BASF
We create chemistry

NEURAL
INFORMATION
PROCESSING
SYSTEMS

ChemTEB: Chemical Text Embedding Benchmark, an Overview of Embedding Models Performance & Efficiency on a Specific Domain

Ali Shiraee Kasmaee, Mohammad Khodadad, Mohammad Arshi Saloot, Nick Sherck, Stephen Dokas, Hamidreza Mahyar, Soheila Samiee

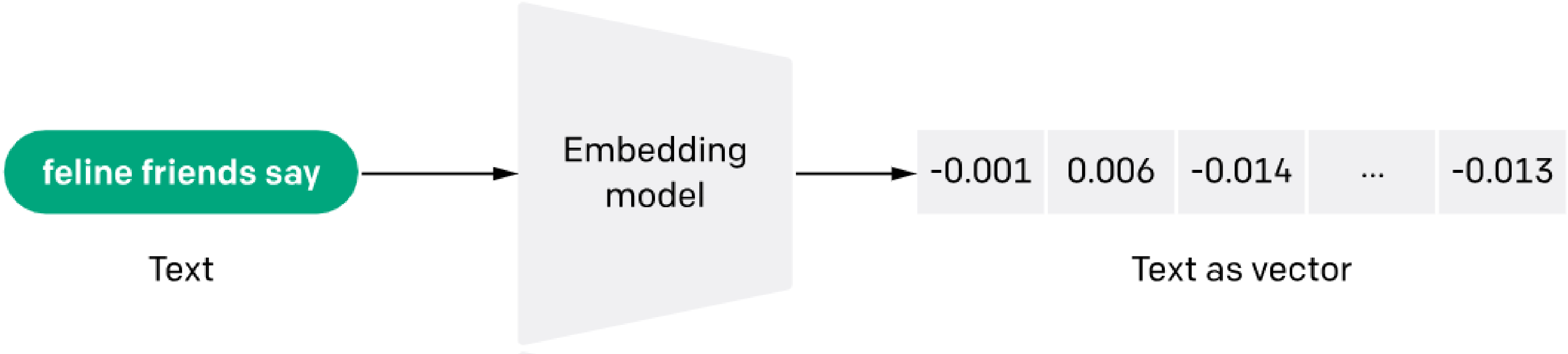
NeurIPS ENLSP 2024

INTRODUCTION & MOTIVATIONS

- Text Embedding Models
- Natural Language Processing Benchmarks
- Chemical Text Embedding Model

Text Embedding Models

Numerical Representation for Text



Natural Language Processing Benchmarks

Evaluation language understanding capabilities of the models

MTEB

Massive Text
Embedding Benchmark

SuperGLUE

GLUE

Beir

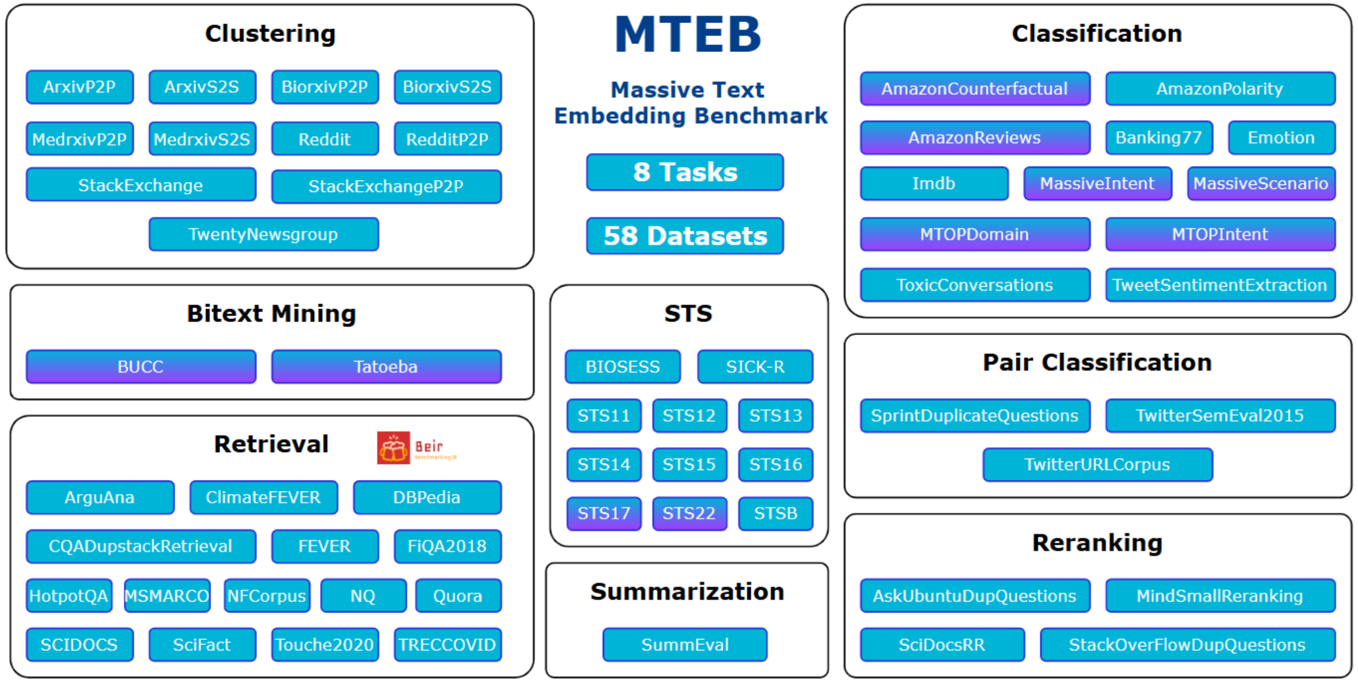
Benchmarking IR

Chemical Text Embedding Model

To address the need for domain specificity

General English

Chemistry



Clustering
ArxivP2P, ArxivS2S, BiorxivP2P, BiorxivS2S, MedrxivP2P, MedrxivS2S, Reddit, RedditP2P, StackExchange, StackExchangeP2P, TwentyNewsgroup

Bitext Mining
BUCC, Tatoeba

Retrieval
Beir
ArguAna, ClimateFEVER, DBPedia, CQADupstackRetrieval, FEVER, FiQA2018, HotpotQA, MSMARCO, NFCorpus, NQ, Quora, SCIDOCS, SciFact, Touche2020, TRECCOVID

MTEB
Massive Text Embedding Benchmark
8 Tasks
58 Datasets

STS
BIOSESS, SICK-R, STS11, STS12, STS13, STS14, STS15, STS16, STS17, STS22, STSB

Summarization
SummEval

Classification
AmazonCounterfactual, AmazonPolarity, AmazonReviews, Banking77, Emotion, Imdb, MassiveIntent, MassiveScenario, MTOPDomain, MTOPIntent, ToxicConversations, TweetSentimentExtraction

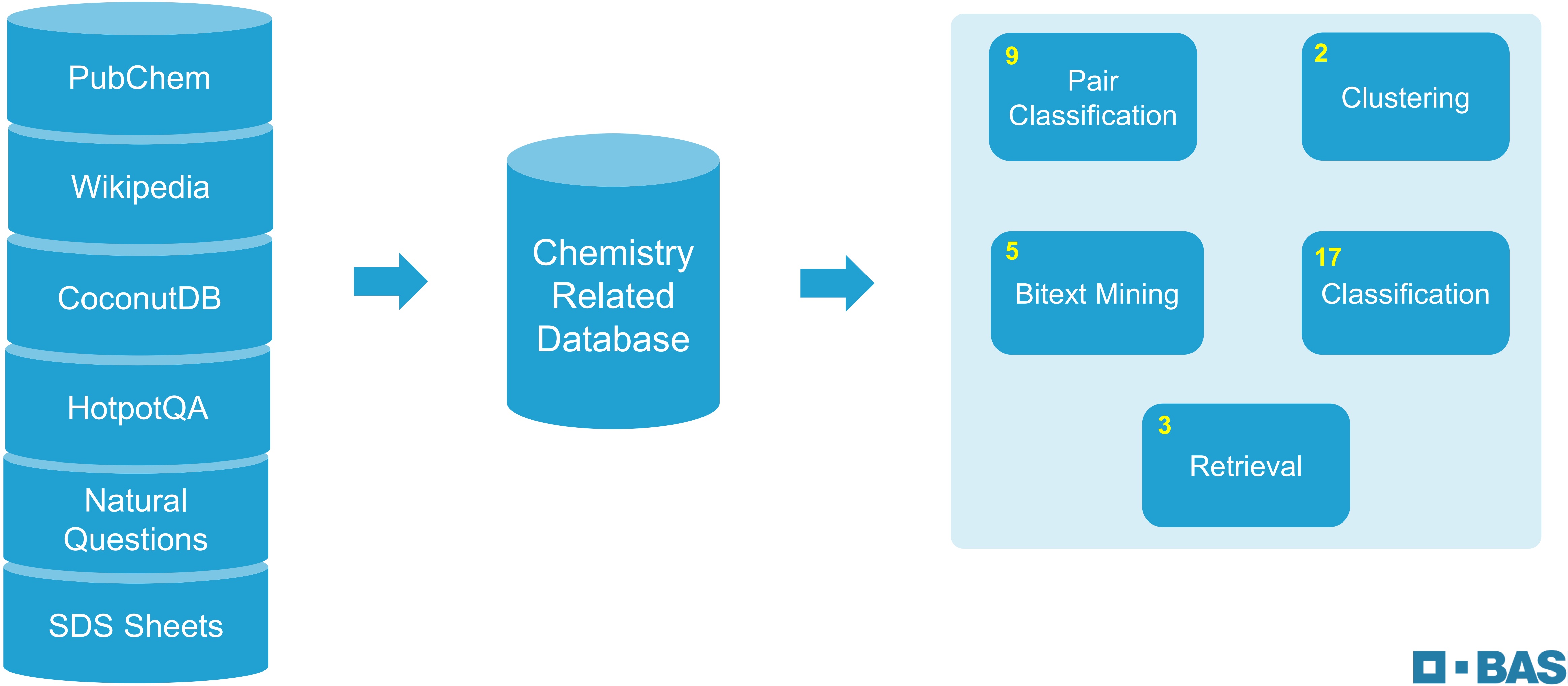
Pair Classification
SprintDuplicateQuestions, TwitterSemEval2015, TwitterURLCorpus

Reranking
AskUbuntuDupQuestions, MindSmallReranking, SciDocsRR, StackOverFlowDupQuestions

TASKS & MODELS

- Data Sources
- Tasks
- Evaluated Models

Data Sources and Tasks



PubChem

Wikipedia

CoconutDB

HotpotQA

Natural Questions

SDS Sheets

Chemistry Related Database

9 Pair Classification

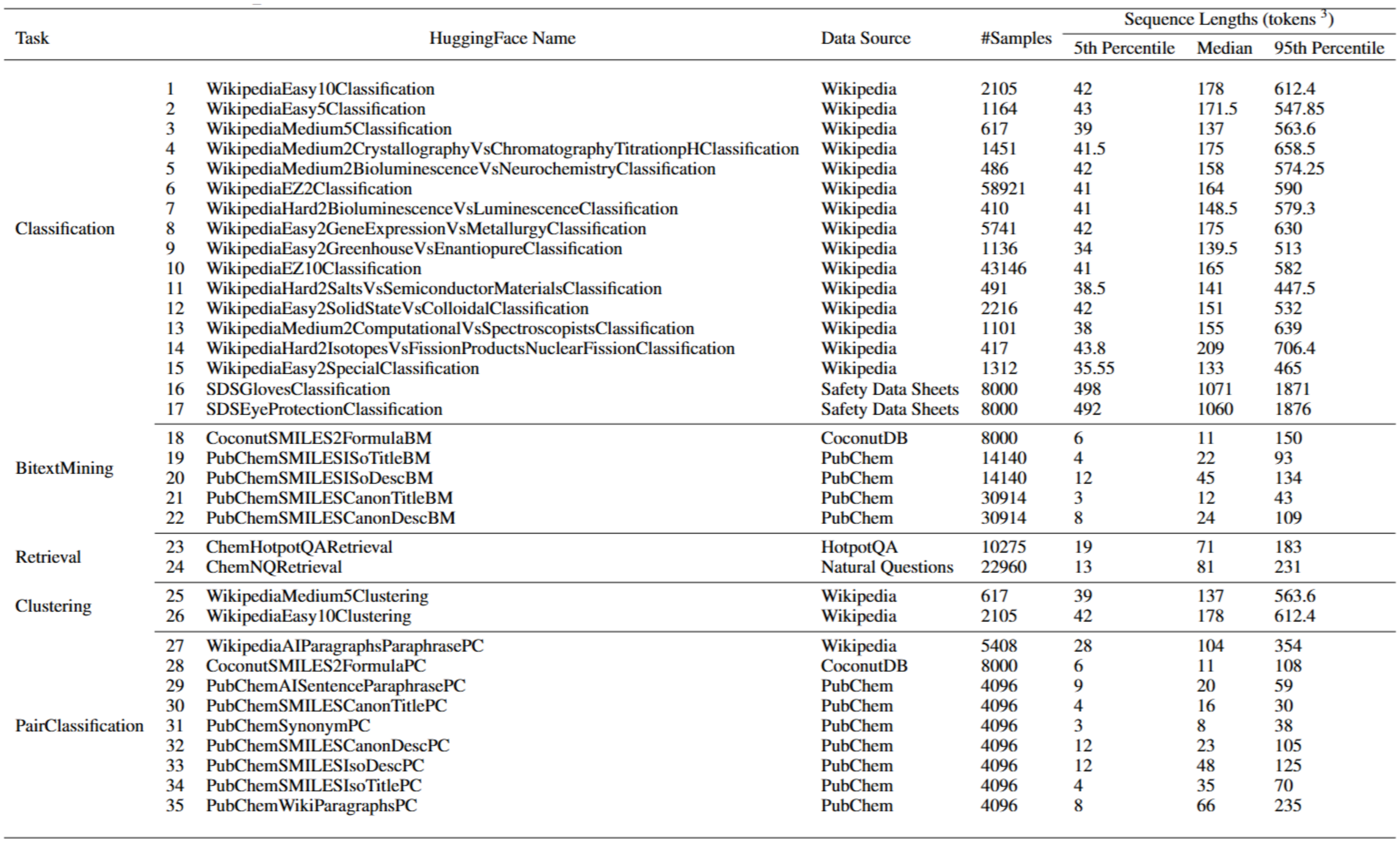
2 Clustering

5 Bitext Mining

17 Classification

3 Retrieval

Datasets & Their Statistics



Task		HuggingFace Name	Data Source	#Samples	Sequence Lengths (tokens [3])		
					5th Percentile	Median	95th Percentile
Classification	1	WikipediaEasy10Classification	Wikipedia	2105	42	178	612.4
	2	WikipediaEasy5Classification	Wikipedia	1164	43	171.5	547.85
	3	WikipediaMedium5Classification	Wikipedia	617	39	137	563.6
	4	WikipediaMedium2CrystallographyVsChromatographyTitrationpHClassification	Wikipedia	1451	41.5	175	658.5
	5	WikipediaMedium2BioluminescenceVsNeurochemistryClassification	Wikipedia	486	42	158	574.25
	6	WikipediaEZ2Classification	Wikipedia	58921	41	164	590
	7	WikipediaHard2BioluminescenceVsLuminescenceClassification	Wikipedia	410	41	148.5	579.3
	8	WikipediaEasy2GeneExpressionVsMetallurgyClassification	Wikipedia	5741	42	175	630
	9	WikipediaEasy2GreenhouseVsEnantiopureClassification	Wikipedia	1136	34	139.5	513
	10	WikipediaEZ10Classification	Wikipedia	43146	41	165	582
	11	WikipediaHard2SaltsVsSemiconductorMaterialsClassification	Wikipedia	491	38.5	141	447.5
	12	WikipediaEasy2SolidStateVsColloidalClassification	Wikipedia	2216	42	151	532
	13	WikipediaMedium2ComputationalVsSpectroscopistsClassification	Wikipedia	1101	38	155	639
	14	WikipediaHard2IsotopesVsFissionProductsNuclearFissionClassification	Wikipedia	417	43.8	209	706.4
	15	WikipediaEasy2SpecialClassification	Wikipedia	1312	35.55	133	465
	16	SDSGlovesClassification	Safety Data Sheets	8000	498	1071	1871
	17	SDSEyeProtectionClassification	Safety Data Sheets	8000	492	1060	1876
BitextMining	18	CoconutSMILES2FormulaBM	CoconutDB	8000	6	11	150
	19	PubChemSMILESISoTitleBM	PubChem	14140	4	22	93
	20	PubChemSMILESISoDescBM	PubChem	14140	12	45	134
	21	PubChemSMILESCanonTitleBM	PubChem	30914	3	12	43
	22	PubChemSMILESCanonDescBM	PubChem	30914	8	24	109
Retrieval	23	ChemHotpotQARetrieval	HotpotQA	10275	19	71	183
	24	ChemNQRetrieval	Natural Questions	22960	13	81	231
Clustering	25	WikipediaMedium5Clustering	Wikipedia	617	39	137	563.6
	26	WikipediaEasy10Clustering	Wikipedia	2105	42	178	612.4
PairClassification	27	WikipediaAIParagraphsParaphrasePC	Wikipedia	5408	28	104	354
	28	CoconutSMILES2FormulaPC	CoconutDB	8000	6	11	108
	29	PubChemAISentenceParaphrasePC	PubChem	4096	9	20	59
	30	PubChemSMILESCanonTitlePC	PubChem	4096	4	16	30
	31	PubChemSynonymPC	PubChem	4096	3	8	38
	32	PubChemSMILESCanonDescPC	PubChem	4096	12	23	105
	33	PubChemSMILESIsoDescPC	PubChem	4096	12	48	125
	34	PubChemSMILESIsoTitlePC	PubChem	4096	4	35	70
	35	PubChemWikiParagraphsPC	PubChem	4096	8	66	235

Evaluated Models

Open-Source & Proprietary Models

BERT

bert-base-uncased
chemical-bert-uncased
scibert_scivocab_uncased
nomic-bert-2048
Matscibert

Sentence Transformer

all-MiniLM-L6-v2
all-MiniLM-L12-v2
all-mpnet-base-v2
multi-qa-mpnet-base-dot-v1

BGE

bge-m3
bge-{small,base,large}-en
bge-{small,base,large}-en-v1.5

Nomic AI

nomic-embed-text-v1
nomic-embed-text-v1.5

E5

e5-{small,base,large}
e5-{small,base,large}
multilingual-e5-{small,base,large}

OpenAI

text-embedding-ada-002
text-embedding-3-small
text-embedding-3-large

Amazon

amazon.titan-embed-text-v1
amazon.titan-embed-text-v2:0

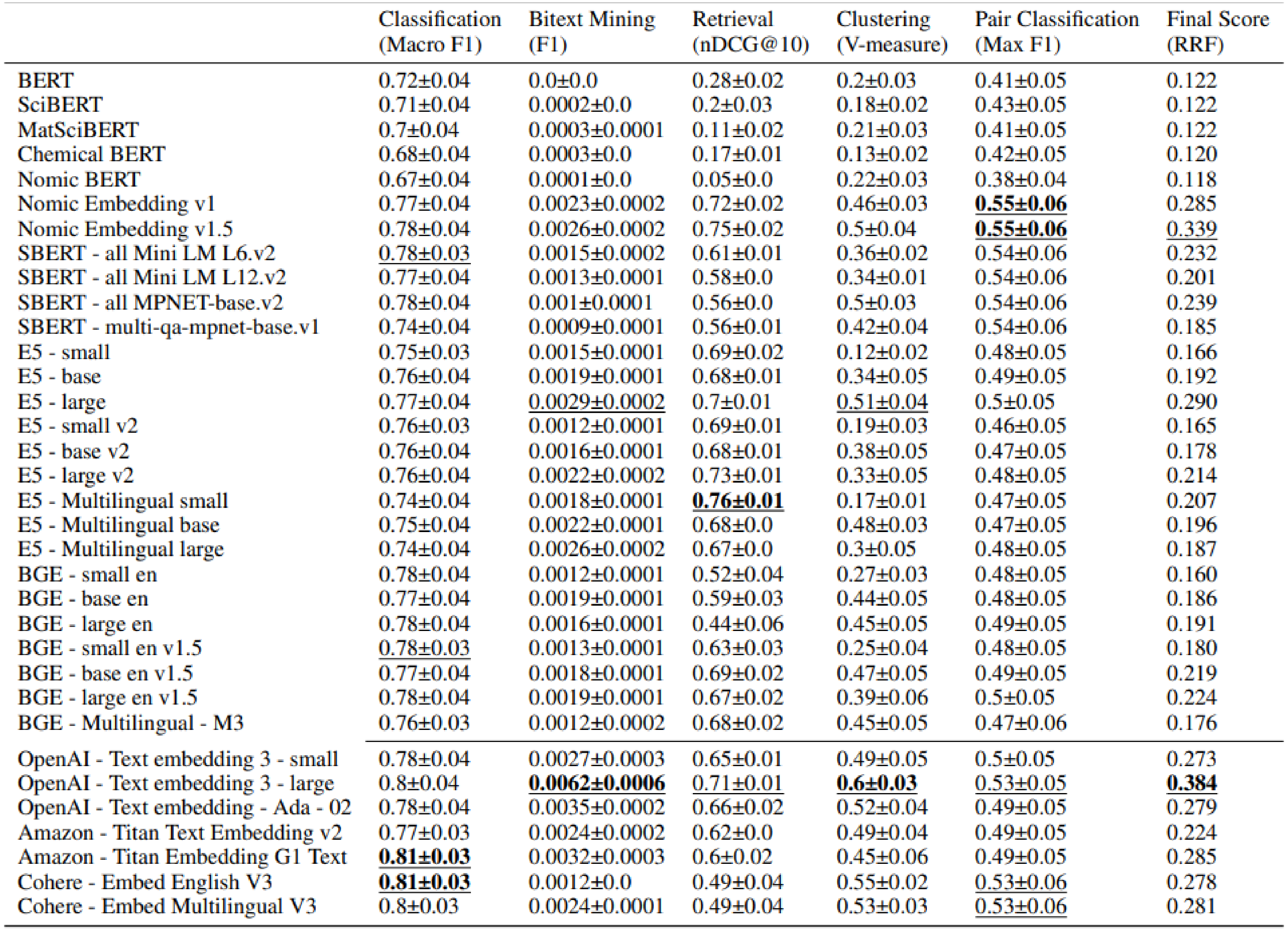
Cohere

cohere.embed-english-v3
cohere-embed-multilingual-v3

Results and Analysis

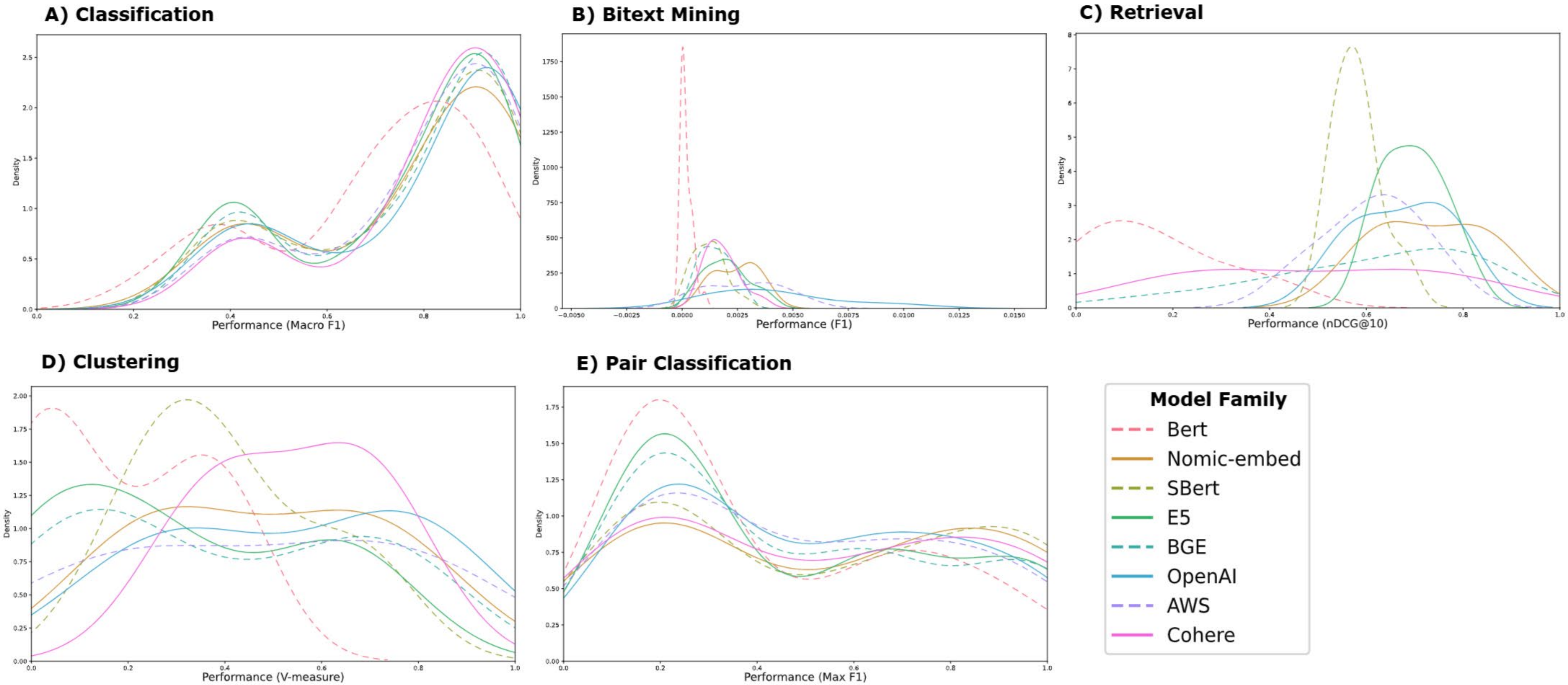
- Overall Model Performance
- Performance Across Different Tasks
- Efficiency Considerations

Overall Model Performance

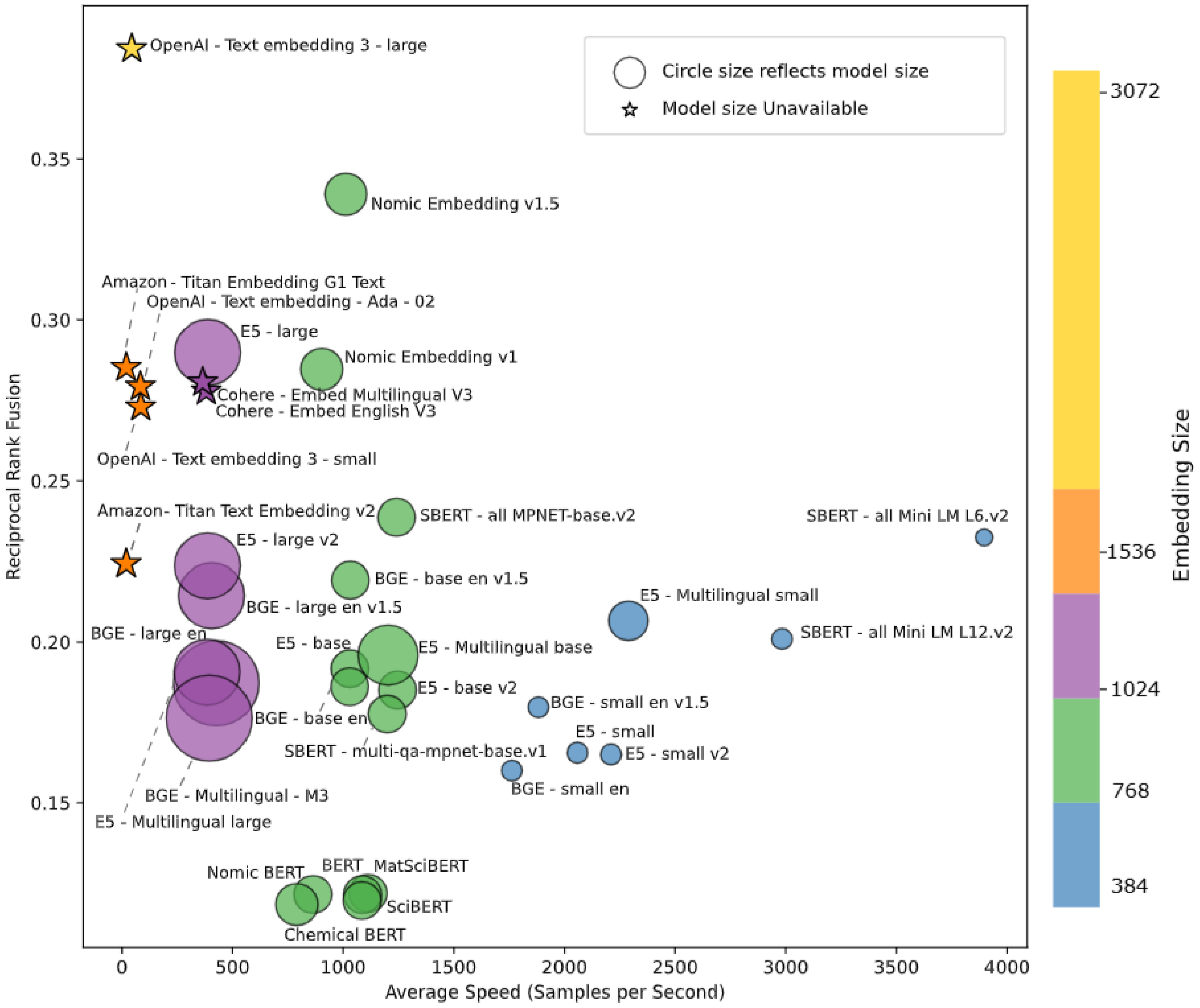


	Classification (Macro F1)	Bitext Mining (F1)	Retrieval (nDCG@10)	Clustering (V-measure)	Pair Classification (Max F1)	Final Score (RRF)
BERT	0.72±0.04	0.0±0.0	0.28±0.02	0.2±0.03	0.41±0.05	0.122
SciBERT	0.71±0.04	0.0002±0.0	0.2±0.03	0.18±0.02	0.43±0.05	0.122
MatSciBERT	0.7±0.04	0.0003±0.0001	0.11±0.02	0.21±0.03	0.41±0.05	0.122
Chemical BERT	0.68±0.04	0.0003±0.0	0.17±0.01	0.13±0.02	0.42±0.05	0.120
Nomic BERT	0.67±0.04	0.0001±0.0	0.05±0.0	0.22±0.03	0.38±0.04	0.118
Nomic Embedding v1	0.77±0.04	0.0023±0.0002	0.72±0.02	0.46±0.03	**0.55±0.06**	0.285
Nomic Embedding v1.5	0.78±0.04	0.0026±0.0002	0.75±0.02	0.5±0.04	**0.55±0.06**	0.339
SBERT - all Mini LM L6.v2	0.78±0.03	0.0015±0.0002	0.61±0.01	0.36±0.02	0.54±0.06	0.232
SBERT - all Mini LM L12.v2	0.77±0.04	0.0013±0.0001	0.58±0.0	0.34±0.01	0.54±0.06	0.201
SBERT - all MPNET-base.v2	0.78±0.04	0.001±0.0001	0.56±0.0	0.5±0.03	0.54±0.06	0.239
SBERT - multi-qa-mpnet-base.v1	0.74±0.04	0.0009±0.0001	0.56±0.01	0.42±0.04	0.54±0.06	0.185
E5 - small	0.75±0.03	0.0015±0.0001	0.69±0.02	0.12±0.02	0.48±0.05	0.166
E5 - base	0.76±0.04	0.0019±0.0001	0.68±0.01	0.34±0.05	0.49±0.05	0.192
E5 - large	0.77±0.04	0.0029±0.0002	0.7±0.01	0.51±0.04	0.5±0.05	0.290
E5 - small v2	0.76±0.03	0.0012±0.0001	0.69±0.01	0.19±0.03	0.46±0.05	0.165
E5 - base v2	0.76±0.04	0.0016±0.0001	0.68±0.01	0.38±0.05	0.47±0.05	0.178
E5 - large v2	0.76±0.04	0.0022±0.0002	0.73±0.01	0.33±0.05	0.48±0.05	0.214
E5 - Multilingual small	0.74±0.04	0.0018±0.0001	**0.76±0.01**	0.17±0.01	0.47±0.05	0.207
E5 - Multilingual base	0.75±0.04	0.0022±0.0001	0.68±0.0	0.48±0.03	0.47±0.05	0.196
E5 - Multilingual large	0.74±0.04	0.0026±0.0002	0.67±0.0	0.3±0.05	0.48±0.05	0.187
BGE - small en	0.78±0.04	0.0012±0.0001	0.52±0.04	0.27±0.03	0.48±0.05	0.160
BGE - base en	0.77±0.04	0.0019±0.0001	0.59±0.03	0.44±0.05	0.48±0.05	0.186
BGE - large en	0.78±0.04	0.0016±0.0001	0.44±0.06	0.45±0.05	0.49±0.05	0.191
BGE - small en v1.5	0.78±0.03	0.0013±0.0001	0.63±0.03	0.25±0.04	0.48±0.05	0.180
BGE - base en v1.5	0.77±0.04	0.0018±0.0001	0.69±0.02	0.47±0.05	0.49±0.05	0.219
BGE - large en v1.5	0.78±0.04	0.0019±0.0001	0.67±0.02	0.39±0.06	0.5±0.05	0.224
BGE - Multilingual - M3	0.76±0.03	0.0012±0.0002	0.68±0.02	0.45±0.05	0.47±0.06	0.176
OpenAI - Text embedding 3 - small	0.78±0.04	0.0027±0.0003	0.65±0.01	0.49±0.05	0.5±0.05	0.273
OpenAI - Text embedding 3 - large	0.8±0.04	**0.0062±0.0006**	0.71±0.01	**0.6±0.03**	0.53±0.05	**0.384**
OpenAI - Text embedding - Ada - 02	0.78±0.04	0.0035±0.0002	0.66±0.02	0.52±0.04	0.49±0.05	0.279
Amazon - Titan Text Embedding v2	0.77±0.03	0.0024±0.0002	0.62±0.0	0.49±0.04	0.49±0.05	0.224
Amazon - Titan Embedding G1 Text	**0.81±0.03**	0.0032±0.0003	0.6±0.02	0.45±0.06	0.49±0.05	0.285
Cohere - Embed English V3	**0.81±0.03**	0.0012±0.0	0.49±0.04	0.55±0.02	0.53±0.06	0.278
Cohere - Embed Multilingual V3	0.8±0.03	0.0024±0.0001	0.49±0.04	0.53±0.03	0.53±0.06	0.281

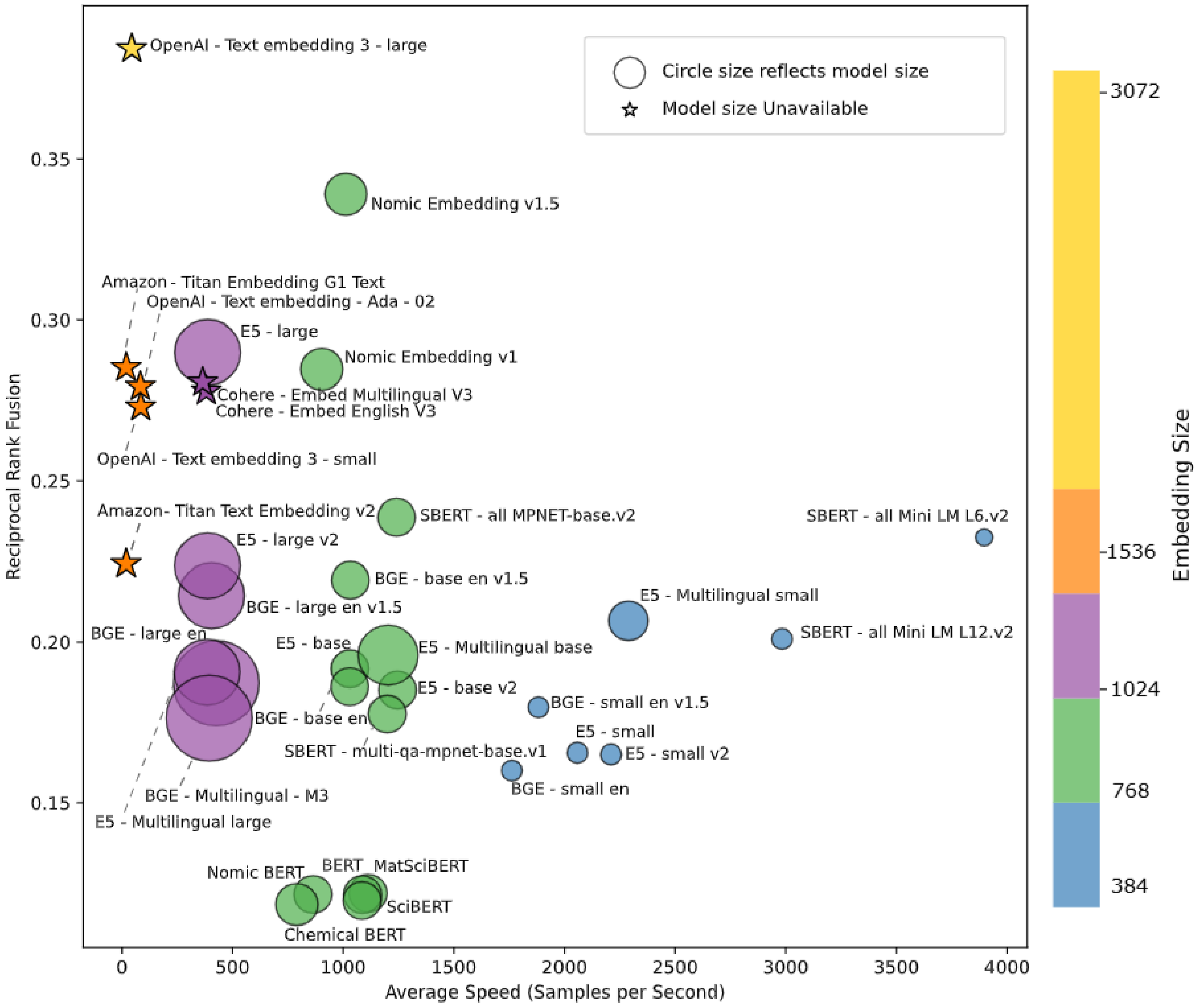
PERFORMANCE ACROSS TASKS



Models Efficiency

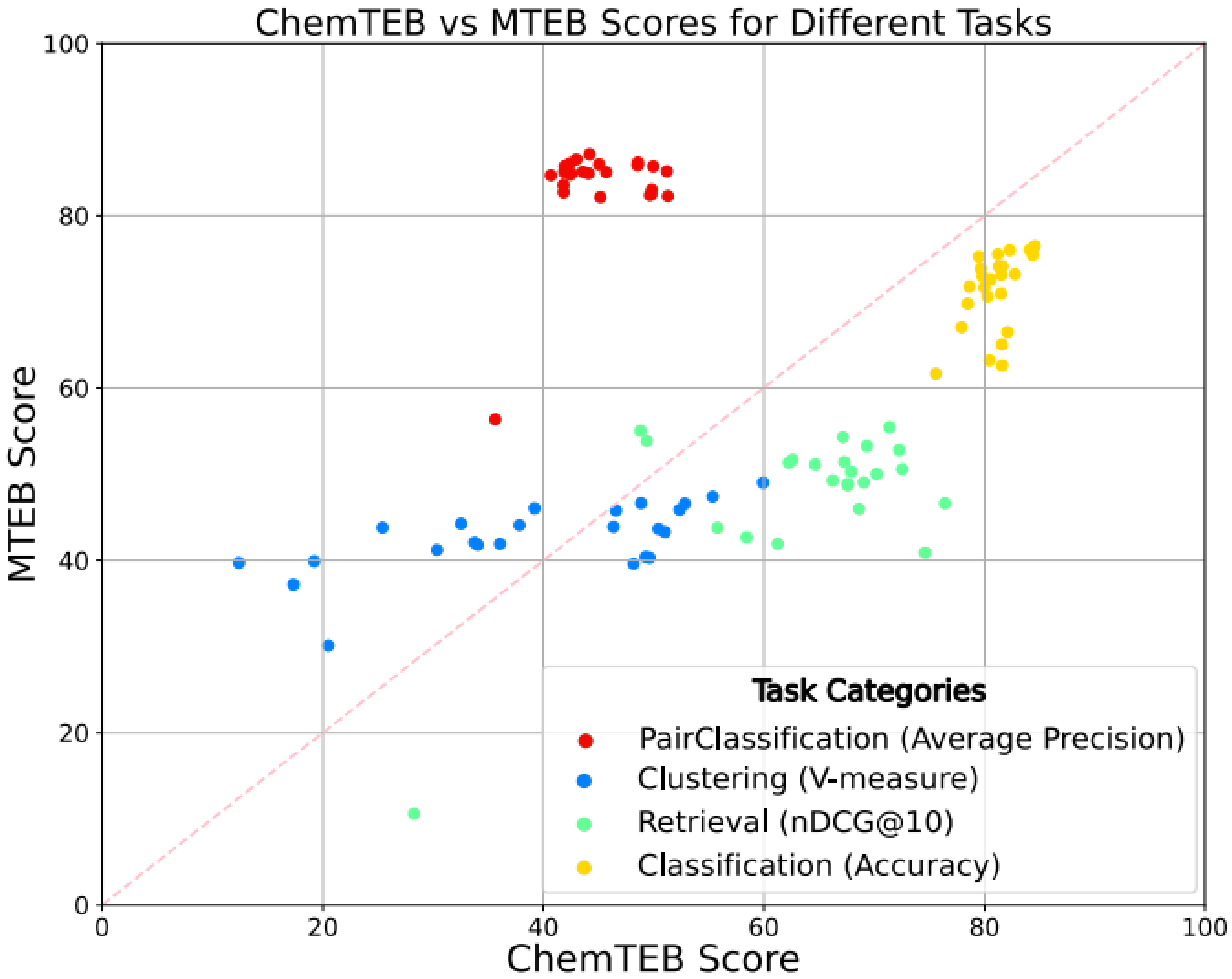


DOMAIN ADAPTATION



We create chemistry

COMPARISON WITH MTEB



Conclusion and Future Work

- CheMTEB Fills a critical gap in evaluating models on domain-specific tasks
- Emphasizes need for stronger, domain-adapted models with efficient architectures.
- Contrastive learning and architectural improvements are key to performance.

THANK YOU!

- Contact: shiraeea@mcmaster.ca

BASF

We create chemistry