0.401	0.465	0.308	0.631
f32	64	2048	0.680	0.422	0.795	0.828	0.325	0.448	0.311	0.598
int	8192	65536	0.747	0.498	0.821	0.861	0.481	0.486	0.308	0.657
int	4096	32768	0.747	0.498	0.821	0.861	0.481	0.486	0.308	0.657
int	2048	16384	0.747	0.498	0.821	0.861	0.481	0.486	0.308	0.657
int	1024	8192	0.747	0.498	0.821	0.861	0.481	0.486	0.308	0.657
int	512	4096	0.736	0.493	0.821	0.859	0.474	0.482	0.311	0.652
int	256	2048	0.717	0.482	0.816	0.853	0.452	0.476	0.313	0.641
int	128	1024	0.681	0.462	0.809	0.843	0.417	0.469	0.320	0.623
int	64	512	0.625	0.418	0.792	0.829	0.350	0.451	0.309	0.587
bin	8192	8192	0.620	0.403	0.766	0.818	0.254	0.430	0.299	0.565
bin	4096	4096	0.620	0.403	0.766	0.818	0.254	0.430	0.299	0.565
bin	2048	2048	0.620	0.403	0.766	0.818	0.254	0.430	0.299	0.565
bin	1024	1024	0.620	0.403	0.766	0.818	0.254	0.430	0.299	0.565
bin	512	512	0.523	0.347	0.724	0.772	0.157	0.399	0.299	0.505
bin	256	256	0.412	0.298	0.666	0.696	0.078	0.369	0.298	0.438
bin	128	128	0.320	0.266	0.571	0.582	0.033	0.338	0.291	0.371
bin	64	64	0.287	0.251	0.456	0.443	0.022	0.312	0.276	0.315

Table 21: Performance for stella_en_400M_v5, averaged per category. Table 2 serves as a legend.

q	m	req. bits	Average Relative classification	Average Relative clustering	Average Relative sts	Average Relative pairclass	Average Relative retrieval	Average Relative rerank	Average Relative summ	Average Relative all
f32	8192	262144	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
f32	4096	131072	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
f32	2048	65536	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
f32	1024	32768	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
f32	512	16384	0.987	0.990	1.000	0.998	0.973	0.996	1.019	0.992
f32	256	8192	0.967	0.966	0.991	0.992	0.915	0.977	1.004	0.973
f32	128	4096	0.924	0.932	0.980	0.981	0.834	0.958	1.006	0.943
f32	64	2048	0.867	0.848	0.963	0.964	0.669	0.916	1.018	0.887
int	8192	65536	0.959	0.996	0.993	1.002	1.006	1.002	1.008	0.990
int	4096	32768	0.959	0.996	0.993	1.002	1.006	1.002	1.008	0.990
int	2048	16384	0.959	0.996	0.993	1.002	1.006	1.002	1.008	0.990
int	1024	8192	0.959	0.996	0.993	1.002	1.006	1.002	1.008	0.990
int	512	4096	0.944	0.987	0.993	1.000	0.992	0.996	1.018	0.982
int	256	2048	0.916	0.967	0.988	0.993	0.942	0.982	1.022	0.964
int	128	1024	0.867	0.927	0.979	0.981	0.867	0.968	1.046	0.934
int	64	512	0.795	0.842	0.959	0.965	0.727	0.927	1.010	0.875
bin	8192	8192	0.786	0.810	0.927	0.951	0.520	0.888	0.977	0.835
bin	4096	4096	0.786	0.810	0.927	0.951	0.520	0.888	0.977	0.835
bin	2048	2048	0.786	0.810	0.927	0.951	0.520	0.888	0.977	0.835
bin	1024	1024	0.786	0.810	0.927	0.951	0.520	0.888	0.977	0.835
bin	512	512	0.664	0.698	0.877	0.897	0.317	0.821	0.977	0.741
bin	256	256	0.532	0.598	0.807	0.810	0.158	0.759	0.975	0.644
bin	128	128	0.426	0.531	0.693	0.681	0.068	0.700	0.951	0.551
bin	64	64	0.387	0.499	0.554	0.523	0.046	0.648	0.904	0.479

Table 22: Relative performance for stella_en_400M_v5, averaged per category. Table 2 serves as a legend.

E. nomic-embed-text-v1.5

This section contains additional figures and tables for the NOMIC model.

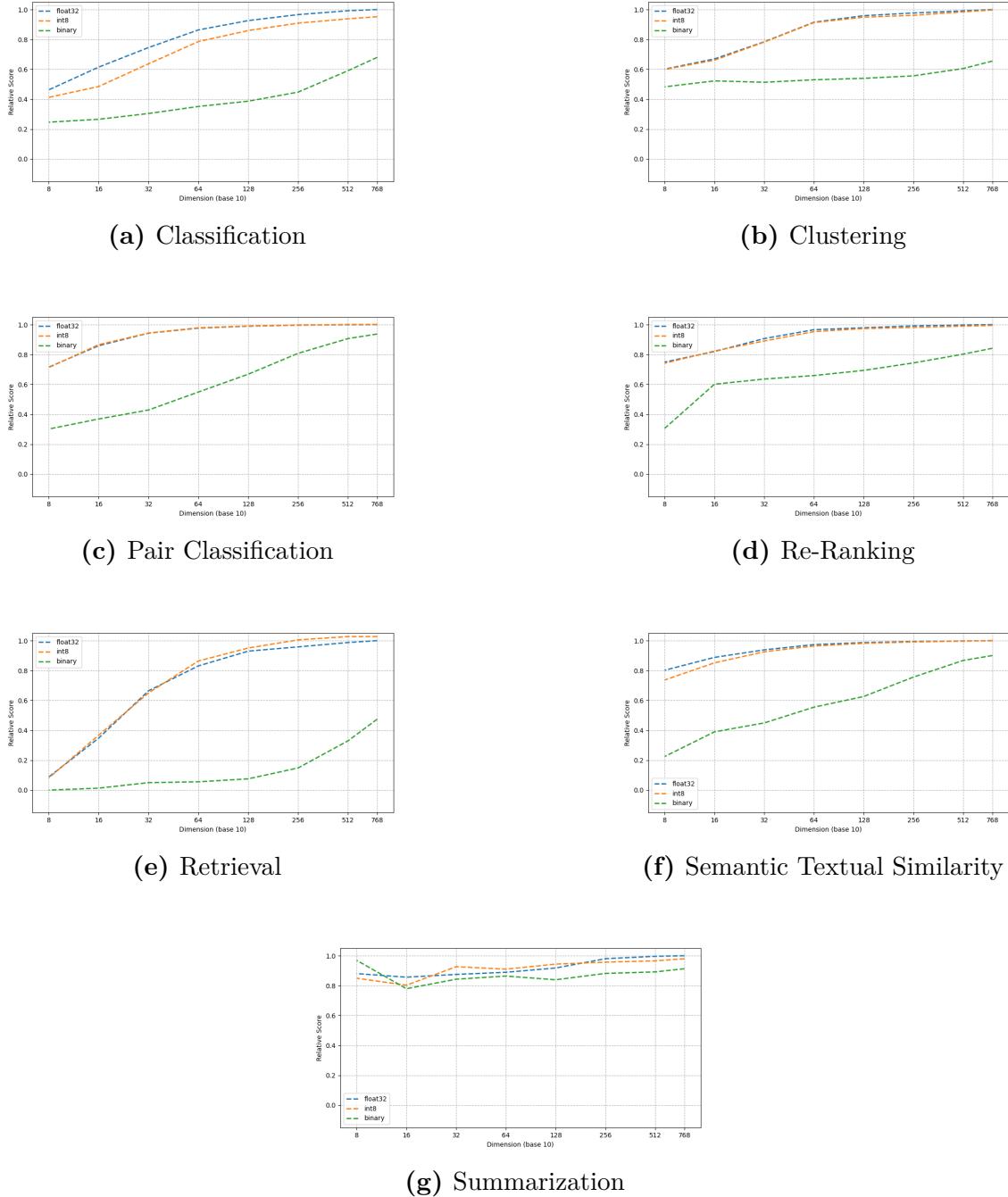


Figure 13: Relative performance of the NOMIC model for each task category in the MTEB. The x-axis of each subfigure represents the embedding size while the y-axis represents the relative performance. For each task the reference point is the performance of the float32 embeddings at the full embedding size. Blue: float32, Orange: int8, Green: binary.

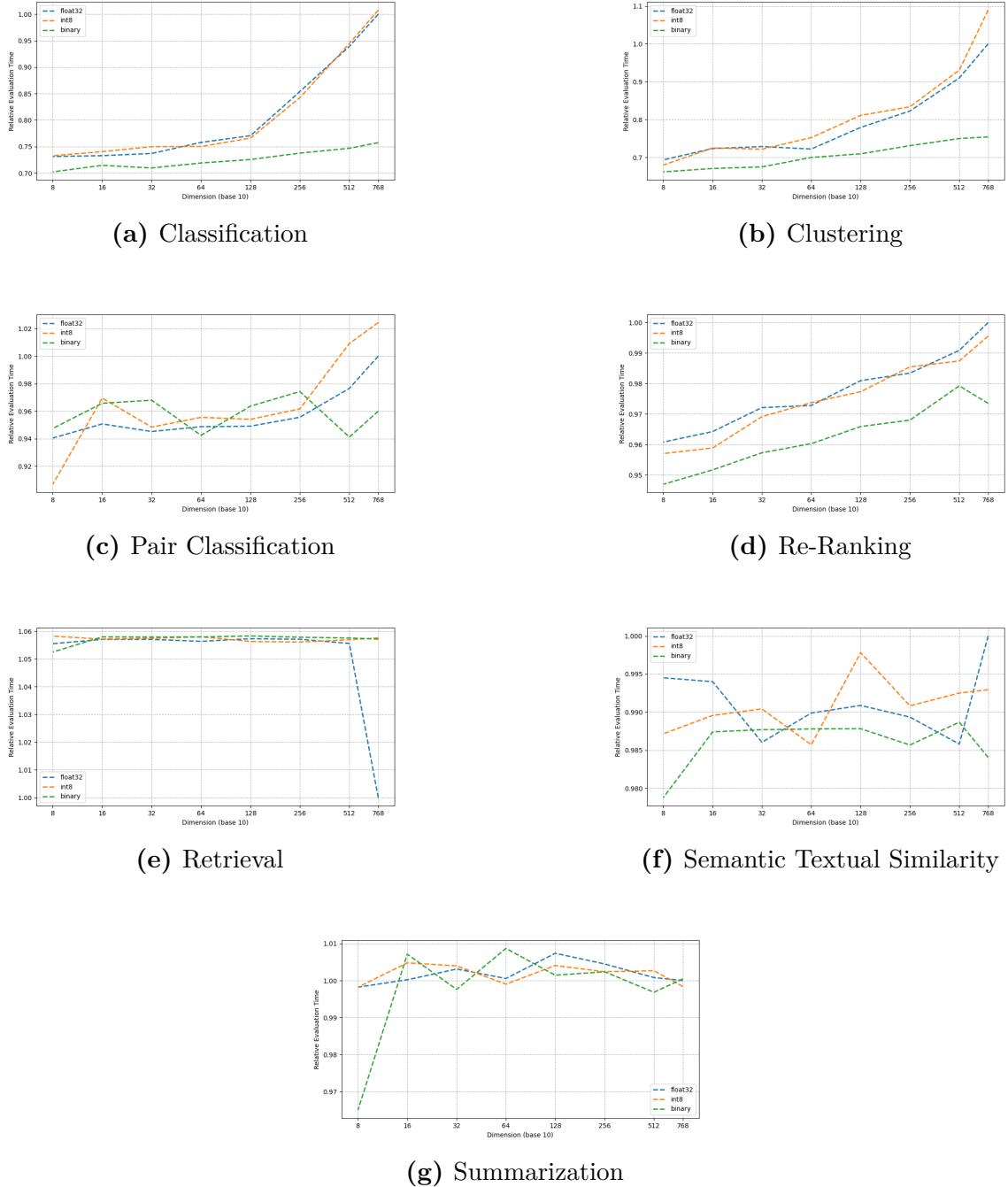


Figure 14: Relative evaluation time of the NOMIC model for each task category in the MTEB. The x-axis of each subfigure represents the embedding size while the y-axis represents the relative evaluation time. For each task the reference point is the evaluation time of the float32 embeddings at the full embedding size. Blue: float32, Orange: int8, Green: binary.

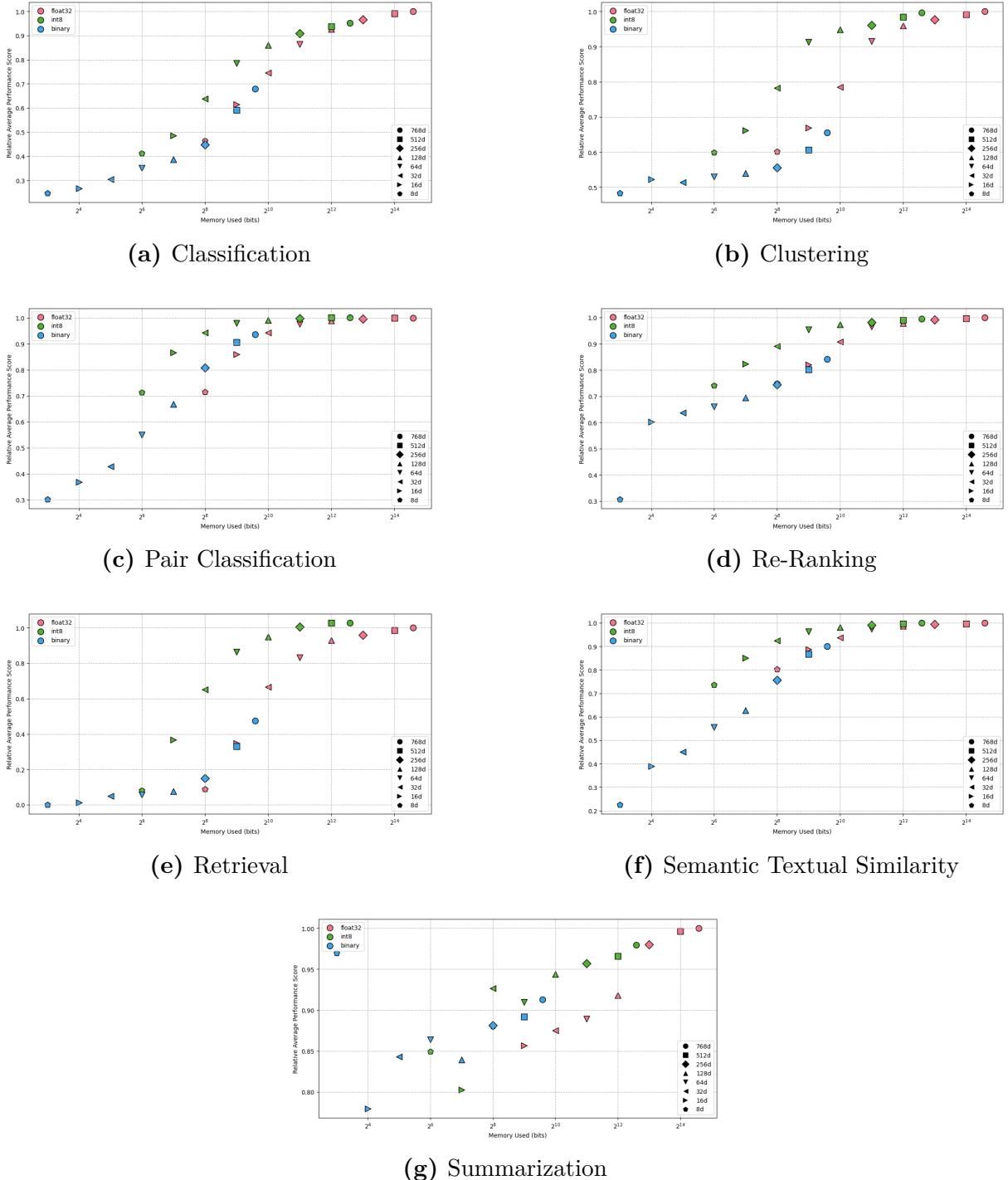


Figure 15: Relative accuracy compute trade-off of the NOMIC model, averaged over all selected MTEB tasks within each category. The x-axis represents memory consumption (log scale) and the y-axis represents relative performance (averaged over tasks within the category). Reference point: float32 embeddings at full size.

Table 23: Absolute evaluation time for nomic-embed-text-v1.5 on all tasks. Table 2 serves as a legend.

q	m	CLA	CLU	STS	PCL	PCL	PCL	RET	RET	RER	RER	RER	SUM																											
		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	1	2	3					
f32	768	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000								
f32	512	0.945	0.890	1.067	1.047	0.892	0.853	0.963	0.856	0.934	0.774	1.008	0.827	nan	0.843	0.982	0.978	0.964	0.877	0.999	0.997	0.992	1.000	0.997	0.990	0.979	0.988	0.931	nan	0.953	0.984	0.992	1.056	nan	1.000	1.000	1.000	1.000	1.000	
f32	256	0.828	0.921	1.097	1.033	0.479	0.852	0.921	0.700	0.666	0.512	0.999	0.837	nan	0.659	0.968	0.932	0.929	0.911	1.023	0.995	0.998	1.014	0.997	0.994	0.991	0.975	0.917	nan	0.881	1.000	0.986	1.057	nan	0.985	0.963	1.002	1.004	0.997	
f32	128	0.556	0.878	1.069	1.041	0.441	0.870	0.882	0.428	0.562	0.647	0.993	0.684	nan	0.645	0.964	0.935	0.896	0.685	1.010	1.001	0.992	1.015	0.994	0.993	0.984	0.978	0.951	nan	0.862	0.995	0.990	1.057	nan	0.985	0.958	1.000	1.000	0.007	
f32	64	0.528	0.897	1.086	1.032	0.414	0.833	0.883	0.390	0.516	0.321	0.985	0.688	nan	0.561	0.961	0.913	0.896	0.658	1.004	0.998	0.994	1.010	1.000	1.003	0.988	0.980	0.932	nan	0.864	0.987	0.996	1.056	nan	0.982	0.937	1.000	1.001	0.001	
f32	32	0.526	0.840	1.039	1.033	0.390	0.734	0.940	0.394	0.566	0.323	0.981	0.731	nan	0.545	0.959	0.911	0.897	0.645	1.019	0.988	0.983	1.010	0.996	1.001	0.989	1.002	0.886	nan	0.862	0.986	0.988	1.057	nan	0.982	0.935	0.999	1.003	0.001	
f32	16	0.515	0.851	1.024	1.030	0.400	0.792	0.870	0.381	0.464	0.335	0.988	0.644	nan	0.581	0.958	0.930	0.882	0.725	1.047	1.000	0.996	1.016	0.994	0.990	0.996	0.976	0.931	nan	0.888	0.978	0.986	1.057	nan	0.988	0.907	0.998	1.000	0.001	
f32	8	0.507	0.851	1.030	0.171	0.397	0.802	0.874	0.373	0.419	0.296	0.980	0.618	nan	0.544	0.954	0.910	0.881	0.637	1.050	0.999	0.994	1.014	0.993	0.998	0.996	0.977	0.929	nan	0.862	0.975	0.975	1.056	nan	0.983	0.901	0.998	0.998	0.001	
int	768	1.063	0.898	1.075	1.059	1.027	0.923	0.959	0.157	1.193	1.309	1.028	1.145	nan	1.018	1.014	1.042	0.995	1.057	1.009	1.008	1.002	1.012	1.010	1.004	0.997	1.004	0.892	nan	1.070	0.999	1.004	1.058	nan	0.988	0.993	1.006	0.998	0.001	
int	512	0.960	0.855	1.047	1.121	0.896	0.875	0.944	0.859	1.070	0.774	1.005	0.843	nan	0.873	0.981	0.960	0.939	0.931	1.050	1.014	0.995	1.012	1.004	0.993	1.001	0.978	0.887	nan	1.035	1.003	0.990	1.057	nan	0.985	0.974	1.003	1.003	0.001	
int	256	0.835	0.868	1.076	1.047	0.449	0.807	0.954	0.705	0.729	0.552	0.991	0.830	nan	0.660	0.974	0.936	0.905	0.927	1.015	0.999	0.992	1.004	0.997	0.996	0.987	0.988	0.940	nan	0.896	1.002	0.987	1.056	nan	0.988	0.966	1.002	1.002	0.001	
int	128	0.582	0.844	1.043	1.121	0.438	0.753	0.793	0.916	0.431	0.760	0.462	0.987	0.710	nan	0.702	0.972	0.937	0.889	0.870	1.056	1.000	1.001	1.020	0.993	0.998	1.000	0.979	0.932	nan	0.880	0.999	0.983	1.056	nan	0.981	0.949	1.002	1.004	0.001
int	64	0.541	0.852	1.046	1.070	0.402	0.478	0.916	0.397	0.700	0.346	0.985	0.739	nan	0.557	0.962	0.912	0.901	0.667	1.014	0.991	0.986	1.004	1.003	1.001	0.992	1.002	0.879	nan	0.890	0.985	0.985	1.058	nan	0.987	0.934	1.000	0.999	0.001	
int	32	0.521	0.846	1.041	1.137	0.397	0.734	0.930	0.391	0.578	0.277	0.982	0.723	nan	0.530	0.957	0.906	0.879	0.679	1.064	0.988	1.000	1.000	1.002	1.002	0.981	0.984	0.954	nan	0.870	0.986	0.989	1.058	nan	0.981	0.927	1.000	1.004	0.001	
int	16	0.519	0.841	1.043	1.050	0.394	0.812	0.888	0.378	0.532	0.370	0.995	0.688	nan	0.553	0.967	0.905	0.882	0.631	0.999	1.004	0.987	1.010	0.993	0.999	1.017	0.987	0.910	nan	0.942	0.983	0.984	1.057	nan	0.978	0.899	0.999	1.005	0.001	
int	8	0.519	0.824	1.027	1.015	0.383	0.826	0.888	0.379	0.456	0.296	0.981	0.627	nan	0.490	0.921	0.891	0.862	0.593	0.992	0.997	0.996	0.992	0.998	0.982	1.001	0.934	nan	0.811	0.928	0.981	1.058	nan	0.983	0.890	0.998	0.998	0.001		
bin	768	0.552	0.834	1.042	1.083	0.431	0.818	0.886	0.414	0.554	0.409	0.988	0.750	nan	0.569	0.963	0.919	0.885	0.748	1.023	0.999	0.990	1.004	0.995	0.996	1.006	0.975	0.869	nan	0.915	0.984	0.981	1.057	nan	0.985	0.936	1.000	1.000	0.001	
bin	512	0.527	0.863	1.067	1.040	0.402	0.816	0.875	0.385	0.498	0.424	0.992	0.669	nan	0.620	0.956	0.918	0.881	0.791	1.052	0.998	1.000	1.019	0.990	0.994	0.996	0.974	0.875	nan	0.860	0.980	0.975	1.058	nan	0.987	0.949	1.001	0.997	0.001	
bin	256	0.532	0.837	1.056	1.032	0.386	0.795	0.873	0.391	0.480	0.473	0.988	0.649	nan	0.590	0.958	0.920	0.882	0.638	1.009	1.004	0.990	1.005	0.994	0.999	0.993	0.995	0.884	nan	0.937	1.008	0.977	1.058	nan	0.977	0.930	0.998	1.002	0.001	
bin	128	0.527	0.815	1.011	1.025	0.380	0.788	0.870	0.389	0.493	0.302	0.981	0.649	nan	0.559	0.956	0.919	0.877	0.649	1.024	1.006	0.999	0.996	1.006	0.985	0.971	0.905	1.008	0.978	1.058	nan	0.985	0.912	1.000	1.001	0.001				
bin	64	0.517	0.808	1.007	1.037	0.379	0.733	0.935	0.383	0.469	0.270	0.981	0.681	nan	0.511	0.955	0.903	0.875	0.653	1.005	0.995	1.001	0.999	0.997	1.000	0.986	0.996	0.912	nan	0.863	0.973	0.992	1.058	nan	0.980	0.903	0.999	1.009	0.001	
bin	32	0.521	0.808	0.998	1.021	0.376	0.742	0.834	0.378	0.418	0.209	0.979	0.611	nan	0.512	0.959	0.919	0.874	0.593	1.009	1.003	0.991	1.006	1.001	1.002	1.012	0.983	0.884	nan	0.941	0.981	0.982	1.058	nan	0.988	0.885	0.999	0.998	0.001	
bin	16	0.527	0.814	1.016	1.010	0.385	0.767	0.820	0.376	0.389	0.240	0.979	0.612	nan	0.510	0.959	0.903	0.871	0.572	1.019	1.006	0.989	1.012	1.001	1.001	1.010	0.973	0.878	nan	0.928	0.981	0.988	1.058	nan	0.983	0.873	0.999	1.007	0.001	
bin	8	0.503	0.791	1.028	1.018	0.377	0.717	0.813	0.366	0.434	0.166	0.976	0.624	nan	0.460	0.948	0.903	0.869	0.576	1.001	0.984	0.988	0.999	0.998	0.996	0.987	0.872	nan	0.889	0.974	0.980	1.052	nan	0.982	0.861	0.997	0.965	0.001		

Table 24: Relative evaluation time for nomic-embed-text-v1.5 on all tasks. Table 2 serves as a legend.

q	m	Total Time	Average All Tasks	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ
f32	768	1449.71	42.64	20.00	54.20	8.20	52.10	212.30	113.96	17.58
f32	512	1426.14	41.95	18.27	52.55	8.15	51.36	224.11	112.64	17.59
f32	256	1379.38	40.57	15.28	50.49	8.16	50.78	224.43	111.62	17.66
f32	128	1345.21	39.57	12.10	49.66	8.17	50.75	224.48	111.19	17.71
f32	64	1330.67	39.14	11.64	48.88	8.18	50.90	224.27	109.81	17.59
f32	32	1326.94	39.03	11.37	48.87	8.14	50.58	224.42	109.69	17.63
f32	16	1320.92	38.85	11.25	48.86	8.18	50.62	224.43	107.88	17.58
f32	8	1312.62	38.61	11.18	48.26	8.18	50.38	224.09	107.50	17.55
int	768	1482.13	43.59	20.64	55.71	8.24	52.74	224.54	113.71	17.55
int	512	1425.17	41.92	18.35	52.13	8.23	51.99	224.40	112.39	17.62
int	256	1377.82	40.52	15.12	50.36	8.17	50.93	224.22	111.84	17.62
int	128	1345.79	39.58	12.10	49.92	8.19	50.64	224.26	110.72	17.65
int	64	1332.37	39.19	11.57	49.22	8.16	50.71	224.62	109.70	17.56
int	32	1323.42	38.92	11.43	48.51	8.19	50.67	224.51	109.18	17.65
int	16	1320.31	38.83	11.31	48.79	8.19	50.95	224.45	107.38	17.66
int	8	1299.73	38.23	11.22	47.29	8.15	49.50	224.67	106.75	17.55
bin	768	1333.40	39.22	11.83	49.10	8.16	50.67	224.45	109.76	17.59
bin	512	1330.04	39.12	11.47	48.92	8.17	50.11	224.53	110.71	17.52
bin	256	1326.59	39.02	11.38	48.80	8.18	50.91	224.59	109.26	17.62
bin	128	1319.24	38.80	11.22	48.47	8.20	50.72	224.69	108.32	17.60
bin	64	1312.76	38.61	11.08	48.18	8.17	50.62	224.61	107.58	17.73
bin	32	1309.42	38.51	10.99	48.17	8.19	50.87	224.60	106.49	17.54
bin	16	1306.26	38.42	11.08	47.92	8.20	51.01	224.62	105.72	17.70
bin	8	1294.65	38.08	10.81	47.66	8.11	50.36	223.45	104.90	16.96

Table 25: Evaluation time for nomic-embed-text-v1.5, averaged per category. Table 2 serves as a legend.

q	m	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ	Average all
f32	768	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
f32	512	0.939	0.910	0.986	0.977	1.056	0.991	1.001	0.957
f32	256	0.854	0.823	0.989	0.956	1.057	0.983	1.004	0.912
f32	128	0.771	0.779	0.991	0.949	1.057	0.981	1.007	0.881
f32	64	0.758	0.722	0.990	0.949	1.056	0.973	1.001	0.862
f32	32	0.737	0.729	0.986	0.945	1.057	0.972	1.003	0.857
f32	16	0.733	0.723	0.994	0.951	1.057	0.964	1.000	0.857
f32	8	0.731	0.693	0.995	0.941	1.056	0.961	0.998	0.847
int	768	1.008	1.089	0.993	1.024	1.058	0.996	0.998	1.027
int	512	0.945	0.931	0.993	1.009	1.057	0.987	1.003	0.968
int	256	0.842	0.834	0.991	0.962	1.056	0.986	1.002	0.914
int	128	0.766	0.811	0.998	0.954	1.056	0.977	1.004	0.890
int	64	0.750	0.752	0.986	0.956	1.058	0.974	0.999	0.867
int	32	0.750	0.722	0.990	0.948	1.058	0.969	1.004	0.860
int	16	0.741	0.725	0.990	0.970	1.057	0.959	1.005	0.859
int	8	0.733	0.680	0.987	0.907	1.058	0.957	0.998	0.839
bin	768	0.757	0.754	0.984	0.960	1.057	0.974	1.000	0.869
bin	512	0.747	0.750	0.989	0.941	1.058	0.979	0.997	0.866
bin	256	0.738	0.731	0.986	0.974	1.058	0.968	1.002	0.860
bin	128	0.726	0.709	0.988	0.964	1.058	0.966	1.001	0.851
bin	64	0.719	0.700	0.988	0.943	1.058	0.960	1.009	0.845
bin	32	0.710	0.675	0.988	0.968	1.058	0.957	0.998	0.837
bin	16	0.715	0.671	0.987	0.966	1.058	0.952	1.007	0.837
bin	8	0.702	0.662	0.979	0.948	1.052	0.947	0.965	0.826

Table 26: Relative evaluation time for nomic-embed-text-v1.5, averaged per category. Table 2 serves as a legend.

Table 27: Absolute performance for nomic-embed-text-v1.5 on all tasks. Table 2 serves as a legend.

Table 28: Relative performance for nomic-embed-text-v1.5 on all tasks. Table 2 serves as a legend

q	m	req. bits	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ	Average all
f32	768	24576	0.724	0.444	0.832	0.851	0.536	0.477	0.325	0.650
f32	512	16384	0.718	0.441	0.829	0.850	0.529	0.476	0.324	0.647
f32	256	8192	0.701	0.434	0.827	0.848	0.514	0.474	0.319	0.640
f32	128	4096	0.673	0.427	0.821	0.842	0.498	0.468	0.298	0.628
f32	64	2048	0.629	0.408	0.809	0.831	0.445	0.463	0.289	0.606
f32	32	1024	0.542	0.351	0.780	0.803	0.356	0.435	0.284	0.555
f32	16	512	0.444	0.301	0.737	0.729	0.186	0.394	0.278	0.492
f32	8	256	0.325	0.272	0.665	0.603	0.047	0.358	0.286	0.419
int	768	6144	0.690	0.442	0.831	0.852	0.551	0.475	0.318	0.642
int	512	4096	0.681	0.437	0.829	0.851	0.550	0.473	0.314	0.638
int	256	2048	0.661	0.427	0.824	0.849	0.539	0.468	0.311	0.628
int	128	1024	0.626	0.422	0.816	0.843	0.509	0.465	0.307	0.615
int	64	512	0.574	0.407	0.801	0.833	0.463	0.455	0.296	0.591
int	32	256	0.464	0.350	0.769	0.804	0.348	0.426	0.301	0.533
int	16	128	0.346	0.297	0.707	0.736	0.197	0.393	0.261	0.460
int	8	64	0.286	0.271	0.611	0.601	0.044	0.353	0.276	0.394
bin	768	768	0.496	0.294	0.749	0.799	0.254	0.404	0.297	0.515
bin	512	512	0.429	0.273	0.722	0.773	0.177	0.385	0.290	0.480
bin	256	256	0.320	0.253	0.629	0.689	0.080	0.355	0.286	0.412
bin	128	128	0.270	0.246	0.521	0.565	0.040	0.329	0.273	0.355
bin	64	64	0.244	0.243	0.462	0.460	0.030	0.308	0.281	0.321
bin	32	32	0.208	0.235	0.373	0.355	0.027	0.297	0.274	0.276
bin	16	16	0.177	0.239	0.322	0.303	0.007	0.281	0.253	0.249
bin	8	8	0.163	0.221	0.185	0.245	0.000	0.151	0.315	0.190

Table 29: Performance for nomic-embed-text-v1.5, averaged per category. Table 2 serves as a legend.

q	m	req. bits	Average classification	Average Relative classification	Average clustering	Average Relative clustering	Average sts	Average Relative sts	Average pairclass	Average Relative pairclass	Average retrieval	Average Relative retrieval	Average rerank	Average Relative rerank	Average summ	Average Relative summ	Average all	Average Relative all
f32	768	24576	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
f32	512	16384	0.992	0.991	0.997	1.000	0.987	0.998	0.996	0.994								
f32	256	8192	0.966	0.977	0.994	0.997	0.958	0.992	0.980	0.982								
f32	128	4096	0.926	0.959	0.987	0.990	0.930	0.979	0.918	0.961								
f32	64	2048	0.864	0.915	0.974	0.977	0.831	0.966	0.889	0.925								
f32	32	1024	0.746	0.784	0.938	0.943	0.664	0.907	0.875	0.840								
f32	16	512	0.615	0.670	0.888	0.859	0.348	0.820	0.857	0.741								
f32	8	256	0.463	0.602	0.802	0.716	0.088	0.750	0.880	0.638								
int	768	6144	0.952	0.997	1.000	1.001	1.027	0.995	0.980	0.988								
int	512	4096	0.938	0.984	0.997	1.001	1.027	0.990	0.966	0.980								
int	256	2048	0.909	0.961	0.991	0.998	1.005	0.981	0.957	0.963								
int	128	1024	0.860	0.949	0.982	0.991	0.951	0.974	0.944	0.942								
int	64	512	0.786	0.913	0.964	0.979	0.864	0.953	0.910	0.904								
int	32	256	0.637	0.782	0.925	0.944	0.650	0.891	0.927	0.810								
int	16	128	0.485	0.661	0.851	0.866	0.367	0.824	0.803	0.698								
int	8	64	0.412	0.600	0.737	0.713	0.081	0.742	0.850	0.606								
bin	768	768	0.680	0.655	0.900	0.937	0.474	0.842	0.913	0.770								
bin	512	512	0.591	0.606	0.868	0.907	0.330	0.803	0.892	0.716								
bin	256	256	0.448	0.556	0.756	0.808	0.148	0.744	0.882	0.620								
bin	128	128	0.387	0.540	0.626	0.669	0.075	0.694	0.839	0.547								
bin	64	64	0.352	0.530	0.555	0.549	0.056	0.659	0.864	0.504								
bin	32	32	0.304	0.513	0.450	0.429	0.050	0.636	0.843	0.447								
bin	16	16	0.266	0.523	0.390	0.369	0.013	0.602	0.779	0.413								
bin	8	8	0.247	0.483	0.225	0.301	0.000	0.306	0.969	0.327								

Table 30: Relative performance for nomic-embed-text-v1.5, averaged per category. Table 2 serves as a legend.

F. mxbai-embed-2d-large-v1

This section contains additional figures and tables for the MBAI-2D model.

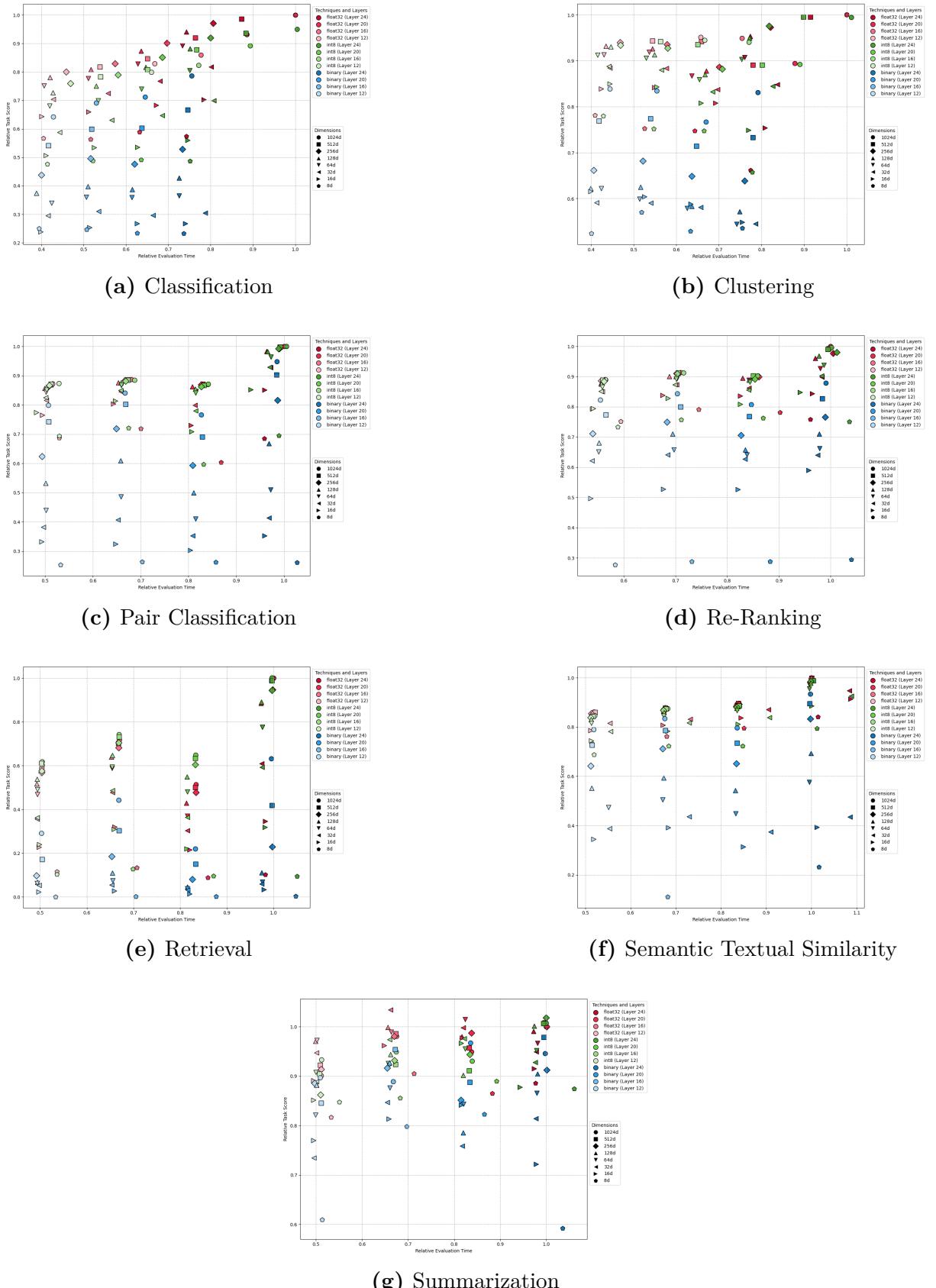


Figure 16: Trade-off between memory consumption, evaluation time, and performance of MBAI-2D, averaged over all selected MTEB tasks for each task category. The x-axis represents the relative evaluation time while the y-axis represents the relative performance. The color of the markers represent the quantization technique and inference layer, while the shape represents the embedding size of the respective MR-quantization combination.

Table 31: Absolute evaluation time for mxbai-embed-2d-large-v1 on all tasks using float quantization. Table 2 serves as a legend.

Table 32: Absolute evaluation time for mxbai-embed-2d-large-v1 on all tasks using int quantization. Table 2 serves as a legend.

Table 33: Absolute evaluation time for mxbai-embed-2d-large-v1 on all tasks using binary quantization. Table 2 serves as a legend.

Table 34: Relative evaluation time for mxbai-embed-2d-large-v1 on all tasks using float quantization. Table 2 serves as a legend.

q	m	1	CLA	CLU	STS	PCL	PCL	PCL	RET	RET	RER	RER	RER	SUM																												
		1	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	1	2	3	1					
int	1024	24	1.032	0.970	0.988	1.012	1.043	0.990	0.958	1.039	0.986	0.939	1.008	1.015	0.983	1.098	1.003	1.008	0.993	1.071	0.995	1.008	1.006	1.001	1.004	1.005	1.009	1.008	1.013	0.989	1.010	1.004	1.001	0.991	0.999	1.000	1.007	0.994	1.004	1.000		
int	1024	20	0.924	0.851	0.844	0.854	0.950	0.920	0.836	0.971	0.899	0.896	0.836	0.995	0.840	1.020	0.843	0.850	0.847	0.872	0.820	0.847	0.839	0.834	0.845	0.847	0.849	0.839	0.852	0.838	0.848	0.842	0.835	0.830	0.834	0.836	0.843	0.897	0.836	0.839		
int	1024	16	0.835	0.709	0.684	0.705	0.847	0.814	0.701	0.876	0.854	0.965	0.684	0.797	0.693	0.832	0.689	0.696	0.699	0.794	0.674	0.682	0.682	0.665	0.670	0.687	0.698	0.681	0.712	0.667	0.702	0.685	0.675	0.666	0.669	0.671	0.681	0.792	0.673	0.674		
int	1024	12	0.714	0.557	0.558	0.564	0.785	0.733	0.571	0.797	1.131	0.528	0.664	0.536	0.725	0.526	0.540	0.567	0.654	0.510	0.522	0.527	0.516	0.506	0.519	0.522	0.515	0.557	0.514	0.557	0.520	0.509	0.503	0.505	0.513	0.676	0.503	0.512				
int	512	24	0.850	0.982	0.998	1.038	0.864	0.562	0.979	0.797	0.730	0.742	0.968	0.942	0.965	0.892	0.986	1.003	1.007	0.994	1.000	1.006	1.005	1.046	0.987	0.971	1.006	0.998	1.004	0.953	1.028	0.994	1.024	0.983	0.844	0.874	0.843	0.844	0.843	0.847	0.860	0.833
int	512	16	0.672	0.698	0.680	0.712	0.706	0.414	0.674	0.649	0.573	0.603	0.666	0.664	0.671	0.728	0.667	0.668	0.646	0.619	0.669	0.680	0.678	0.663	0.662	0.683	0.691	0.681	0.704	0.670	0.670	0.681	0.663	0.665	0.671	0.677	0.745	0.693	0.673			
int	512	12	0.590	0.551	0.537	0.583	0.636	0.315	0.522	0.583	0.591	0.719	0.514	0.605	0.522	0.644	0.507	0.507	0.509	0.516	0.515	0.522	0.517	0.501	0.515	0.520	0.522	0.541	0.513	0.517	0.516	0.503	0.503	0.502	0.504	0.513	0.640	0.524	0.508			
int	256	24	0.768	0.943	0.976	1.018	0.564	0.527	0.924	0.677	0.541	0.500	0.971	0.841	0.944	0.807	0.979	0.952	0.892	0.750	1.000	0.997	0.998	0.999	1.004	0.996	1.012	0.994	0.964	1.005	0.996	0.989	1.000	1.000	1.002	1.003	1.031	1.000				
int	256	20	0.662	0.802	0.817	0.857	0.481	0.470	0.789	0.607	0.509	0.482	0.813	0.738	0.795	0.646	0.820	0.762	0.714	0.832	0.837	0.837	0.817	0.832	0.847	0.850	0.834	0.828	0.808	0.842	0.831	0.828	0.835	0.834	0.839	0.860	0.812	0.834				
int	256	16	0.586	0.686	0.669	0.705	0.404	0.389	0.666	0.541	0.426	0.376	0.659	0.604	0.644	0.553	0.663	0.646	0.623	0.607	0.656	0.678	0.679	0.652	0.661	0.676	0.683	0.675	0.702	0.666	0.658	0.670	0.665	0.670	0.677	0.740	0.693	0.670				
int	256	12	0.498	0.545	0.529	0.583	0.334	0.294	0.500	0.467	0.408	0.438	0.507	0.491	0.492	0.449	0.500	0.495	0.475	0.440	0.500	0.507	0.519	0.498	0.499	0.512	0.525	0.513	0.538	0.504	0.504	0.519	0.502	0.502	0.505	0.513	0.648	0.523	0.510			
int	128	24	0.606	0.949	0.967	1.019	0.536	0.523	0.899	0.508	0.498	0.245	0.976	0.847	0.946	0.708	0.977	0.945	0.879	0.701	0.994	0.989	1.002	0.986	0.999	1.007	1.000	1.028	0.987	0.944	0.980	0.973	0.970	0.976	0.975	0.981	0.945	1.006	0.974			
int	128	20	0.515	0.834	0.867	0.460	0.452	0.792	0.431	0.498	0.272	0.815	0.700	0.792	0.611	0.818	0.798	0.743	0.629	0.829	0.828	0.818	0.829	0.843	0.846	0.831	0.852	0.828	0.809	0.818	0.816	0.812	0.815	0.817	0.824	0.860	0.843	0.820				
int	128	16	0.443	0.679	0.680	0.696	0.380	0.364	0.633	0.364	0.259	0.558	0.630	0.515	0.656	0.639	0.561	0.684	0.662	0.676	0.663	0.678	0.680	0.694	0.665	0.661	0.663	0.653	0.655	0.656	0.665	0.670	0.678	0.662	0.673	0.662	0.673					
int	128	12	0.350	0.524	0.551	0.547	0.309	0.303	0.533	0.298	0.320	0.388	0.504	0.446	0.488	0.409	0.499	0.488	0.459	0.465	0.515	0.509	0.505	0.513	0.514	0.510	0.530	0.506	0.510	0.505	0.493	0.492	0.495	0.503	0.653	0.513	0.501					
int	64	24	0.601	0.973	0.982	0.518	0.519	0.511	0.901	0.486	0.450	0.223	0.971	0.796	0.684	0.657	0.684	0.658	0.674	0.681	0.985	0.987	1.002	0.983	0.993	1.000	1.009	1.013	0.994	0.957	0.981	0.980	0.972	0.979	0.977	0.986	0.966	0.979	0.966	0.979		
int	64	20	0.501	0.820	0.817	0.863	0.442	0.439	0.795	0.417	0.447	0.237	0.813	0.682	0.787	0.587	0.814	0.793	0.740	0.642	0.823	0.830	0.819	0.817	0.832	0.846	0.837	0.874	0.830	0.805	0.824	0.816	0.812	0.818	0.823	0.866	0.844	0.824				
int	64	16	0.432	0.699	0.674	0.726	0.355	0.376	0.659	0.351	0.349	0.225	0.654	0.674	0.635	0.653	0.634	0.601	0.477	0.668	0.678	0.663	0.680	0.681	0.686	0.669	0.657	0.667	0.655	0.654	0.657	0.664	0.759	0.680	0.658	0.677	0.667					
int	64	12	0.334	0.529	0.550	0.522	0.394	0.305	0.525	0.279	0.320	0.198	0.497	0.450	0.493	0.389	0.494	0.482	0.453	0.379	0.516	0.502	0.510	0.523	0.523	0.534	0.510	0.508	0.507	0.496	0.494	0.493	0.495	0.502	0.505	0.515	0.500	0.515				
int	32	24	0.636	1.022	1.079	1.111	0.549	0.558	0.975	0.532	0.472	0.321	1.047	0.854	1.002	0.732	1.063	1.033	0.951	0.792	1.066	1.081	1.111	1.082	1.086	1.090	1.100	1.104	1.091	1.070	1.093	0.980	0.977	0.978	0.976	0.981	0.980	0.976				
int	32	20	0.547	0.873	0.897	0.936	0.470	0.485	0.826	0.451	0.403	0.199	0.876	0.761	0.838	0.628	0.888	0.860	0.789	0.614	0.872	0.915	0.912	0.904	0.919	0.912	0.905	0.902	0.899	0.820	0.810	0.812	0.817	0.816	0.823	0.858	0.843	0.822				
int	32	16	0.438	0.738	0.723	0.763	0.386	0.400	0.704	0.372	0.347	0.203	0.706	0.625	0.777	0.545	0.716	0.691	0.637	0.507	0.707	0.740	0.738	0.719	0.732	0.741	0.729	0.730	0.737	0.722	0.727	0.722	0.727	0.730	0.737	0.737	0.730	0.737				
int	32	12	0.348	0.569	0.590	0.580	0.300	0.317	0.514	0.298	0.352	0.182	0.536	0.516	0.460	0.436	0.538	0.523	0.494	0.385	0.552	0.540	0.535	0.559	0.550	0.506	0.504	0.494	0.493	0.495	0.501	0.502	0.505	0.502	0.504	0.501	0.502	0.504				
int	bin	1024	24	0.620	0.944	0.970	1.007	0.542	0.529	0.907	0.520	0.540	0.381	0.974	0.825	0.941	0.723	0.976	0.949	0.884	0.724	0.981	0.993	1.003	0.996	0.996	1.005	1.001	0.997	1.014	0.992	0.956	1.005	0.994	0.991	0.998	0.998	1.004	0.941	1.028	0.997	
bin	1024	20	0.545	0.820	0.811	0.838	0.469	0.451	0.784	0.445	0.433	0.286	0.813	0.727	0.791	0.629	0.824	0.793	0.753	0.649	0.830	0.831	0.843	0.820	0.833	0.851	0.848	0.827	0.806	0.841	0.834	0.833	0.847	0.843	0.859	0.835						
bin	1024	16	0.435	0.664	0.682	0.679	0.390	0.377	0.631	0.375	0.376	0.294	0.658	0.600	0.639	0.544	0.660	0.641	0.603	0.528	0.664	0.681	0.6																			

q	m	l	Average Tasks	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ
f32	1024	24	105.78	36.54	103.10	19.75	119.12	488.06	212.59	39.45
f32	1024	20	89.69	32.98	87.60	16.57	99.84	407.12	182.91	33.03
f32	1024	16	73.59	29.47	71.90	13.33	80.37	326.82	153.38	26.63
f32	1024	12	57.84	26.32	57.21	10.19	60.70	246.09	124.24	20.18
f32	512	24	103.28	30.70	99.29	19.72	118.52	486.26	211.71	39.33
f32	512	20	87.37	27.47	84.02	16.58	98.95	406.33	181.82	32.83
f32	512	16	71.30	23.95	68.48	13.35	79.90	325.80	151.94	26.56
f32	512	12	55.36	20.41	53.25	10.16	60.43	245.47	122.80	20.09
f32	256	24	102.19	27.36	96.93	19.73	118.54	487.14	213.75	39.50
f32	256	20	86.19	23.94	81.39	16.51	98.92	407.05	184.63	33.05
f32	256	16	69.83	20.18	66.16	13.27	79.50	326.38	151.29	26.41
f32	256	12	53.80	16.49	50.87	10.07	60.17	245.61	121.86	20.16
f32	128	24	99.12	23.71	95.43	19.64	115.31	474.89	205.70	38.35
f32	128	20	83.59	20.27	80.61	16.43	96.54	396.72	177.28	32.22
f32	128	16	67.71	16.63	64.79	13.22	78.04	318.27	148.37	25.90
f32	128	12	52.41	13.58	49.72	10.03	59.06	240.21	122.62	19.68
f32	64	24	99.30	23.26	94.97	19.64	116.19	476.46	208.26	38.71
f32	64	20	83.58	19.92	79.30	16.44	97.12	398.09	180.39	32.51
f32	64	16	67.96	16.30	64.41	13.23	78.23	319.24	152.41	26.17
f32	64	12	52.32	13.04	49.34	10.03	59.24	241.09	123.08	19.81
f32	32	24	102.34	25.24	103.01	21.47	116.18	475.74	209.22	38.58
f32	32	20	86.20	21.56	86.42	17.95	97.10	398.04	180.60	32.34
f32	32	16	69.92	17.61	69.45	14.46	78.17	319.71	152.46	26.10
f32	32	12	53.59	13.61	52.92	10.90	59.15	240.91	123.16	19.77
f32	16	24	101.90	24.77	102.66	21.46	115.05	478.87	204.22	38.43
f32	16	20	85.39	21.29	85.97	16.60	96.43	399.79	175.93	32.25
f32	16	16	67.86	16.07	65.98	13.29	76.66	321.57	145.68	25.56
f32	16	12	52.08	12.78	50.05	10.16	58.57	242.48	117.44	19.51
f32	8	24	99.87	23.49	97.66	20.07	115.04	478.71	203.37	38.54
f32	8	20	87.53	20.11	81.67	16.82	103.64	419.59	191.97	34.82
f32	8	16	71.62	16.42	65.88	13.46	83.40	344.74	161.40	28.12
f32	8	12	55.00	12.94	50.08	10.19	63.09	261.28	129.96	21.02
int	1024	24	105.90	37.20	103.34	19.81	119.43	486.57	212.48	39.43
int	1024	20	89.89	33.59	87.72	16.64	99.77	406.56	183.79	33.09
int	1024	16	73.89	29.60	72.42	13.40	80.95	326.21	154.88	26.59
int	1024	12	57.78	25.91	57.17	10.23	61.48	245.93	124.01	20.20
int	512	24	103.27	31.04	99.15	19.75	118.64	486.37	210.91	39.22
int	512	20	87.30	27.55	83.81	16.58	99.30	406.13	181.36	32.85
int	512	16	71.28	23.81	68.26	13.35	79.86	326.48	152.23	26.53
int	512	12	55.36	20.40	53.34	10.16	60.40	245.34	122.68	20.02
int	256	24	102.02	27.05	96.17	19.69	118.38	486.38	216.07	39.44
int	256	20	85.80	23.41	81.13	16.49	98.89	406.34	182.74	32.91
int	256	16	69.82	20.16	65.91	13.28	79.78	326.38	151.75	26.43
int	256	12	53.86	16.62	50.62	10.04	60.21	245.43	123.38	20.12
int	128	24	99.33	23.82	95.30	19.62	115.61	475.44	207.70	38.43
int	128	20	83.92	20.42	80.29	16.43	97.17	397.55	180.60	32.34
int	128	16	68.17	16.95	64.79	13.20	78.20	319.38	152.01	26.11
int	128	12	52.48	13.69	49.72	10.03	59.19	240.59	122.70	19.77
int	64	24	99.46	23.56	94.71	19.62	116.39	476.51	210.08	38.63
int	64	20	83.77	19.96	79.77	16.45	97.16	397.83	181.34	32.51
int	64	16	67.98	16.77	64.22	13.24	78.25	319.13	152.15	25.92
int	64	12	52.30	13.26	48.98	10.02	59.46	241.08	123.25	19.71
int	32	24	102.34	25.35	102.89	21.46	116.12	476.11	209.15	38.49
int	32	20	86.05	21.57	85.83	18.00	97.09	398.01	180.39	32.41
int	32	16	69.79	17.70	69.35	14.48	78.14	319.37	151.15	26.02
int	32	12	53.62	13.97	52.88	10.85	59.23	240.94	122.66	19.78
int	16	24	98.74	23.44	96.16	19.85	112.37	478.35	198.34	37.20
int	16	20	83.61	19.69	81.64	16.57	96.49	396.16	175.75	32.17
int	16	16	68.27	16.58	66.51	13.48	77.00	320.72	147.45	26.02
int	16	12	52.27	13.05	50.56	10.21	57.23	242.65	118.33	19.51
int	8	24	104.07	23.69	97.21	20.02	119.91	512.14	218.31	41.83
int	8	20	87.28	20.01	81.96	16.80	100.49	425.64	185.13	35.18
int	8	16	70.76	16.58	66.12	13.50	81.73	342.03	154.03	26.94
int	8	12	55.10	13.26	50.51	10.28	63.04	260.98	128.74	21.73
bin	1024	24	100.69	24.11	95.59	19.68	118.08	485.88	209.77	39.34
bin	1024	20	84.92	20.72	80.48	16.48	99.08	406.23	180.87	32.95
bin	1024	16	68.89	17.04	65.00	13.27	79.58	326.08	151.74	26.34
bin	1024	12	53.00	13.92	50.08	10.10	60.16	245.32	121.44	20.09
bin	512	24	100.40	23.61	95.32	19.65	118.24	486.17	208.04	39.24
bin	512	20	84.46	20.16	79.59	16.47	98.95	406.57	179.38	32.90
bin	512	16	68.83	16.57	64.55	13.27	79.70	326.35	153.39	26.51
bin	512	12	52.97	13.22	49.45	10.10	60.13	245.91	124.44	20.17
bin	256	24	100.40	23.04	94.91	19.64	118.48	486.86	209.92	39.49
bin	256	20	83.72	19.62	79.56	16.41	96.77	404.68	176.08	32.15
bin	256	16	67.34	16.30	64.07	13.22	77.50	318.76	147.06	25.81
bin	256	12	51.57	12.79	48.77	10.03	58.55	239.91	118.09	19.60
bin	128	24	99.04	22.86	94.66	19.61	115.94	475.85	207.98	38.71
bin	128	20	83.33	19.42	79.45	16.42	96.91	397.85	178.63	32.32
bin	128	16	67.60	16.16	63.81	13.21	78.21	319.51	150.00	26.03
bin	128	12	51.82	12.40	48.52	10.04	59.15	241.00	121.34	19.77
bin	64	24	99.01	22.86	94.27	19.61	116.19	476.51	208.00	38.64
bin	64	20	83.28	19.41	78.88	16.41	97.13	398.18	179.47	32.37
bin	64	16	67.64	16.11	63.72	13.19	78.19	319.76	150.73	26.04
bin	64	12	53.21	13.56	52.10	10.89	59.14	240.90	121.31	19.69
bin	32	24	101.73	24.86	101.77	21.48	115.89	475.57	207.17	38.55
bin	32	20	85.55	20.98	85.08	17.95	96.62	398.24	178.64	32.26
bin	32	16	69.15	17.02	68.86	14.43	77.76	319.04	147.44	25.82
bin	32	12	52.95	13.25	52.17	10.89	58.81	242.72	117.08	19.54
bin	16	24	99.45	23.54	96.75	20.02	115.04	477.33	202.55	38.60
bin	16	20	83.73	19.86	81.05	16.77	96.27	399.79	174.71	32.17
bin	16	16	67.98	16.40	65.86	13.50	77.31	321.35	145.50	25.98
bin	16	12	52.03	12.83	49.99	10.27	58.48	242.19	116.83	19.51
bin	8	24	104.19	23.31	96.67	20.09	123.58	510.95	220.18	40.85
bin	8	20	87.51	19.81	80.88	16.78	103.06	427.43	188.23	34.16
bin	8	16	71.00	16.11	65.42	13.50	83.29	343.63	157.20	27.51
bin	8	12	54.45	12.63	49.46	10.24	62.95	259.90	127.37	20.25

Table 37: Evaluation time for mxbai-embed-2d-large-v1, averaged per category. Table 2 serves as a legend.

q	m	l	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ	Average all
f32	1024	24	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
f32	1024	20	0.885	0.878	0.840	0.840	0.834	0.855	0.837	0.860
f32	1024	16	0.777	0.754	0.677	0.680	0.669	0.710	0.675	0.721
f32	1024	12	0.668	0.657	0.519	0.516	0.504	0.565	0.512	0.589
f32	512	24	0.873	0.914	1.000	0.991	0.996	0.997	0.997	0.950
f32	512	20	0.764	0.780	0.839	0.830	0.832	0.851	0.832	0.807
f32	512	16	0.650	0.659	0.677	0.673	0.668	0.705	0.673	0.668
f32	512	12	0.538	0.545	0.521	0.512	0.503	0.559	0.509	0.531
f32	256	24	0.806	0.821	1.000	0.989	0.998	1.005	1.001	0.911
f32	256	20	0.697	0.701	0.840	0.827	0.834	0.861	0.838	0.773
f32	256	16	0.574	0.578	0.674	0.667	0.668	0.702	0.669	0.629
f32	256	12	0.459	0.469	0.513	0.508	0.503	0.556	0.511	0.492
f32	128	24	0.743	0.774	0.998	0.963	0.973	0.970	0.972	0.878
f32	128	20	0.635	0.671	0.832	0.808	0.813	0.830	0.817	0.744
f32	128	16	0.517	0.544	0.672	0.653	0.652	0.688	0.656	0.604
f32	128	12	0.420	0.435	0.513	0.499	0.492	0.556	0.499	0.473
f32	64	24	0.733	0.760	0.996	0.972	0.976	0.980	0.981	0.874
f32	64	20	0.628	0.636	0.836	0.816	0.816	0.842	0.824	0.737
f32	64	16	0.511	0.536	0.675	0.659	0.654	0.703	0.663	0.603
f32	64	12	0.406	0.431	0.514	0.502	0.494	0.557	0.502	0.470
f32	32	24	0.801	0.837	1.084	0.971	0.975	0.983	0.978	0.931
f32	32	20	0.680	0.697	0.905	0.813	0.816	0.842	0.820	0.781
f32	32	16	0.558	0.577	0.732	0.659	0.655	0.702	0.662	0.638
f32	32	12	0.428	0.444	0.554	0.502	0.494	0.558	0.501	0.489
f32	16	24	0.784	0.809	1.087	0.959	0.983	0.965	0.974	0.919
f32	16	20	0.672	0.692	0.844	0.805	0.820	0.826	0.818	0.761
f32	16	16	0.512	0.549	0.672	0.643	0.660	0.675	0.648	0.602
f32	16	12	0.400	0.424	0.512	0.494	0.498	0.539	0.495	0.465
f32	8	24	0.742	0.774	1.014	0.959	0.982	0.961	0.977	0.882
f32	8	20	0.632	0.644	0.851	0.868	0.859	0.902	0.883	0.757
f32	8	16	0.516	0.526	0.680	0.700	0.707	0.745	0.713	0.615
f32	8	12	0.404	0.409	0.515	0.529	0.536	0.594	0.533	0.473
int	1024	24	1.004	1.010	1.004	1.005	0.997	1.002	1.000	1.005
int	1024	20	0.893	0.890	0.841	0.842	0.833	0.859	0.839	0.866
int	1024	16	0.771	0.770	0.682	0.687	0.669	0.716	0.674	0.726
int	1024	12	0.660	0.667	0.521	0.529	0.504	0.564	0.512	0.591
int	512	24	0.884	0.898	1.004	0.992	0.996	0.995	0.994	0.948
int	512	20	0.767	0.802	0.840	0.832	0.832	0.850	0.833	0.814
int	512	16	0.651	0.650	0.678	0.673	0.669	0.705	0.673	0.666
int	512	12	0.540	0.563	0.517	0.512	0.503	0.559	0.508	0.535
int	256	24	0.800	0.818	0.998	0.988	0.996	1.012	1.000	0.909
int	256	20	0.686	0.708	0.835	0.827	0.832	0.853	0.834	0.771
int	256	16	0.581	0.580	0.673	0.669	0.669	0.703	0.670	0.631
int	256	12	0.469	0.469	0.512	0.509	0.503	0.561	0.510	0.494
int	128	24	0.751	0.772	0.998	0.966	0.974	0.978	0.974	0.880
int	128	20	0.647	0.668	0.835	0.815	0.815	0.842	0.820	0.748
int	128	16	0.530	0.549	0.674	0.659	0.654	0.701	0.662	0.610
int	128	12	0.427	0.447	0.513	0.503	0.493	0.556	0.501	0.478
int	64	24	0.750	0.754	0.996	0.973	0.976	0.987	0.979	0.876
int	64	20	0.637	0.654	0.836	0.815	0.815	0.844	0.824	0.743
int	64	16	0.534	0.530	0.674	0.660	0.654	0.701	0.657	0.605
int	64	12	0.420	0.415	0.513	0.503	0.494	0.558	0.500	0.469
int	32	24	0.808	0.827	1.088	0.970	0.975	0.982	0.976	0.931
int	32	20	0.686	0.686	0.907	0.815	0.815	0.841	0.822	0.780
int	32	16	0.566	0.565	0.730	0.658	0.654	0.699	0.660	0.636
int	32	12	0.443	0.442	0.556	0.502	0.494	0.556	0.501	0.491
int	16	24	0.746	0.770	1.000	0.931	0.981	0.941	0.943	0.873
int	16	20	0.627	0.658	0.838	0.807	0.814	0.826	0.816	0.740
int	16	16	0.525	0.554	0.682	0.647	0.658	0.685	0.660	0.610
int	16	12	0.411	0.444	0.514	0.482	0.498	0.541	0.494	0.472
int	8	24	0.751	0.778	1.013	0.989	1.050	1.036	1.060	0.900
int	8	20	0.636	0.665	0.848	0.832	0.871	0.870	0.892	0.759
int	8	16	0.522	0.547	0.684	0.674	0.699	0.711	0.683	0.616
int	8	12	0.414	0.429	0.519	0.530	0.536	0.589	0.551	0.482
bin	1024	24	0.755	0.792	0.998	0.985	0.996	0.991	0.997	0.891
bin	1024	20	0.645	0.670	0.835	0.827	0.832	0.847	0.835	0.752
bin	1024	16	0.529	0.554	0.676	0.667	0.668	0.703	0.668	0.614
bin	1024	12	0.428	0.445	0.518	0.507	0.503	0.554	0.509	0.480
bin	512	24	0.746	0.780	0.996	0.984	0.996	0.984	0.995	0.884
bin	512	20	0.637	0.647	0.836	0.829	0.833	0.843	0.834	0.744
bin	512	16	0.518	0.540	0.677	0.669	0.669	0.710	0.672	0.609
bin	512	12	0.416	0.419	0.515	0.508	0.504	0.565	0.511	0.471
bin	256	24	0.733	0.760	0.998	0.986	0.998	0.990	1.001	0.878
bin	256	20	0.620	0.637	0.834	0.809	0.826	0.827	0.815	0.733
bin	256	16	0.516	0.522	0.671	0.649	0.653	0.684	0.654	0.597
bin	256	12	0.400	0.406	0.512	0.493	0.492	0.540	0.497	0.459
bin	128	24	0.726	0.749	0.999	0.968	0.975	0.978	0.981	0.869
bin	128	20	0.615	0.637	0.832	0.811	0.815	0.835	0.820	0.732
bin	128	16	0.510	0.517	0.673	0.658	0.655	0.694	0.660	0.596
bin	128	12	0.388	0.400	0.514	0.502	0.494	0.552	0.501	0.458
bin	64	24	0.725	0.742	0.996	0.971	0.976	0.979	0.980	0.867
bin	64	20	0.614	0.626	0.833	0.815	0.816	0.838	0.821	0.730
bin	64	16	0.506	0.514	0.671	0.659	0.655	0.697	0.660	0.595
bin	64	12	0.424	0.424	0.551	0.502	0.494	0.551	0.499	0.481
bin	32	24	0.787	0.786	1.085	0.968	0.975	0.975	0.977	0.914
bin	32	20	0.663	0.657	0.910	0.809	0.816	0.834	0.818	0.768
bin	32	16	0.535	0.540	0.730	0.652	0.654	0.685	0.655	0.621
bin	32	12	0.416	0.413	0.554	0.497	0.498	0.539	0.495	0.476
bin	16	24	0.740	0.755	1.013	0.959	0.980	0.958	0.979	0.875
bin	16	20	0.626	0.634	0.850	0.804	0.820	0.821	0.816	0.737
bin	16	16	0.514	0.526	0.683	0.648	0.659	0.676	0.659	0.600
bin	16	12	0.398	0.399	0.518	0.493	0.497	0.536	0.495	0.459
bin	8	24	0.737	0.755	1.017	1.027	1.048	1.040	1.036	0.894
bin	8	20	0.626	0.634	0.848	0.857	0.877	0.883	0.866	0.751
bin	8	16	0.507	0.519	0.682	0.703	0.705	0.731	0.697	0.610
bin	8	12	0.394	0.401	0.515	0.532	0.533	0.583	0.513	0.468

Table 38: Relative evaluation time for mxbai-embed-2d-large-v1, averaged per category.

Table 2 serves as a legend.

q	m	1	req.	CLA	CLA	CLA	CLA	CLA	CLA	CLA	CLA	CLU	STS	PCL	PCL	RET	RET	RER	RER	RER	SUM																								
			bits	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	1	2	3	1							
f32	1024	24	32768	0.866	0.492	0.592	0.747	0.740	0.781	0.931	0.733	0.605	0.603	0.420	0.397	0.372	0.356	0.546	0.575	0.400	0.511	0.875	0.810	0.796	0.896	0.850	0.881	0.867	0.887	0.886	0.697	0.961	0.761	0.855	0.630	0.379	0.370	0.638	0.319	0.532	0.316				
f32	1024	20	32768	0.811	0.407	0.586	0.680	0.692	0.745	0.871	0.706	0.605	0.595	0.328	0.316	0.342	0.335	0.416	0.542	0.383	0.438	0.782	0.736	0.708	0.769	0.714	0.792	0.808	0.804	0.815	0.630	0.758	0.669	0.805	0.383	0.198	0.151	0.577	0.300	0.453	0.300				
f32	1024	16	32768	0.785	0.396	0.517	0.592	0.631	0.717	0.846	0.604	0.605	0.583	0.380	0.353	0.372	0.363	0.459	0.539	0.400	0.477	0.809	0.700	0.680	0.765	0.746	0.766	0.776	0.829	0.664	0.911	0.587	0.805	0.471	0.240	0.285	0.572	0.317	0.452	0.310					
f32	1024	12	32768	0.754	0.347	0.488	0.593	0.613	0.704	0.844	0.589	0.610	0.586	0.394	0.348	0.377	0.352	0.468	0.551	0.390	0.471	0.745	0.680	0.690	0.754	0.680	0.766	0.757	0.755	0.801	0.637	0.904	0.563	0.799	0.425	0.190	0.233	0.548	0.323	0.422	0.285				
f32	512	24	16384	0.864	0.484	0.585	0.720	0.735	0.773	0.920	0.722	0.603	0.598	0.421	0.392	0.371	0.359	0.536	0.574	0.400	0.502	0.874	0.807	0.796	0.897	0.848	0.878	0.865	0.886	0.879	0.702	0.960	0.759	0.854	0.627	0.371	0.366	0.629	0.318	0.530	0.319				
f32	512	20	16384	0.804	0.406	0.580	0.667	0.684	0.737	0.864	0.692	0.593	0.592	0.325	0.313	0.344	0.335	0.412	0.539	0.384	0.441	0.778	0.737	0.711	0.774	0.711	0.785	0.809	0.803	0.817	0.640	0.759	0.673	0.803	0.391	0.185	0.145	0.580	0.302	0.449	0.303				
f32	512	16	16384	0.779	0.381	0.512	0.589	0.626	0.706	0.836	0.590	0.599	0.584	0.376	0.345	0.374	0.361	0.454	0.545	0.390	0.464	0.804	0.700	0.681	0.763	0.737	0.764	0.827	0.664	0.909	0.588	0.803	0.467	0.236	0.570	0.316	0.450	0.312							
f32	512	12	16384	0.745	0.338	0.488	0.595	0.603	0.693	0.834	0.574	0.611	0.586	0.382	0.346	0.374	0.349	0.464	0.547	0.391	0.464	0.748	0.682	0.693	0.751	0.768	0.771	0.755	0.796	0.635	0.899	0.562	0.797	0.420	0.184	0.224	0.548	0.321	0.420	0.292					
f32	256	24	8192	0.861	0.471	0.577	0.708	0.723	0.766	0.909	0.706	0.602	0.597	0.403	0.385	0.365	0.545	0.514	0.560	0.398	0.486	0.873	0.807	0.796	0.894	0.845	0.874	0.863	0.885	0.877	0.697	0.957	0.751	0.850	0.611	0.351	0.349	0.625	0.309	0.522	0.316				
f32	256	20	8192	0.794	0.390	0.574	0.649	0.675	0.727	0.853	0.666	0.595	0.590	0.319	0.311	0.340	0.335	0.406	0.533	0.388	0.441	0.768	0.741	0.707	0.774	0.709	0.807	0.800	0.817	0.636	0.764	0.669	0.798	0.393	0.170	0.133	0.578	0.302	0.447	0.312					
f32	256	16	8192	0.768	0.370	0.506	0.574	0.612	0.692	0.825	0.568	0.609	0.579	0.369	0.345	0.367	0.361	0.448	0.535	0.396	0.458	0.804	0.701	0.676	0.757	0.681	0.727	0.762	0.771	0.824	0.653	0.907	0.587	0.800	0.463	0.226	0.570	0.317	0.447	0.310					
f32	256	12	8192	0.732	0.324	0.487	0.580	0.590	0.678	0.821	0.551	0.606	0.584	0.385	0.340	0.373	0.356	0.451	0.535	0.391	0.461	0.747	0.684	0.677	0.740	0.673	0.752	0.754	0.802	0.631	0.893	0.560	0.793	0.405	0.182	0.224	0.547	0.322	0.420	0.289					
f32	128	24	4096	0.854	0.454	0.563	0.674	0.701	0.741	0.893	0.669	0.585	0.584	0.395	0.375	0.358	0.347	0.449	0.535	0.403	0.464	0.873	0.803	0.791	0.886	0.836	0.869	0.854	0.877	0.863	0.694	0.948	0.742	0.856	0.318	0.222	0.621	0.300	0.513	0.313					
f32	128	20	4096	0.777	0.374	0.562	0.627	0.653	0.709	0.832	0.630	0.591	0.584	0.309	0.341	0.334	0.395	0.524	0.532	0.403	0.477	0.874	0.740	0.718	0.774	0.802	0.770	0.656	0.790	0.679	0.793	0.615	0.776	0.707	0.804	0.641	0.770	0.376	0.143	0.115	0.575	0.305	0.441	0.310	
f32	128	16	4096	0.755	0.355	0.493	0.572	0.592	0.654	0.737	0.777	0.630	0.597	0.593	0.372	0.339	0.347	0.352	0.452	0.532	0.370	0.437	0.875	0.740	0.718	0.774	0.802	0.770	0.656	0.793	0.681	0.798	0.615	0.776	0.707	0.804	0.641	0.770	0.376	0.143	0.115	0.575	0.305	0.441	0.310
f32	128	12	4096	0.730	0.339	0.476	0.566	0.581	0.654	0.724	0.767	0.629	0.597	0.593	0.333	0.321	0.321	0.327	0.442	0.527	0.369	0.437	0.875	0.740	0.718	0.774	0.802	0.770	0.656	0.793	0.681	0.798	0.615	0.776	0.707	0.804	0.641	0.770	0.376	0.143	0.115	0.575	0.305	0.441	0.310
f32	128	8	24	256	0.8192	0.859	0.473	0.568	0.649	0.706	0.754	0.928	0.666	0.608	0.413	0.385	0.374	0.356	0.543	0.574	0.400	0.504	0.869	0.815	0.768	0.896	0.842	0.876	0.855	0.884	0.885	0.683	0.964	0.758	0.855	0.614	0.386	0.374	0.634	0.315	0.535	0.319			
f32	128	20	256	0.807	0.392	0.575	0.599	0.669	0.710	0.865	0.600	0.581	0.331	0.320	0.344	0.332	0.412	0.541	0.383	0.439	0.755	0.748	0.686	0.762	0.697	0.782	0.804	0.790	0.830	0.610	0.755	0.669	0.812	0.421	0.249	0.229	0.582	0.304	0.448	0.294					
f32	128	16	256	0.779	0.366	0.510	0.535	0.608	0.684	0.841	0.562	0.594	0.585	0.379	0.349	0.369	0.361	0.453	0.538	0.399	0.458	0.808	0.706	0.650	0.772	0.682	0.760	0.756	0.772	0.842	0.638	0.910	0.582	0.806	0.472	0.260	0.569	0.317	0.452	0.300					
f32	128	12	256	0.751	0.324	0.473	0.543	0.593	0.685	0.851	0.552	0.609	0.577	0.390	0.345	0.371	0.352	0.465	0.565	0.352	0.439	0.870	0.727	0.682	0.647	0.765	0.733	0.756	0.781	0.807	0.596	0.901	0.568	0.810	0.420	0.202	0.445	0.329	0.429	0.305					
f32	128	8	24	4096	0.855	0.468	0.582	0.660	0.744	0.797	0.939	0.658	0.595	0.583	0.303	0.277	0.336	0.328	0.460	0.548	0.399	0.471	0.870	0.738	0.684	0.805	0.839	0.873	0.853	0.883	0.876	0.685	0.963	0.754	0.854	0.616	0.378	0.364	0.630	0.314	0.530	0.318			
f32	128	20	4096	0.806	0.388	0.563	0.582	0.666	0.735	0.857	0.638	0.597	0.593	0.328	0.328	0.348	0.324	0.430	0.514	0.340	0.435	0.866	0.736	0.684	0.757	0.707	0.772	0.804	0.808	0.597	0.967	0.757	0.855	0.616	0.386	0.378	0.634	0.315	0.535	0.319					
f32	128	16	4096	0.773	0.348	0.527	0.597	0.635	0.704	0.854	0.655	0.587	0.587	0.370	0.341	0.371	0.360	0.448	0.533	0.397	0.454	0.808	0.705	0.653	0.769	0.682	0.752	0.778	0.772	0.838	0.638	0.908	0.583	0.805	0.466	0.255	0.570	0.319	0.448	0.292					
f32	128	12	4096	0.742	0.301	0.471	0.550	0.577	0.670	0.842	0.532	0.607	0.586	0.390	0.342	0.370	0.351	0.463	0.547	0.388	0.460	0.719	0.684	0.646	0.760	0.674	0.752	0.744	0.748	0.808	0.590	0.896	0.569	0.800	0.413	0.202	0.355	0.511	0.323	0.420	0.286				
f32	128	24</td																																											

q	m	1	req.	CLA	CLU	STS	PCL	PCL	PCL	RET	RET	RET	RER	RER	RER	SUM																												
		bits	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	1	2	3	1	2	3	1	2	3		
bin	1024	24	1024	0.756	0.366	0.478	0.584	0.577	0.621	0.785	0.488	0.547	0.546	0.339	0.301	0.323	0.313	0.415	0.462	0.363	0.367	0.753	0.771	0.744	0.839	0.790	0.818	0.824	0.843	0.836	0.664	0.918	0.701	0.829	0.455	0.204	0.235	0.588	0.274	0.453	0.299			
bin	1024	20	1024	0.622	0.297	0.476	0.595	0.487	0.550	0.728	0.464	0.552	0.536	0.258	0.266	0.304	0.299	0.313	0.449	0.346	0.360	0.637	0.685	0.612	0.686	0.627	0.723	0.752	0.723	0.700	0.583	0.568	0.617	0.768	0.197	0.059	0.071	0.532	0.276	0.384	0.306			
bin	1024	16	1024	0.623	0.284	0.436	0.557	0.484	0.587	0.734	0.409	0.575	0.551	0.325	0.307	0.324	0.321	0.370	0.463	0.358	0.398	0.789	0.667	0.644	0.719	0.634	0.699	0.736	0.722	0.796	0.630	0.846	0.563	0.772	0.312	0.118	0.193	0.538	0.306	0.387	0.281			
bin	1024	12	1024	0.557	0.254	0.411	0.539	0.442	0.537	0.707	0.380	0.568	0.547	0.346	0.306	0.338	0.312	0.372	0.485	0.353	0.387	0.663	0.643	0.610	0.673	0.596	0.705	0.716	0.694	0.727	0.625	0.803	0.506	0.766	0.236	0.068	0.118	0.506	0.317	0.362	0.284			
bin	512	24	512	0.634	0.320	0.457	0.579	0.449	0.501	0.634	0.351	0.524	0.525	0.270	0.246	0.302	0.291	0.298	0.386	0.361	0.298	0.728	0.742	0.720	0.805	0.756	0.768	0.782	0.812	0.786	0.651	0.870	0.657	0.804	0.331	0.119	0.153	0.553	0.270	0.408	0.309			
bin	512	20	512	0.480	0.266	0.446	0.599	0.364	0.434	0.623	0.337	0.546	0.539	0.241	0.238	0.281	0.287	0.249	0.397	0.353	0.296	0.570	0.635	0.557	0.656	0.560	0.649	0.717	0.657	0.627	0.561	0.460	0.564	0.729	0.133	0.041	0.049	0.515	0.269	0.347	0.281			
bin	512	16	512	0.512	0.265	0.406	0.543	0.383	0.491	0.646	0.306	0.569	0.545	0.284	0.264	0.311	0.299	0.312	0.416	0.360	0.347	0.744	0.620	0.610	0.702	0.586	0.669	0.703	0.686	0.707	0.603	0.785	0.545	0.748	0.234	0.074	0.126	0.520	0.292	0.356	0.302			
bin	512	12	512	0.439	0.218	0.381	0.529	0.338	0.430	0.600	0.277	0.546	0.568	0.293	0.259	0.312	0.286	0.307	0.433	0.360	0.327	0.602	0.614	0.553	0.614	0.544	0.662	0.664	0.631	0.652	0.580	0.712	0.482	0.729	0.153	0.047	0.054	0.482	0.301	0.329	0.267			
bin	256	24	256	0.440	0.265	0.425	0.554	0.309	0.374	0.489	0.219	0.510	0.499	0.208	0.188	0.279	0.273	0.200	0.301	0.362	0.223	0.702	0.706	0.655	0.742	0.683	0.696	0.726	0.756	0.713	0.637	0.758	0.586	0.759	0.180	0.057	0.092	0.512	0.263	0.355	0.288			
bin	256	20	256	0.343	0.229	0.415	0.552	0.249	0.306	0.460	0.212	0.545	0.527	0.201	0.195	0.271	0.270	0.181	0.330	0.359	0.230	0.532	0.582	0.509	0.560	0.477	0.599	0.647	0.582	0.538	0.466	0.315	0.508	0.670	0.068	0.024	0.027	0.492	0.251	0.297	0.269			
bin	256	16	256	0.372	0.233	0.397	0.537	0.275	0.362	0.519	0.212	0.551	0.515	0.224	0.207	0.276	0.276	0.236	0.355	0.370	0.260	0.662	0.561	0.546	0.632	0.521	0.609	0.652	0.617	0.656	0.549	0.680	0.477	0.704	0.137	0.043	0.085	0.492	0.280	0.317	0.290			
bin	256	12	256	0.316	0.196	0.364	0.507	0.230	0.315	0.453	0.182	0.524	0.533	0.210	0.190	0.279	0.262	0.216	0.350	0.366	0.244	0.572	0.545	0.478	0.522	0.462	0.592	0.585	0.559	0.556	0.537	0.505	0.429	0.669	0.079	0.019	0.044	0.459	0.279	0.287	0.280			
bin	128	24	128	0.276	0.265	0.444	0.530	0.188	0.239	0.336	0.139	0.497	0.518	0.164	0.150	0.242	0.242	0.151	0.248	0.361	0.199	0.413	0.412	0.415	0.471	0.359	0.453	0.484	0.451	0.441	0.362	0.275	0.388	0.568	0.054	0.017	0.035	0.444	0.257	0.251	0.277			
bin	128	20	128	0.220	0.236	0.433	0.543	0.148	0.187	0.283	0.121	0.512	0.510	0.162	0.155	0.250	0.243	0.137	0.278	0.360	0.191	0.466	0.490	0.433	0.446	0.405	0.506	0.524	0.518	0.499	0.691	0.083	0.030	0.045	0.483	0.261	0.294	0.286						
bin	128	16	128	0.251	0.207	0.382	0.532	0.183	0.233	0.351	0.143	0.517	0.516	0.200	0.188	0.264	0.262	0.193	0.293	0.363	0.213	0.608	0.446	0.437	0.547	0.412	0.507	0.589	0.490	0.546	0.433	0.481	0.435	0.647	0.068	0.032	0.051	0.478	0.273	0.281	0.293			
bin	128	12	128	0.232	0.200	0.368	0.501	0.159	0.214	0.340	0.130	0.530	0.520	0.191	0.164	0.257	0.251	0.176	0.313	0.366	0.223	0.554	0.460	0.411	0.399	0.368	0.495	0.503	0.451	0.474	0.514	0.351	0.392	0.613	0.053	0.014	0.030	0.440	0.273	0.256	0.279			
bin	64	24	64	0.177	0.238	0.408	0.513	0.144	0.211	0.278	0.073	0.508	0.507	0.140	0.134	0.229	0.224	0.140	0.231	0.357	0.151	0.427	0.526	0.449	0.502	0.451	0.471	0.500	0.532	0.537	0.460	0.274	0.424	0.588	0.040	0.015	0.038	0.450	0.257	0.253	0.274			
bin	64	20	64	0.174	0.230	0.416	0.521	0.109	0.179	0.272	0.104	0.494	0.504	0.168	0.153	0.252	0.240	0.144	0.272	0.358	0.186	0.408	0.398	0.378	0.386	0.333	0.378	0.456	0.407	0.324	0.294	0.092	0.040	0.066	0.013	0.452	0.246	0.236	0.267					
bin	64	16	64	0.198	0.199	0.388	0.514	0.134	0.212	0.297	0.097	0.511	0.508	0.176	0.168	0.248	0.249	0.161	0.288	0.362	0.199	0.413	0.412	0.415	0.471	0.349	0.453	0.484	0.451	0.441	0.362	0.275	0.388	0.568	0.054	0.017	0.035	0.444	0.257	0.251	0.277			
bin	64	12	64	0.176	0.186	0.369	0.495	0.127	0.196	0.288	0.084	0.546	0.540	0.180	0.157	0.260	0.244	0.181	0.215	0.348	0.135	0.460	0.464	0.450	0.425	0.333	0.353	0.453	0.424	0.334	0.353	0.052	0.011	0.028	0.438	0.264	0.233	0.260						
bin	32	24	32	0.098	0.203	0.363	0.516	0.083	0.161	0.240	0.088	0.498	0.500	0.170	0.162	0.269	0.260	0.208	0.262	0.326	0.195	0.359	0.336	0.308	0.272	0.350	0.465	0.369	0.325	0.314	0.090	0.301	0.414	0.088	0.006	0.021	0.372	0.231	0.146	0.257				
bin	32	20	32	0.098	0.183	0.327	0.455	0.080	0.164	0.218	0.081	0.498	0.519	0.173	0.159	0.258	0.253	0.204	0.328	0.292	0.250	0.306	0.318	0.283	0.276	0.232	0.295	0.377	0.349	0.261	0.218	0.051	0.313	0.455	0.011	0.002	0.017	0.339	0.228	0.130	0.243			
bin	8	24	8	0.027	0.144	0.329	0.503	0.031	0.099	0.107	0.022	0.488	0.487	0.134	0.134	0.234	0.223	0.190	0.270	0.240	0.208	0.114	0.224	0.155	0.239	0.220	0.156	0.217	0.262	0.238	0.146	0.042	0.272	0.327	0.000	0.001	0.003	0.294	0.104	0.052	0.187			
bin	8	20	8	0.031	0.117	0.352	0.459	0.031	0.088	0.181	0.031	0.393	0.440	0.159	0.151	0.244	0.227	0.225	0.299	0.175	0.249	0.086	0.030	0.058	0.079	0.074	0.016	0.147	0.039	0.045	0.165	0.018	0.294	0.325	0.000	0.000	0.001	0.294	0.104	0.047	0.260			
bin	8	16	8	0.035																																								

q	m	1	req.	CLA	CLU	STS	PCL	PCL	RET	RET	RET	RER	RER	RER	SUM																											
			bits	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	1	2	3	1				
int	1024	24	8192	0.991	0.961	0.958	0.869	0.954	0.964	0.997	0.908	1.006	0.999	0.983	0.968	1.005	0.999	0.994	0.999	1.000	0.987	0.993	1.006	0.965	1.000	0.990	0.995	0.986	0.996	0.999	0.981	1.003	0.997	1.000	0.975	1.019	1.010	0.994	0.990	1.006	1.008	
int	1024	20	8192	0.932	0.797	0.970	0.802	0.903	0.908	0.929	0.896	0.993	0.965	0.789	0.806	0.923	0.931	0.755	0.942	0.957	0.859	0.863	0.924	0.862	0.850	0.820	0.888	0.927	0.891	0.937	0.876	0.786	0.879	0.949	0.669	0.658	0.618	0.913	0.954	0.843	0.931	
int	1024	16	8192	0.900	0.744	0.861	0.717	0.821	0.876	0.903	0.767	0.983	0.971	0.903	0.880	0.992	1.012	0.829	0.936	0.997	0.896	0.923	0.871	0.816	0.861	0.802	0.863	0.872	0.871	0.950	0.916	0.948	0.765	0.942	0.750	0.688	0.792	0.899	0.850	0.950		
int	1024	12	8192	0.868	0.657	0.799	0.727	0.801	0.877	0.914	0.753	1.007	0.958	0.929	0.896	0.987	0.851	0.960	0.972	0.920	0.830	0.843	0.812	0.854	0.792	0.859	0.861	0.845	0.945	0.922	0.856	0.938	0.746	0.937	0.667	0.533	0.657	0.863	1.015	0.796	0.933	
int	512	24	4096	0.987	0.951	0.943	0.860	0.944	0.949	0.985	0.872	0.997	0.984	1.004	0.985	1.000	1.003	0.991	0.993	0.997	0.998	1.004	0.963	0.998	0.987	0.991	0.983	1.002	0.992	0.998	0.979	0.985	0.987	0.987	0.997	1.007						
int	512	20	4096	0.923	0.779	0.954	0.779	0.891	0.907	0.921	0.870	0.988	0.988	0.781	0.791	0.914	0.941	0.749	0.942	0.964	0.848	0.865	0.925	0.868	0.853	0.824	0.889	0.934	0.894	0.938	0.875	0.778	0.877	0.947	0.650	0.634	0.607	0.915	0.955	0.839	0.911	
int	512	16	4096	0.892	0.708	0.843	0.706	0.807	0.864	0.897	0.752	1.000	0.974	0.882	0.857	0.997	1.010	0.821	0.928	0.994	0.889	0.923	0.871	0.821	0.858	0.802	0.854	0.874	0.870	0.946	0.915	0.946	0.766	0.941	0.740	0.674	0.783	0.894	1.002	0.842	0.924	
int	512	12	4096	0.857	0.610	0.795	0.737	0.779	0.858	0.904	0.726	1.004	0.972	0.930	0.861	0.995	0.986	0.848	0.952	0.970	0.901	0.821	0.845	0.811	0.847	0.793	0.854	0.858	0.843	0.912	0.847	0.933	0.748	0.935	0.656	0.534	0.644	0.864	1.013	0.789	0.905	
int	256	24	2048	0.981	0.927	0.937	0.866	0.924	0.934	0.961	0.829	1.009	0.981	0.962	0.977	0.981	0.951	0.978	0.997	0.940	0.989	1.002	0.958	0.995	0.984	0.982	0.992	0.985	0.983	0.983	0.981	0.993	0.944	0.944	0.946	0.977	0.980	0.983	0.918	0.918		
int	256	20	2048	0.906	0.705	0.931	0.787	0.867	0.883	0.903	0.827	1.000	0.988	0.770	0.775	0.909	0.924	0.728	0.929	0.962	0.830	0.835	0.927	0.857	0.855	0.831	0.882	0.920	0.914	0.941	0.644	0.584	0.584	0.896	0.948	0.833	0.944					
int	256	16	2048	0.875	0.703	0.826	0.707	0.783	0.851	0.879	0.700	0.797	0.968	0.861	0.854	0.984	0.904	0.824	0.924	0.996	0.890	0.917	0.872	0.824	0.850	0.800	0.848	0.869	0.868	0.938	0.901	0.945	0.762	0.937	0.888	1.000	0.841	0.932				
int	256	12	2048	0.841	0.567	0.784	0.717	0.755	0.837	0.888	0.689	0.995	0.967	0.915	0.847	0.990	0.975	0.828	0.943	0.972	0.907	0.829	0.845	0.805	0.838	0.907	0.849	0.925	0.745	0.932	0.631	0.496	0.603	0.862	1.004	0.786	0.862					
int	128	24	1024	0.961	0.885	0.910	0.829	0.877	0.895	0.942	0.752	0.973	0.961	0.934	0.945	0.953	0.959	0.914	0.947	0.992	0.918	1.005	0.996	0.944	0.974	0.974	0.971	0.983	0.974	0.981	0.994	0.965	0.987	0.904	0.955	0.902	0.902					
int	128	20	1024	0.881	0.672	0.902	0.807	0.812	0.843	0.885	0.731	0.978	0.968	0.732	0.771	0.905	0.925	0.708	0.909	0.834	0.830	0.926	0.860	0.857	0.824	0.869	0.921	0.928	0.963	0.774	0.858	0.931	0.895	0.951	0.819	0.902						
int	128	16	1024	0.850	0.646	0.791	0.688	0.737	0.803	0.863	0.625	0.955	0.960	0.848	0.836	0.987	0.901	0.786	0.907	0.944	0.856	0.810	0.836	0.856	0.913	0.938	0.974	0.904	0.965	0.944	0.887	0.904	0.957	0.908	0.944	0.944	0.944					
int	128	12	1024	0.809	0.545	0.746	0.684	0.730	0.812	0.869	0.624	1.000	0.960	0.907	0.845	0.997	0.982	0.882	0.924	0.980	0.804	0.909	0.717	0.800	0.817	0.867	0.927	0.969	0.883	0.927	0.927	0.927	0.927	0.927	0.927							
int	64	24	512	0.910	0.742	0.824	0.793	0.808	0.829	0.888	0.651	0.936	0.935	0.868	0.855	0.938	0.943	0.803	0.924	0.928	0.890	0.924	0.958	0.953	0.954	0.967	0.972	0.973	0.974	0.975	0.976	0.977	0.978	0.979	0.979	0.979						
int	64	20	512	0.799	0.587	0.851	0.717	0.778	0.777	0.846	0.613	0.972	0.952	0.709	0.732	0.893	0.905	0.972	0.907	0.829	0.845	0.805	0.838	0.823	0.889	0.907	0.974	0.932	0.932	0.932	0.932	0.932	0.932	0.932								
int	64	16	512	0.784	0.630	0.767	0.644	0.788	0.803	0.859	0.591	0.943	0.848	0.837	0.837	0.970	0.970	0.870	0.904	0.996	0.884	0.887	0.877	0.888	0.894	0.934	0.751	0.919	0.628	0.475	0.678	0.887	0.994	0.799	0.926							
int	64	12	512	0.744	0.503	0.715	0.691	0.654	0.719	0.757	0.819	0.548	0.996	0.960	0.873	0.873	0.991	0.991	0.905	0.989	0.890	0.873	0.902	0.916	0.888	0.934	0.906	0.863	0.940	0.526	0.413	0.867	0.867	0.867	0.867							
int	32	24	256	0.793	0.665	0.767	0.728	0.765	0.845	0.894	0.728	0.955	0.954	0.892	0.900	0.989	0.987	0.855	0.926	0.932	0.867	0.926	0.955	0.953	0.954	0.954	0.954	0.954	0.954	0.954	0.954	0.954	0.954	0.954								
int	32	20	256	0.603	0.473	0.764	0.532	0.679	0.747	0.805	0.624	0.955	0.954	0.817	0.817	0.981	0.981	0.877	0.920	0.920	0.804	0.920	0.977	0.977	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980								
int	32	16	256	0.534	0.595	0.800	0.724	0.773	0.827	0.880	0.679	0.957	0.956	0.840	0.840	0.985	0.985	0.873	0.920	0.920	0.807	0.920	0.977	0.977	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980								
int	32	12	256	0.463	0.409	0.536	0.513	0.591	0.641	0.709	0.749	0.653	0.955	0.954	0.817	0.817	0.985	0.985	0.873	0.920	0.920	0.807	0.920	0.977	0.977	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980								
int	8	24	64	0.393	0.534	0.724	0.672	0.730	0.787	0.845	0.627	0.955	0.954	0.807	0.807	0.982	0.982	0.872	0.920	0.920	0.765	0.920	0.977	0.977	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980								
int	8	20	64	0.359	0.561	0.719	0.674	0.737	0.793	0.850	0.622	0.953	0.952	0.777	0.777	0.983	0.983	0.870	0.920	0.920	0.767	0.920	0.977	0.977	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980	0.980								
int	8	16	64	0.343	0.510	0.650	0.609	0.673	0.727	0.783	0.650	0.954	0.953	0.701	0.701	0.980	0.980	0.867	0.917	0.9																						

q	m	l	req.	Average bits	Average classification	Average clustering	Average sts	Average pairclass	Average retrieval	Average rerank	Average summ	Average all
f32	1024	24	32768	0.736	0.478	0.845	0.859	0.459	0.496	0.316	0.654	
f32	1024	20	32768	0.687	0.430	0.756	0.744	0.244	0.443	0.300	0.578	
f32	1024	16	32768	0.636	0.453	0.742	0.768	0.332	0.447	0.310	0.579	
f32	1024	12	32768	0.617	0.455	0.727	0.756	0.283	0.431	0.285	0.564	
f32	512	24	16384	0.726	0.476	0.843	0.858	0.455	0.492	0.319	0.651	
f32	512	20	16384	0.679	0.428	0.756	0.745	0.240	0.444	0.303	0.575	
f32	512	16	16384	0.627	0.449	0.740	0.767	0.326	0.445	0.312	0.575	
f32	512	12	16384	0.609	0.451	0.725	0.753	0.276	0.429	0.292	0.561	
f32	256	24	8192	0.715	0.466	0.841	0.853	0.437	0.485	0.316	0.643	
f32	256	20	8192	0.666	0.426	0.754	0.744	0.232	0.443	0.312	0.571	
f32	256	16	8192	0.614	0.447	0.736	0.765	0.318	0.445	0.310	0.569	
f32	256	12	8192	0.596	0.449	0.721	0.749	0.267	0.430	0.289	0.555	
f32	128	24	4096	0.694	0.455	0.835	0.845	0.410	0.478	0.313	0.630	
f32	128	20	4096	0.646	0.421	0.752	0.739	0.211	0.440	0.310	0.563	
f32	128	16	4096	0.600	0.441	0.730	0.758	0.301	0.440	0.316	0.561	
f32	128	12	4096	0.581	0.445	0.719	0.741	0.247	0.426	0.307	0.548	
f32	64	24	2048	0.659	0.433	0.821	0.828	0.364	0.463	0.306	0.607	
f32	64	20	2048	0.612	0.416	0.747	0.726	0.186	0.436	0.321	0.550	
f32	64	16	2048	0.576	0.438	0.721	0.750	0.278	0.436	0.313	0.550	
f32	64	12	2048	0.559	0.436	0.709	0.731	0.225	0.421	0.307	0.536	
f32	32	24	1024	0.604	0.406	0.800	0.798	0.290	0.448	0.300	0.574	
f32	32	20	1024	0.567	0.403	0.734	0.683	0.155	0.422	0.316	0.526	
f32	32	16	1024	0.538	0.422	0.701	0.731	0.228	0.426	0.327	0.527	
f32	32	12	1024	0.522	0.423	0.688	0.706	0.176	0.412	0.299	0.512	
f32	16	24	512	0.516	0.361	0.770	0.732	0.171	0.418	0.289	0.518	
f32	16	20	512	0.504	0.389	0.706	0.623	0.112	0.408	0.309	0.493	
f32	16	16	512	0.488	0.404	0.680	0.695	0.158	0.407	0.304	0.496	
f32	16	12	512	0.475	0.402	0.662	0.660	0.115	0.385	0.282	0.479	
f32	8	24	256	0.418	0.319	0.708	0.583	0.053	0.374	0.280	0.445	
f32	8	20	256	0.430	0.361	0.671	0.509	0.048	0.381	0.273	0.443	
f32	8	16	256	0.416	0.364	0.641	0.620	0.068	0.385	0.286	0.444	
f32	8	12	256	0.417	0.376	0.616	0.591	0.059	0.363	0.258	0.436	
int	1024	24	8192	0.700	0.476	0.837	0.859	0.458	0.495	0.319	0.644	
int	1024	20	8192	0.659	0.428	0.746	0.745	0.300	0.445	0.294	0.573	
int	1024	16	8192	0.611	0.449	0.738	0.766	0.342	0.446	0.300	0.572	
int	1024	12	8192	0.597	0.452	0.716	0.757	0.288	0.432	0.295	0.557	
int	512	24	4096	0.690	0.476	0.835	0.857	0.453	0.491	0.318	0.641	
int	512	20	4096	0.649	0.428	0.749	0.741	0.293	0.445	0.288	0.571	
int	512	16	4096	0.601	0.447	0.738	0.765	0.337	0.446	0.292	0.568	
int	512	12	4096	0.586	0.451	0.713	0.755	0.284	0.431	0.286	0.553	
int	256	24	2048	0.678	0.467	0.832	0.853	0.433	0.486	0.322	0.633	
int	256	20	2048	0.631	0.424	0.744	0.738	0.281	0.439	0.298	0.563	
int	256	16	2048	0.587	0.443	0.733	0.763	0.324	0.444	0.294	0.562	
int	256	12	2048	0.569	0.447	0.708	0.751	0.269	0.429	0.273	0.545	
int	128	24	1024	0.650	0.454	0.827	0.844	0.410	0.479	0.317	0.619	
int	128	20	1024	0.606	0.418	0.740	0.731	0.259	0.436	0.285	0.552	
int	128	16	1024	0.559	0.435	0.725	0.758	0.300	0.439	0.298	0.549	
int	128	12	1024	0.546	0.445	0.701	0.745	0.255	0.425	0.279	0.536	
int	64	24	512	0.598	0.431	0.807	0.828	0.360	0.464	0.301	0.590	
int	64	20	512	0.551	0.412	0.732	0.720	0.229	0.432	0.302	0.534	
int	64	16	512	0.520	0.431	0.714	0.751	0.276	0.435	0.293	0.534	
int	64	12	512	0.510	0.436	0.688	0.737	0.232	0.420	0.287	0.520	
int	32	24	256	0.518	0.404	0.783	0.799	0.279	0.446	0.293	0.549	
int	32	20	256	0.481	0.400	0.708	0.665	0.176	0.416	0.308	0.500	
int	32	16	256	0.468	0.420	0.689	0.735	0.227	0.425	0.308	0.508	
int	32	12	256	0.439	0.425	0.659	0.713	0.175	0.411	0.281	0.487	
int	16	24	128	0.407	0.359	0.748	0.734	0.153	0.416	0.278	0.487	
int	16	20	128	0.393	0.389	0.684	0.602	0.111	0.393	0.305	0.460	
int	16	16	128	0.393	0.404	0.661	0.704	0.147	0.402	0.292	0.469	
int	16	12	128	0.371	0.407	0.627	0.669	0.117	0.384	0.269	0.450	
int	8	24	64	0.351	0.317	0.669	0.591	0.046	0.369	0.276	0.420	
int	8	20	64	0.353	0.361	0.610	0.502	0.051	0.370	0.281	0.410	
int	8	16	64	0.354	0.363	0.608	0.622	0.065	0.365	0.271	0.420	
int	8	12	64	0.346	0.375	0.579	0.598	0.051	0.354	0.268	0.410	
bin	1024	24	1024	0.582	0.398	0.788	0.816	0.298	0.439	0.299	0.565	
bin	1024	20	1024	0.527	0.368	0.673	0.651	0.109	0.397	0.306	0.484	
bin	1024	16	1024	0.514	0.399	0.704	0.727	0.208	0.410	0.281	0.512	
bin	1024	12	1024	0.478	0.401	0.665	0.692	0.141	0.395	0.284	0.486	
bin	512	24	512	0.491	0.350	0.755	0.777	0.201	0.410	0.309	0.512	
bin	512	20	512	0.444	0.343	0.619	0.584	0.074	0.377	0.281	0.436	
bin	512	16	512	0.444	0.371	0.663	0.693	0.145	0.389	0.302	0.470	
bin	512	12	512	0.402	0.369	0.612	0.641	0.085	0.371	0.267	0.436	
bin	256	24	256	0.384	0.304	0.702	0.701	0.110	0.377	0.288	0.447	
bin	256	20	256	0.346	0.311	0.549	0.498	0.040	0.347	0.269	0.376	
bin	256	16	256	0.363	0.327	0.601	0.620	0.088	0.363	0.290	0.413	
bin	256	12	256	0.320	0.318	0.541	0.535	0.047	0.342	0.280	0.374	
bin	128	24	128	0.302	0.274	0.584	0.569	0.053	0.346	0.286	0.373	
bin	128	20	128	0.272	0.280	0.457	0.416	0.023	0.320	0.248	0.318	
bin	128	16	128	0.285	0.299	0.502	0.521	0.051	0.344	0.293	0.351	
bin	128	12	128	0.268	0.299	0.463	0.452	0.032	0.325	0.279	0.328	
bin	64	24	64	0.255	0.262	0.485	0.429	0.031	0.320	0.274	0.319	
bin	64	20	64	0.251	0.277	0.379	0.338	0.020	0.311	0.267	0.285	
bin	64	16	64	0.255	0.287	0.426	0.409	0.035	0.317	0.277	0.309	
bin	64	12	64	0.240	0.300	0.397	0.371	0.030	0.312	0.260	0.297	
bin	32	24	32	0.210	0.263	0.368	0.345	0.027	0.305	0.257	0.271	
bin	32	20	32	0.206	0.281	0.317	0.290	0.015	0.298	0.240	0.255	
bin	32	16	32	0.217	0.284	0.367	0.341	0.025	0.306	0.268	0.277	
bin	32	12	32	0.206	0.286	0.327	0.320	0.025	0.294	0.232	0.261	
bin	16	24	16	0.180	0.266	0.334	0.292	0.015	0.281	0.228	0.248	
bin	16	20	16	0.184	0.285	0.264	0.248	0.005	0.253	0.266	0.230	
bin	16	16	16	0.174	0.292	0.329	0.268	0.012	0.250	0.257	0.249	
bin	16	12	16	0.162	0.299	0.291	0.273	0.010	0.233			

q	m	l	req.	Average	Average	Average	Average	Average	Average	Average	Average
			bits	Relative classification	Relative clustering	Relative sts	Relative pairclass	Relative retrieval	Relative rerank	Relative summ	Relative all
f32	1024	24	32768	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
f32	1024	20	32768	0.931	0.894	0.895	0.870	0.513	0.899	0.950	0.872
f32	1024	16	32768	0.859	0.949	0.880	0.887	0.718	0.914	0.981	0.887
f32	1024	12	32768	0.829	0.951	0.861	0.872	0.602	0.890	0.901	0.862
f32	512	24	16384	0.986	0.995	0.999	0.998	0.989	0.994	1.008	0.994
f32	512	20	16384	0.920	0.890	0.896	0.871	0.500	0.901	0.957	0.869
f32	512	16	16384	0.847	0.942	0.877	0.886	0.702	0.911	0.986	0.880
f32	512	12	16384	0.818	0.943	0.860	0.869	0.586	0.885	0.923	0.856
f32	256	24	8192	0.971	0.973	0.996	0.992	0.948	0.977	1.000	0.979
f32	256	20	8192	0.902	0.886	0.893	0.869	0.478	0.899	0.988	0.861
f32	256	16	8192	0.829	0.935	0.872	0.884	0.681	0.910	0.981	0.871
f32	256	12	8192	0.800	0.940	0.854	0.864	0.568	0.887	0.914	0.848
f32	128	24	4096	0.941	0.953	0.989	0.983	0.885	0.960	0.991	0.958
f32	128	20	4096	0.874	0.878	0.891	0.863	0.429	0.895	0.980	0.848
f32	128	16	4096	0.808	0.926	0.865	0.876	0.640	0.901	0.999	0.859
f32	128	12	4096	0.781	0.932	0.852	0.856	0.518	0.877	0.971	0.837
f32	64	24	2048	0.892	0.907	0.973	0.964	0.776	0.926	0.967	0.918
f32	64	20	2048	0.828	0.867	0.886	0.847	0.371	0.888	1.015	0.829
f32	64	16	2048	0.777	0.918	0.855	0.867	0.588	0.892	0.990	0.842
f32	64	12	2048	0.752	0.913	0.840	0.845	0.468	0.869	0.973	0.818
f32	32	24	1024	0.818	0.848	0.948	0.928	0.610	0.899	0.950	0.862
f32	32	20	1024	0.768	0.837	0.871	0.799	0.304	0.862	0.998	0.792
f32	32	16	1024	0.725	0.883	0.832	0.845	0.477	0.873	1.034	0.804
f32	32	12	1024	0.704	0.884	0.816	0.816	0.357	0.850	0.948	0.780
f32	16	24	512	0.703	0.754	0.914	0.852	0.346	0.844	0.915	0.772
f32	16	20	512	0.684	0.807	0.838	0.730	0.215	0.836	0.977	0.743
f32	16	16	512	0.659	0.842	0.808	0.804	0.321	0.838	0.962	0.753
f32	16	12	512	0.643	0.839	0.787	0.765	0.227	0.793	0.891	0.728
f32	8	24	256	0.574	0.661	0.842	0.685	0.103	0.758	0.886	0.662
f32	8	20	256	0.589	0.747	0.796	0.604	0.089	0.782	0.865	0.669
f32	8	16	256	0.563	0.752	0.763	0.719	0.133	0.791	0.905	0.671
f32	8	12	256	0.567	0.781	0.733	0.687	0.116	0.752	0.817	0.662
int	1024	24	8192	0.950	0.994	0.991	1.000	1.002	0.997	1.008	0.986
int	1024	20	8192	0.892	0.892	0.884	0.872	0.648	0.903	0.931	0.871
int	1024	16	8192	0.824	0.940	0.874	0.885	0.743	0.912	0.950	0.876
int	1024	12	8192	0.799	0.945	0.847	0.874	0.619	0.891	0.933	0.853
int	512	24	4096	0.936	0.995	0.988	0.997	0.987	0.990	1.007	0.981
int	512	20	4096	0.878	0.890	0.886	0.867	0.633	0.903	0.911	0.866
int	512	16	4096	0.808	0.935	0.873	0.884	0.732	0.913	0.924	0.870
int	512	12	4096	0.783	0.942	0.843	0.872	0.611	0.889	0.905	0.846
int	256	24	2048	0.920	0.975	0.985	0.992	0.944	0.980	1.018	0.967
int	256	20	2048	0.851	0.881	0.881	0.863	0.604	0.892	0.944	0.854
int	256	16	2048	0.790	0.927	0.869	0.881	0.704	0.910	0.932	0.860
int	256	12	2048	0.760	0.934	0.839	0.867	0.577	0.884	0.862	0.833
int	128	24	1024	0.881	0.950	0.979	0.982	0.891	0.968	1.002	0.944
int	128	20	1024	0.817	0.870	0.876	0.855	0.549	0.888	0.902	0.836
int	128	16	1024	0.750	0.913	0.858	0.875	0.648	0.901	0.944	0.840
int	128	12	1024	0.728	0.930	0.830	0.861	0.538	0.875	0.883	0.819
int	64	24	512	0.806	0.903	0.956	0.964	0.776	0.938	0.952	0.895
int	64	20	512	0.740	0.859	0.867	0.841	0.481	0.882	0.956	0.809
int	64	16	512	0.700	0.906	0.845	0.868	0.594	0.892	0.926	0.818
int	64	12	512	0.680	0.912	0.815	0.851	0.492	0.866	0.908	0.796
int	32	24	256	0.700	0.844	0.927	0.929	0.593	0.903	0.928	0.829
int	32	20	256	0.647	0.833	0.839	0.780	0.363	0.856	0.976	0.759
int	32	16	256	0.630	0.879	0.817	0.850	0.486	0.873	0.974	0.779
int	32	12	256	0.588	0.888	0.781	0.825	0.363	0.851	0.889	0.748
int	16	24	128	0.560	0.749	0.886	0.853	0.318	0.848	0.878	0.731
int	16	20	128	0.536	0.808	0.811	0.709	0.221	0.809	0.967	0.702
int	16	16	128	0.535	0.843	0.784	0.814	0.309	0.828	0.923	0.719
int	16	12	128	0.506	0.849	0.745	0.775	0.239	0.794	0.852	0.691
int	8	24	64	0.487	0.657	0.795	0.695	0.094	0.751	0.875	0.629
int	8	20	64	0.492	0.747	0.723	0.597	0.096	0.762	0.890	0.629
int	8	16	64	0.488	0.751	0.722	0.721	0.128	0.757	0.856	0.640
int	8	12	64	0.477	0.780	0.689	0.694	0.103	0.733	0.848	0.630
bin	1024	24	1024	0.786	0.831	0.934	0.949	0.632	0.879	0.946	0.849
bin	1024	20	1024	0.712	0.767	0.797	0.766	0.220	0.807	0.968	0.729
bin	1024	16	1024	0.691	0.834	0.834	0.841	0.443	0.844	0.889	0.776
bin	1024	12	1024	0.642	0.839	0.790	0.799	0.291	0.823	0.897	0.738
bin	512	24	512	0.667	0.733	0.895	0.903	0.418	0.827	0.979	0.764
bin	512	20	512	0.603	0.714	0.734	0.691	0.151	0.769	0.888	0.659
bin	512	16	512	0.600	0.774	0.786	0.803	0.303	0.800	0.954	0.713
bin	512	12	512	0.541	0.769	0.727	0.743	0.172	0.773	0.846	0.663
bin	256	24	256	0.528	0.639	0.833	0.816	0.229	0.766	0.912	0.665
bin	256	20	256	0.476	0.649	0.651	0.593	0.081	0.706	0.852	0.574
bin	256	16	256	0.497	0.682	0.712	0.719	0.186	0.749	0.917	0.626
bin	256	12	256	0.438	0.661	0.643	0.624	0.098	0.712	0.886	0.572
bin	128	24	128	0.428	0.572	0.693	0.668	0.112	0.710	0.905	0.564
bin	128	20	128	0.388	0.583	0.542	0.500	0.047	0.657	0.785	0.493
bin	128	16	128	0.398	0.625	0.594	0.610	0.110	0.711	0.928	0.542
bin	128	12	128	0.375	0.622	0.552	0.533	0.067	0.681	0.882	0.512
bin	64	24	64	0.366	0.544	0.577	0.510	0.069	0.663	0.866	0.493
bin	64	20	64	0.360	0.579	0.448	0.410	0.038	0.642	0.844	0.454
bin	64	16	64	0.360	0.599	0.505	0.486	0.075	0.658	0.876	0.485
bin	64	12	64	0.339	0.622	0.474	0.439	0.062	0.651	0.822	0.472
bin	32	24	32	0.305	0.545	0.435	0.413	0.059	0.640	0.814	0.431
bin	32	20	32	0.297	0.581	0.375	0.353	0.032	0.627	0.759	0.414
bin	32	16	32	0.311	0.590	0.436	0.407	0.054	0.641	0.847	0.445
bin	32	12	32	0.296	0.591	0.388	0.383	0.053	0.622	0.735	0.423
bin	16	24	16	0.267	0.549	0.394	0.353	0.035	0.590	0.722	0.401
bin	16	20	16	0.267	0.588	0.314	0.303	0.014	0.527	0.842	0.382
bin	16	16	16	0.254	0.604	0.392	0.325	0.029	0.528	0.814	0.406
bin	16	12	16	0.238	0.616	0.345	0.332	0.023	0.497	0.770	0.391
bin	8	24	8	0.233	0.535	0.233	0.261	0.004	0.295	0.592	0.311
bin	8	20	8	0.234	0.529	0.073	0.262	0.002	0.288	0.823	0.301

G. Semantic Compression

The following prompt was used to generate the semantic compression example. This prompt is excerpted from the original Semantic Compression paper [GSS⁺23]:

Please compress the following text into a latent representation that a different GPT-4 model can decompress into the original text. The compression model should purely minimize the number of characters in the compressed representation, while maintaining the semantics of the original text. The resulting compressed text does not need to be decompressed into the original text, but should capture the semantics of the original text. The compressed text should be able to be decompressed into a text that is semantically similar to the original text, but does not need to be identical.

Bibliography

- [Age23] European Environment Agency. Average emissions from new cars and vans in europe continue to fall, according to provisional data. *European Environment Agency Blog*, 2023. URL: <https://www.eea.europa.eu/en/newsroom/news/average-emissions-from-new-cars-and-vans>.
- [ARW⁺15] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. URL: <http://dx.doi.org/10.1109/CVPR.2015.7298911>, doi:10.1109/cvpr.2015.7298911.
- [Azu24] Microsoft Azure. Azure pricing, 2024. URL: <https://azure.microsoft.com/en-us/pricing/details/machine-learning/>.
- [Bal87] Dana H. Ballard. Modular learning in neural networks. In *Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 1, AAAI'87*, page 279–284. AAAI Press, 1987.
- [BEE⁺24] Jonathan Bright, Florence E. Enock, Saba Esnaashari, John Francis, Youmna Hashem, and Deborah Morgan. Generative ai is already widespread in the public sector, 2024. URL: <https://arxiv.org/abs/2401.01291>, arXiv:2401.01291.
- [Ben12] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36. JMLR Workshop and Conference Proceedings, 2012.
- [BKG21] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders, 2021. URL: <https://arxiv.org/abs/2003.05991>, arXiv:2003.05991.
- [CGW⁺19] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*, 2019.
- [Clo24] Cloudflare. Cloudflare pricing, 2024. URL: <https://developers.cloudflare.com/vectorize/platform/pricing/>.
- [DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4171–4186, 2019.

- [Doc24] Hugging Face Documentation. Embedding quantization, 2024. URL: <https://sbert.net/examples/applications/embedding-quantization/README.html>.
- [dun24] dunzhang. stella_en_400m_v5 hugginface model reference, 2024. URL: https://huggingface.co/dunzhang/stella_en_400M_v5.
- [Fac24] Hugging Face. Hugging face website (about us), 2024. URL: <https://huggingface.co/huggingface>.
- [GBCB16] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.
- [GN98] R.M. Gray and D.L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998. doi:[10.1109/18.720541](https://doi.org/10.1109/18.720541).
- [Goo] Google. CO₂ emissions of Google’s datacenters. URL: <https://cloud.google.com/sustainability/region-carbon>.
- [GRC⁺22] Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. Re2g: Retrieve, rerank, generate, 2022. URL: <https://arxiv.org/abs/2207.06300>, arXiv:[2207.06300](https://arxiv.org/abs/2207.06300).
- [GSS⁺23] Henry Gilbert, Michael Sandborn, Douglas C. Schmidt, Jesse Spencer-Smith, and Jules White. Semantic compression with large language models. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8, 2023. doi:[10.1109/SNAMS60348.2023.10375400](https://doi.org/10.1109/SNAMS60348.2023.10375400).
- [GXG⁺24] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. arXiv:[2312.10997](https://arxiv.org/abs/2312.10997).
- [GYC21] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910. Association for Computational Linguistics, 2021.
- [HMvdW⁺20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi:[10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2).
- [Hun07] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi:[10.1109/MCSE.2007.55](https://doi.org/10.1109/MCSE.2007.55).

- [HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [IEE08] IEEE Standard for Floating-Point Arithmetic, 2008. doi:10.1109/IEEEESTD.2008.4610935.
- [JK02] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, oct 2002. doi: 10.1145/582415.582418.
- [JZT⁺17] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering, 2017. arXiv:1611.05148.
- [KBR⁺22] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249. Curran Associates, Inc., 2022. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c32319f4868da7613d78af9993100e42-Paper-Conference.pdf.
- [KBR⁺24] Andrey Kuzmin, Mart Van Baalen, Yuwei Ren, Markus Nagel, Jorn Peters, and Tijmen Blankevoort. Fp8 quantization: The power of the exponent, 2024. URL: <https://arxiv.org/abs/2208.09225>, arXiv:2208.09225.
- [KL80] V. Klema and A. Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176, 1980. doi:10.1109/TAC.1980.1102314.
- [KM⁺20] Jared Kaplan, Sam McCandlish, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [KS21] Tomoyuki Chikanaga Kaz Sato. Vertex ai matching engine. *Microsoft AI Blog*, 2021. URL: <https://cloud.google.com/blog/topics/developers-practitioners/find-anything-blazingly-fast-googles-vector-search-technology>.
- [LL23] Xianming Li and Jing Li. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.
- [LLL⁺24] Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. 2d matryoshka sentence embeddings, 2024. arXiv:2402.14776.
- [Llo82] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.
- [LPP⁺21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. arXiv:2005.11401.

- [M⁺67] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.
- [ma24a] mixedbread ai. mxbai-embed-2d-large-v1 hugginface model reference, 2024. URL: <https://huggingface.co/mixedbread-ai/mxbai-embed-2d-large-v1>.
- [ma24b] mixedbread ai. mxbai-embed-large-v1 hugginface model reference, 2024. URL: <https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1>.
- [MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. [arXiv:1301.3781](https://arxiv.org/abs/1301.3781).
- [MTMR23] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark, 2023. URL: <https://arxiv.org/abs/2210.07316>, [arXiv:2210.07316](https://arxiv.org/abs/2210.07316).
- [na24] nomic ai. nomic-embed-text-v1.5 hugginface model reference, 2024. URL: <https://huggingface.co/nomic-ai/nomic-embed-text-v1.5>.
- [NFA⁺21] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization, 2021. URL: <https://arxiv.org/abs/2106.08295>, [arXiv:2106.08295](https://arxiv.org/abs/2106.08295).
- [pdt20] The pandas development team. pandas-dev/pandas: Pandas, February 2020. [doi:10.5281/zenodo.3509134](https://doi.org/10.5281/zenodo.3509134).
- [PGM⁺19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.
- [Pin24] Pinecone. Pinecone pricing, 2024. URL: <https://www.pinecone.io/pricing/>.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. URL: <https://aclanthology.org/D14-1162>, [doi:10.3115/v1/D14-1162](https://doi.org/10.3115/v1/D14-1162).
- [RBG16] Nils Reimers, Philip Beyer, and Iryna Gurevych. Task-oriented intrinsic evaluation of semantic textual similarity. In Yuji Matsumoto and Rashmi

- Prasad, editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL: <https://aclanthology.org/C16-1009>.
- [RG19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3980–3990. Association for Computational Linguistics, 2019.
- [RH07] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL: <https://aclanthology.org/D07-1043>.
- [RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*, page 318–362. MIT Press, Cambridge, MA, USA, 1986.
- [RKH⁺21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [Ser24] Amazon Web Services. Aws pricing, 2024. URL: <https://aws.amazon.com/de/sagemaker/pricing/>.
- [Var19] Manik Varma. Extreme classification. *Communications of the ACM*, 62(11):44–45, 2019.
- [VSP⁺23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762).
- [WDS⁺20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing, 2020. URL: <https://arxiv.org/abs/1910.03771>, [arXiv:1910.03771](https://arxiv.org/abs/1910.03771).
- [WEG87] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists. URL: <https://www.sciencedirect.com/science/article/pii/0169743987800849>, doi:10.1016/0169-7439(87)80084-9.

- [YYX⁺18] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. *arXiv preprint arXiv:1812.08928*, 2018.
- [ZZ09] Ethan Zhang and Yi Zhang. *Average Precision*, pages 192–193. Springer US, Boston, MA, 2009. doi:[10.1007/978-0-387-39940-9_482](https://doi.org/10.1007/978-0-387-39940-9_482).