MOGIC: METADATA-INFUSED ORACLE GUIDANCE FOR IMPROVED EXTREME CLASSIFICATION

ICML 2025

Suchith Chidananda Prabhu, Bhavyajeet Singh, Anshul Mittal, Siddarth Asokan, Shikhar Mohan, Deepak Saini, Yashoteja Prabhu, Lakshya Kuman, Jian jiao, Amit S, Niket Tandon, Manish Gupta, Sumeet Agarwal, Manik Varma

Microsoft

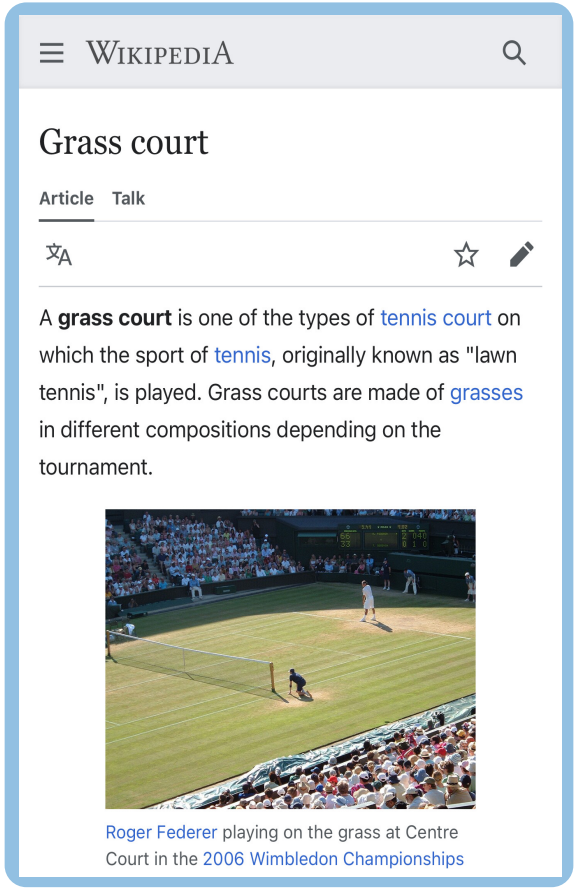
WHAT IS EXTREME MULTI-LABEL CLASSIFICATION (XC)?

It is the task of annotating a datapoint with relevant subset of labels from an extremely large label set.

Application:

a) Product recommendation
b) Document tagging
c) Sponsored Search Ads

In these applications, we deal with short-text inputs *i.e.,* query and webpage titles.



SOME PECULIARITIES AND DEMANDS OF XC

a) **Sparse Label Relevance:** Millions of potential labels L exist for each data point, but only a small fraction $\mathcal{O}(\log L)$ of these labels are truly relevant.

b) **Scalability Challenge:** Training and prediction should scale logarithmically $\mathcal{O}(\log L)$ with L, and not as $\Omega(L)$

c) **Data Scarcity:** Many labels have limited training data, often less than 5, hindering accurate prediction.

d) **Missing labels:** Manual annotation of all data points is impractical.

WHAT IS MISSING IN XC METHODS?

Metadata (or memory) infusion during learning can enhance contextual query and label representation and improve the overall task performance:

a) In **sponsored search ads recommendation**, the query-side metadata can be organic search webpage titles clicked in response to the query.

b) In **Wikipedia categories prediction** the metadata can be titles of linked Wikipedia articles.

CHALLENGES OF USING MEMORY

a) **Sensitivity to retrieved metadata:** Low quality retrieval from memory leads to noisy augmentation to the query, degrading task performance.

b) **Influence of metadata form and fusion layer on Latency:** Text-based metadata offers higher interpretability but incurs a higher inference-time.

CHALLENGES OF USING MEMORY

A comparison of the design choices in popular metadata infusion models in the generative (Gen.) and extreme classification (XC) settings.

	Methods	Retrieved Metadata Quality	Metadata Form	Fusion Depth	Inference Latency
Gen. Models	Retrieval-augmented generation	Variable*	Text	Early	High
	Retrieval-interleaved generation	Query representation	Embedding	Early	Very High
	Unified RAG (URAG)	Memory representation	Embedding	Early	High
	GRIT-LM	Memory representation	Text	Early	High
XC Models	OAK	Variable*	Embedding	Late	Low
	DEXA	–	Embeddings	Late	Low
	MOGIC Oracle (Ours)	Memory representation	Text	Early	High
	MOGIC (OAK) (Ours)	Memory representation	Embedding	Late	Low

EXISTING BASELINE: THE OAK APPROACH

a) OAK[1] is a late-fusion, embedding-based method that uses query-metadata to obtain enhanced representations.

b) Late-fusion in OAK improves generalization and lowers inference latency with noisy predicted metadata (P@1: 33.71 vs. 28.49 for early-fusion).

c) Early-fusion models outperform when ground-truth metadata is available (P@1: 47.63 vs. 38.92 for OAK).

d) Fusion method choice reflects a trade-off—late-fusion handles noisy metadata better, while early-fusion yields higher accuracy with clean inputs.

[1] Mohan et al., "OAK: Enriching document representations using auxiliary knowledge for extreme classification," ICML 2024

OUR PROPOSED APPROACH: MOGIC

a) **MOGIC** is a two-phase method combining early-fusion of textual metadata and late-fusion of memory items while maintaining low latency.
    a) In phase one, we train an early-fusion **oracle** with access to ground-truth query and label metadata in the text form.
    b) In phase two, the oracle guides the training of memory-based XC **disciple** model like OAK via a regularization loss.
b) MOGIC maintains real-world inference latency while improving over state-of-the-art XC models by1-2%.

OUR CONTRIBUTION

a) MOGIC significantly improves accuracy on four benchmark XC datasets.

b) It boosts precision, NDCG, and propensity-scored metrics for both memory-based (OAK) and memory-free (DEXA[2], NGAME[3]) models.

c) MOGIC is robust to missing and noisy metadata.

[2] Dahiya et al., "Deep encoders with auxiliary parameters for extreme classification," SIGKDD 2023

[3] Dahiya et al., "NGAME: Negative mining-aware mini-batching for extreme classification," WSDM 2023

THE MOGIC FRAMEWORK

MOGIC comprises four main components:

a) Disciple ($\mathcal{D}$)

b) Oracle ($\mathcal{O}$)

c) Task-specific loss function ($\mathcal{L}_{\text{Oracle}}, \mathcal{L}_{\text{Disciple}}$)

d) Guidance loss function ($\mathcal{L}_{\text{Alignment}}, \mathcal{L}_{\text{Matching}}$)

PHASE 1: ORACLE TRAINING

To train a highly accurate XC oracle, three components are critical:

(a) The task-specific loss function.

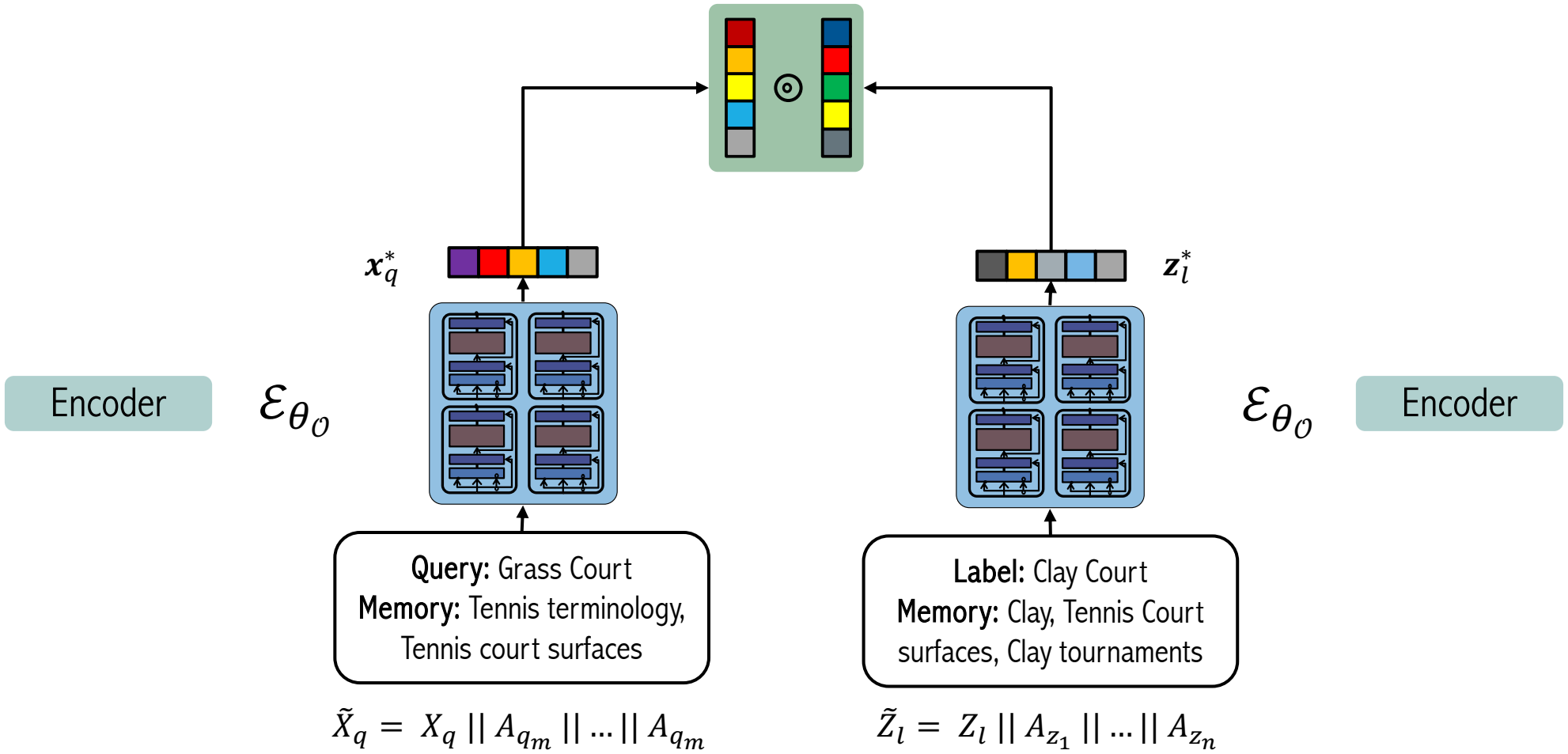
(b) Supervised training data.

(c) Auxiliary metadata, which can enhance the quality of label and query.

ORACLE ($\mathcal{O}$)

Triplet loss

$$\mathcal{L}_{Triplet}\left(\{\boldsymbol{x}_q^*\},\{\boldsymbol{z}_l^*\},\{y_{ql}\}\right) = \sum_{\substack{p:\, y_{pn}=+1 \\ n:\, y_{qn}=-1}} \left[\boldsymbol{x}^{*\top}_q \boldsymbol{z}^*_n \;-\; \boldsymbol{x}^{*\top}_q \boldsymbol{z}^*_p \;+\gamma\right]_+$$



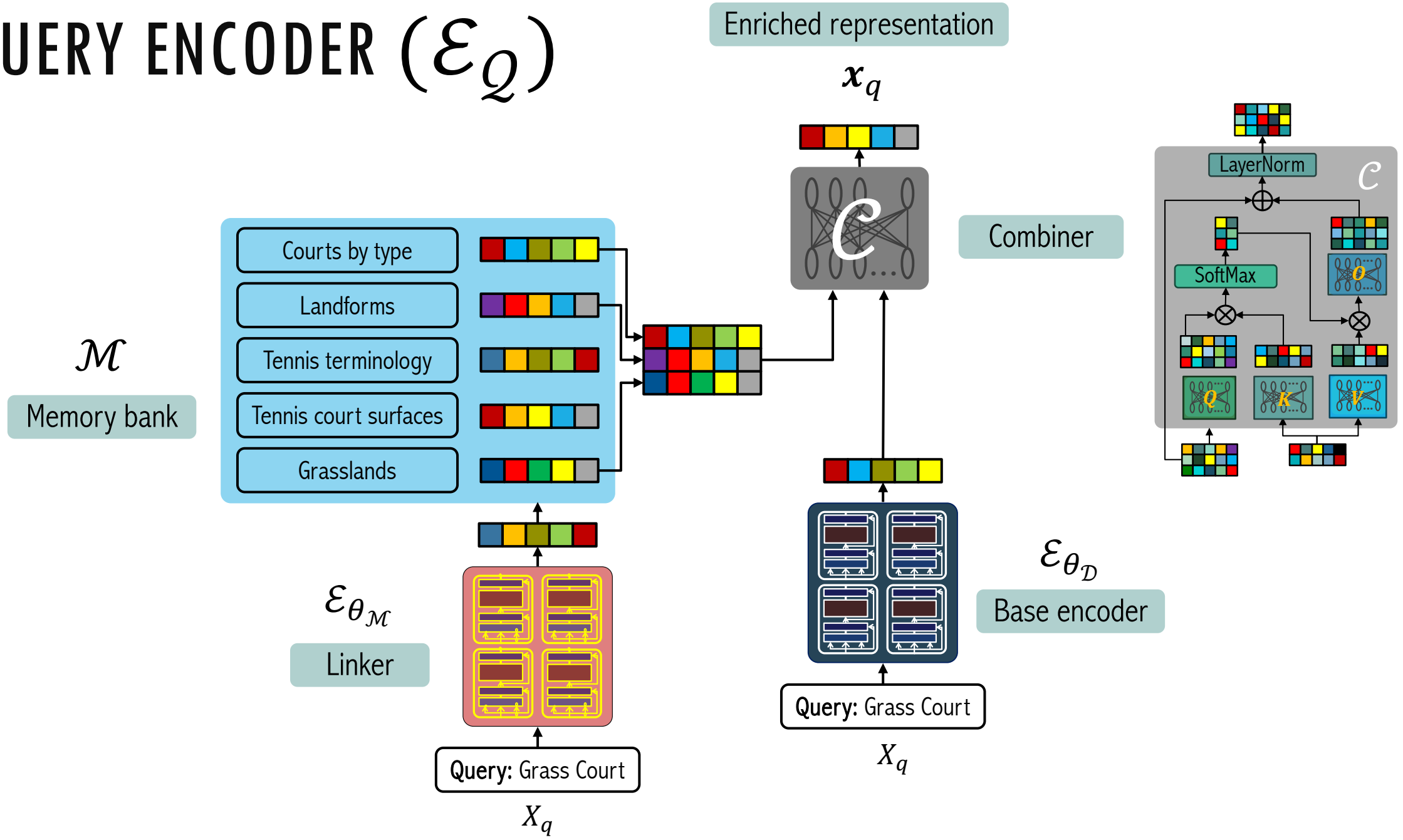
PHASE 2: ORACLE-GUIDED DISCIPLE TRAINING

Disciple training comprises two key components:

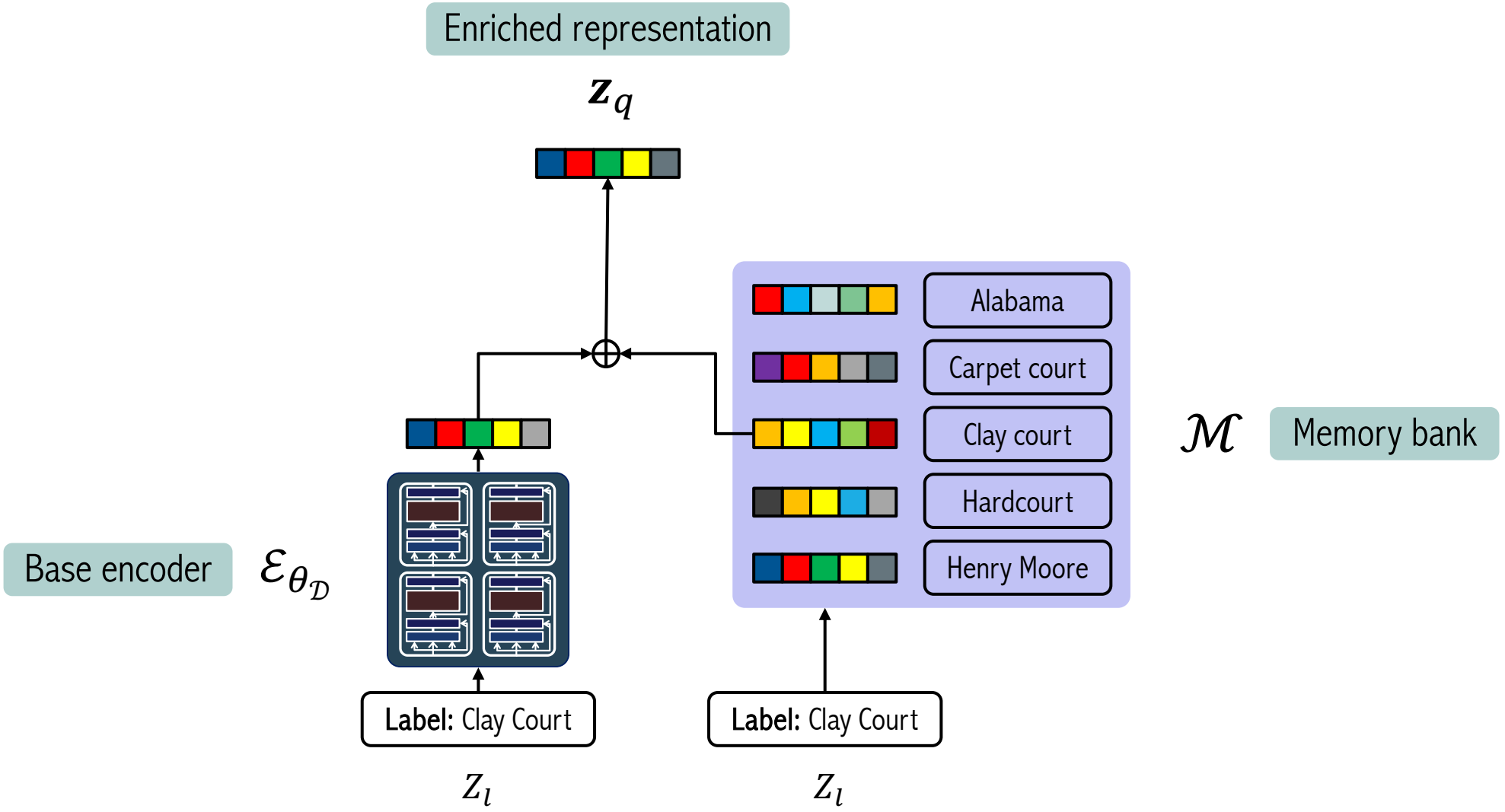
(a) An embedding generator which provides embeddings $\boldsymbol{x}_q$ and $\boldsymbol{z}_l$, associated with a given query X_q and label Z_l

(b) Alignment and Matching losses, together with the task-specific loss, offer oracle-guidance for learning better embeddings.

QUERY ENCODER ($\mathcal{E}_Q$)



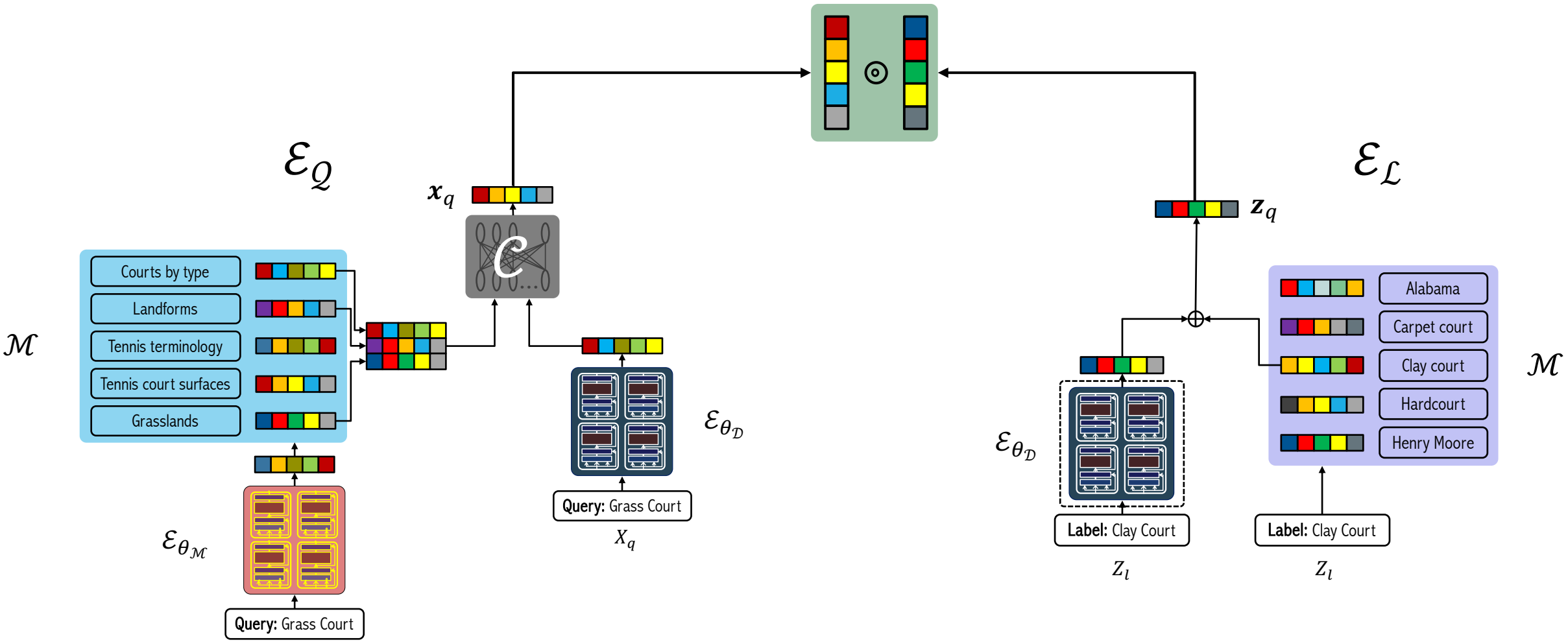
LABEL ENCODER ($\mathcal{E}_{\mathcal{L}}$)



DISCIPLE ($\mathcal{D}$)

Triplet loss

$$\mathcal{L}_{Triplet}\left(\{\boldsymbol{x}_q\},\{\boldsymbol{z}_n\},\{y_{ql}\}\right) = \sum_{\substack{p:\, y_{pn}=+1 \\ n:\, y_{qn}=-1}} \left[\boldsymbol{x}_q^{\top}\boldsymbol{z}_n \;-\; \boldsymbol{x}_q^{\top}\boldsymbol{z}_p \;+\gamma\right]_+$$

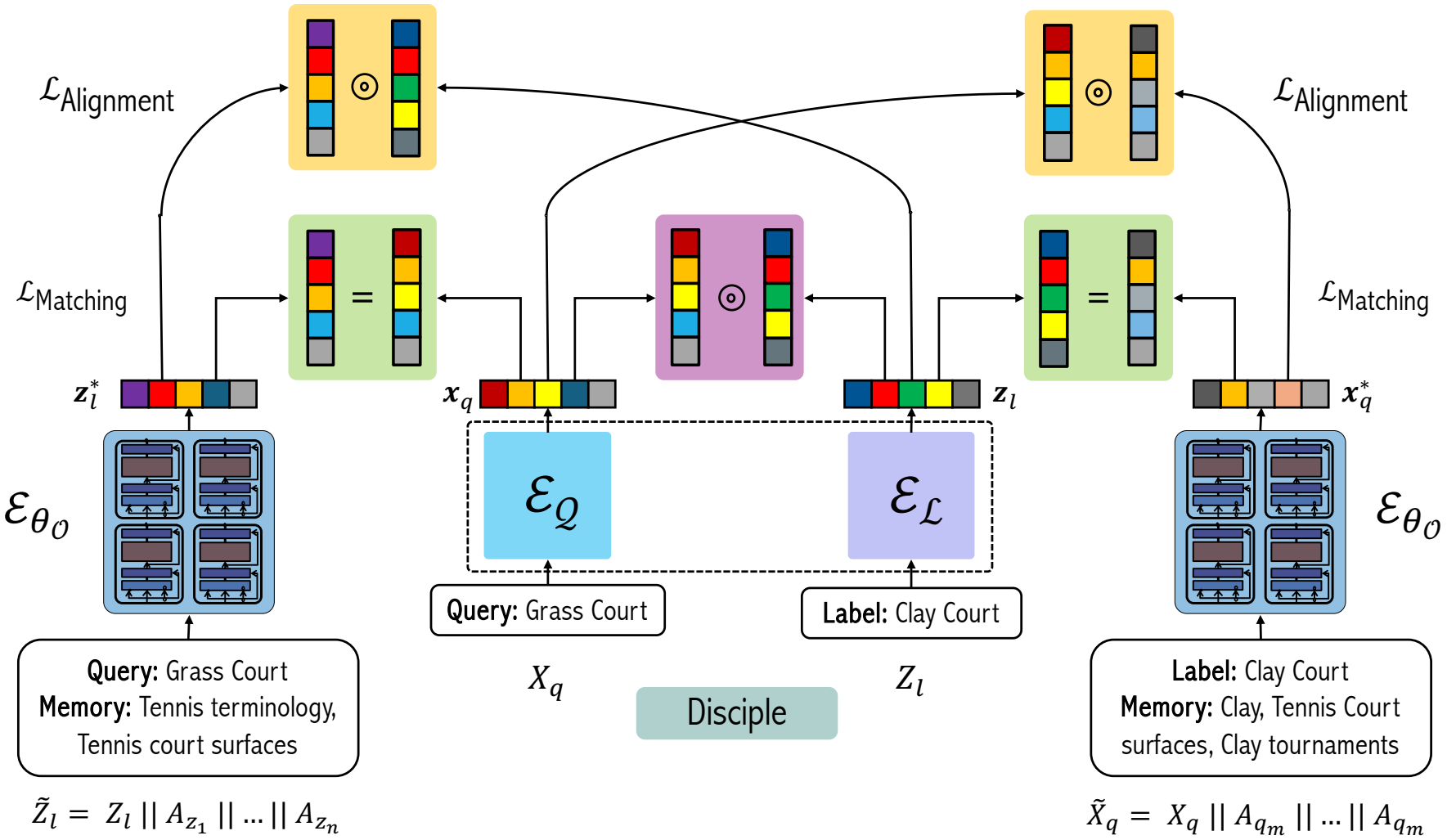


MOGIC

$$\mathcal{L}_{\text{MOGIC}} = \mathcal{L}_{\text{Disciple}} + \alpha \cdot \mathcal{L}_{\text{Alignment}} + \beta \cdot \mathcal{L}_{\text{Matching}}$$

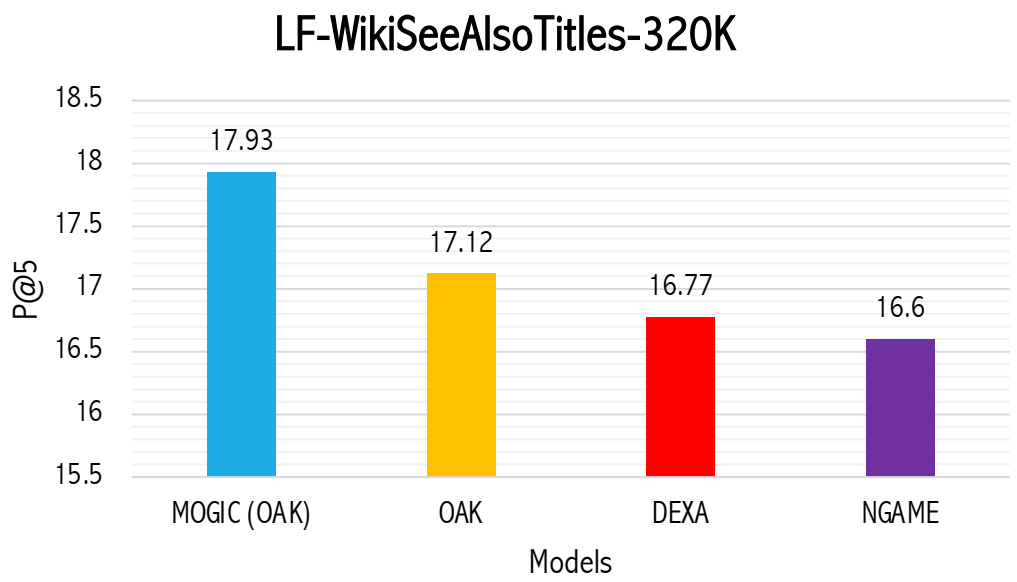
$$\mathcal{L}_{\text{Alignment}} = \mathcal{L}_{\text{Triplet}}(\boldsymbol{x}_q, \boldsymbol{z}_l^*, y_{ql}) + \mathcal{L}_{\text{Triplet}}(\boldsymbol{x}_q^*, \boldsymbol{z}_l, y_{ql})$$

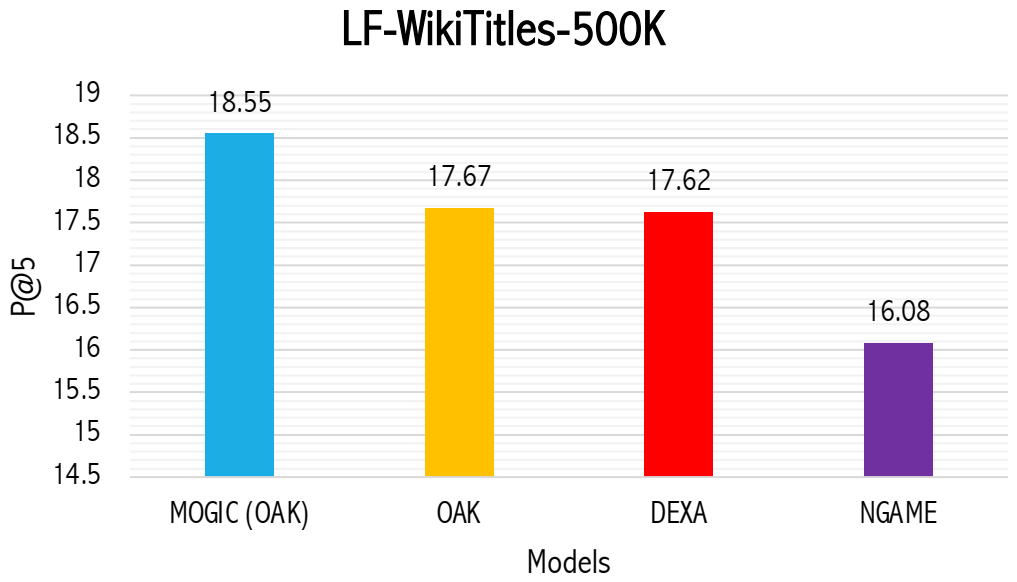
$$\mathcal{L}_{\text{Matching}} = \sum_{q \in Q} \|\boldsymbol{x}_q - \boldsymbol{x}_q^*\|_2 + \sum_{l \in L} \|\boldsymbol{z}_l - \boldsymbol{z}_l^*\|_2$$

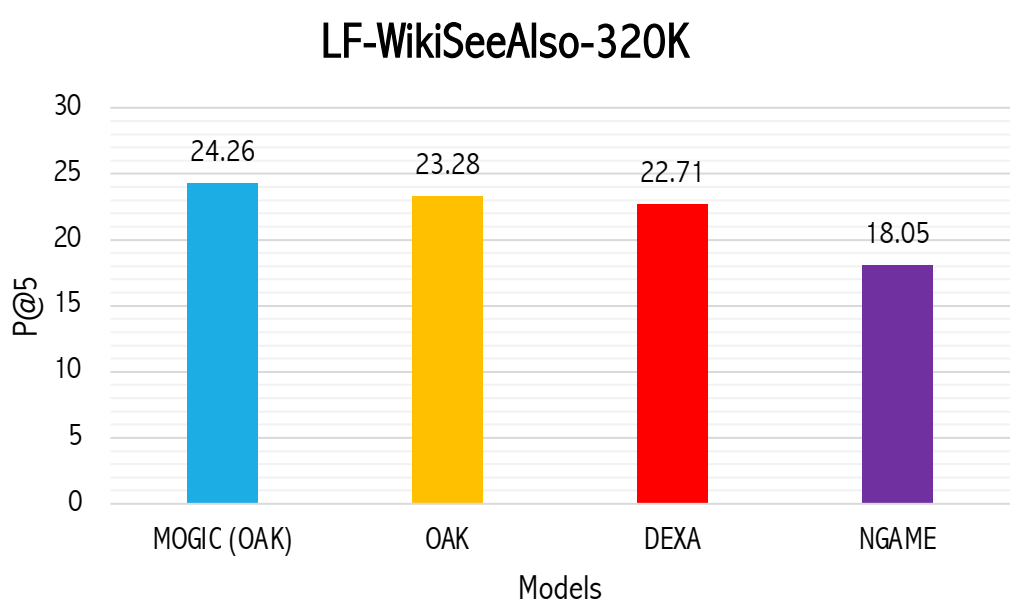


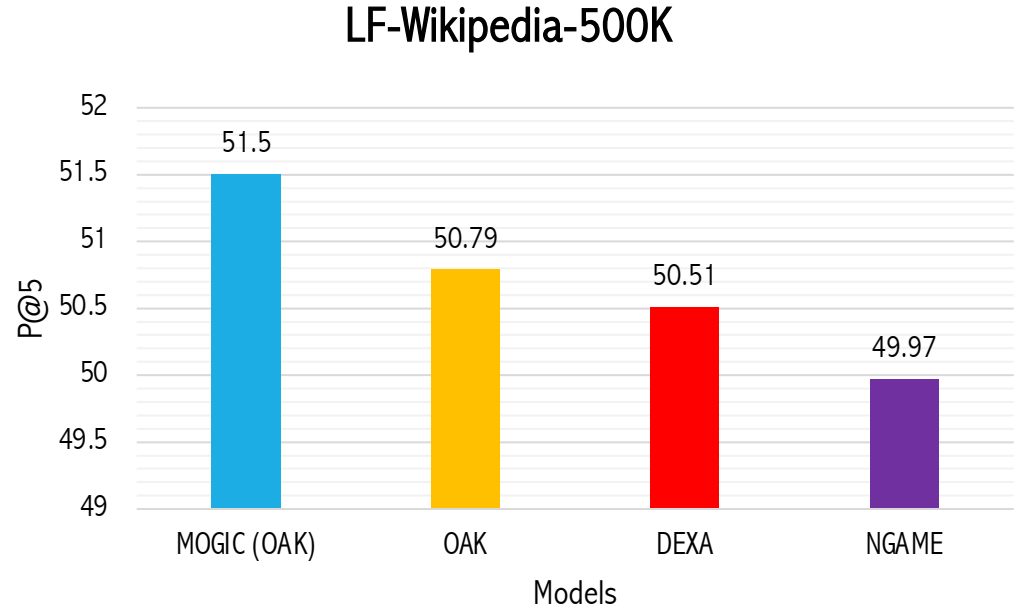
RESULTS

Methods	P@1	P@5	N@5	PSP@1	PSP@5	P@1	P@5	N@5	PSP@1	PSP@5
	LF-WikiSeeAlsoTitles-320K					LF-WikiTitles-500K				
MOGIC (OAK)	34.62	17.93	27.44	35.70	33.18	47.28	18.55	34.97	27.29	26.12
OAK	33.71	17.12	24.53	33.83	30.83	44.82	17.67	33.72	25.79	24.90
DEXA	32.91	16.77	24.63	33.63	29.55	47.41	17.62	33.64	25.27	24.03
NGAME	32.64	16.60	23.44	33.21	29.87	39.04	16.08	30.75	23.12	23.03
	LF-WikiSeeAlso-320K					LF-Wikipedia-500K				
MOGIC (OAK)	49.62	24.26	50.49	36.15	43.17	85.34	51.50	77.85	43.60	61.74
OAK	48.57	23.28	49.16	33.92	40.44	85.23	50.79	77.26	45.28	60.80
DEXA	47.11	22.71	47.62	31.81	38.78	84.92	50.51	76.80	42.59	58.33
NGAME	46.40	18.05	46.64	28.18	33.33	84.01	49.97	75.97	41.25	57.04


LF-WikiSeeAlsoTitles-320K
P@5
18.5
18
17.5
17
16.5
16
15.5
17.93
17.12
16.77
16.6
MOGIC (OAK)
OAK
DEXA
NGAME
Models

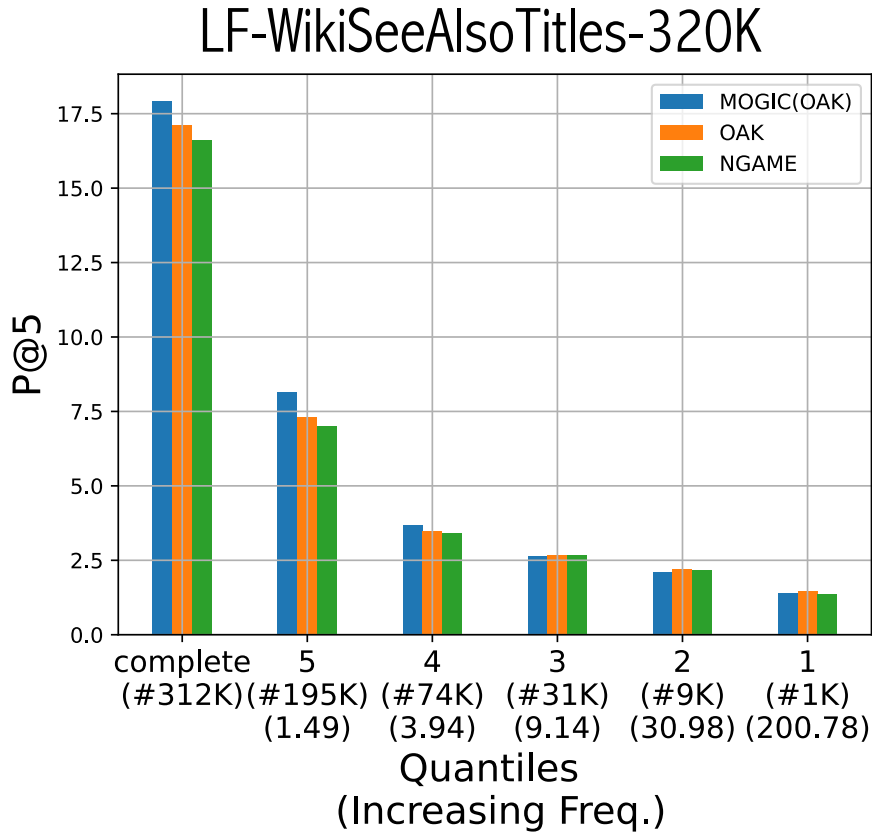

LF-WikiTitles-500K
P@5
19
18.5
18
17.5
17
16.5
16
15.5
15
14.5
18.55
17.67
17.62
16.08
MOGIC (OAK)
OAK
DEXA
NGAME
Models

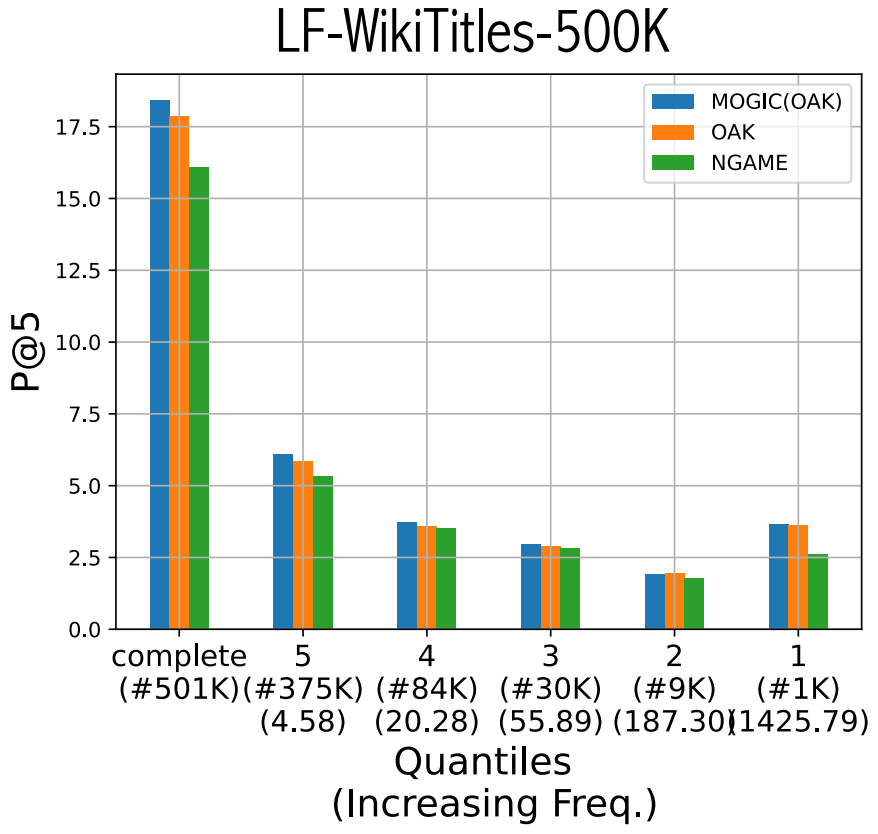

LF-WikiSeeAlso-320K
P@5
30
25
20
15
10
5
0
24.26
23.28
22.71
18.05
MOGIC (OAK)
OAK
DEXA
NGAME
Models


LF-Wikipedia-500K
P@5
52
51.5
51
50.5
50
49.5
49
51.5
50.79
50.51
49.97
MOGIC (OAK)
OAK
DEXA
NGAME
Models

QUANTILE COMPARISON

MOGIC (OAK) gives consistent gains in tail bins and comparable results in head bins





ABLATION

MOGIC on the LF-WikiSeeAlsoTitles-320K dataset with different oracle models.

	MOGIC					Oracle				
Methods	**P@1**	**P@5**	**N@5**	**PSP@1**	**PSP@5**	**P@1**	**P@5**	**N@5**	**PSP@1**	**PSP@5**
DistilBERT	34.62	17.93	27.44	35.70	33.18	42.78	20.53	32.99	43.59	37.57
Phi-2	34.25	17.71	26.97	35.37	32.62	26.84	12.06	24.79	24.49	24.20
LLaMA-2-7b-hf	33.94	17.43	26.87	34.92	32.10	29.57	13.40	27.38	26.69	26.74

MOGIC robust framework can be extended to any XC algorithm and improve its accuracy. In particular, on LF-WikiSeeAlsoTitles-320K, we observe MOGIC can improve accuracy of base algorithm by 1-2% in P@1

Methods	**P@1**	**P@5**	**N@5**	**PSP@1**	**PSP@5**
MOGIC (OAK)	34.62	17.93	27.44	35.70	33.18
OAK	33.71	17.12	24.53	33.83	30.83
MOGIC (NGAME)	32.37	16.38	33.16	26.87	31.08
NGAME	30.72	15.42	31.56	25.18	28.88
MOGIC (DEXA)	32.75	16.92	34.00	26.88	31.82
DEXA	31.57	16.14	32.71	25.64	29.99

MOGIC uses Aligment and Matching losses to regularize base XC algorithm. This table summarizes the impact of each loss term. Here, we show results for the OAK disciple on LF-WikiSeeAlsoTitles-320K. Also, Disciple + Alignment + Matching is same as **MOGIC (OAK)**.

Loss terms in $\mathcal{L}$	**P@1**	**P@5**	**N@5**	**PSP@1**	**PSP@5**
Disciple + Alignment + Matching	34.62	17.93	35.70	27.44	33.18
Disciple + Alignment	34.12	17.66	35.16	26.72	32.57
Disciple + Matching	34.11	17.63	35.24	26.83	32.40
Disciple	33.71	17.12	33.83	24.53	30.83
Alignment + Matching	32.70	16.92	33.60	26.03	31.30

Oracle used in **MOGIC** framework is sensitive to noise is metadata. Introducing noise in metadata used for oracle training can lead to up to 20% reduction in accuracy however, XC

	MOGIC					Oracle				
Noise %	**P@1**	**P@5**	**N@5**	**PSP@1**	**PSP@5**	**P@1**	**P@5**	**N@5**	**PSP@1**	**PSP@5**
0	34.62	17.93	27.44	35.70	33.18	42.78	20.53	32.99	43.59	37.57
20	36.26	18.80	28.66	37.69	34.61	34.80	16.83	26.67	35.64	30.73
40	35.62	18.44	28.36	36.90	34.08	26.75	13.10	20.45	27.56	23.87
60	34.92	18.12	27.94	36.19	33.59	18.65	9.31	14.29	19.44	17.02

QUALITATIVE ANALYSIS

A comparison of predictions from MOGIC(OAK), OAK and the ground-truth, on the Wikipedia See Also prediction task. Legend: Black indicates ground truth, Red indicates incorrect predictions and Green indicates correct predictions.

Document	Predicted Metadata	Ground truth label	MOGIC predictions	OAK predictions
Tangbe	Populated places in Cameroon, Communes of Cameroon, Township divisions of Hebei	Mustang District, Kali Gandaki Gorge, Kali Gandaki River, Upper Mustang, Gandaki River	Mustang District, Kali Gandaki River, Upper Mustang, Gandaki River	Desalpur, Vladivostok, Kitenge, List of currently erupting volcanoes
Gummy candy	Brand name confectionery, Candy, Gummi candies	Jelly bean, Gumdrop, Jelly baby, Swedish Fish, Quince cheese	Jelly bean, Wine gum, Gumdrop, Jelly baby, Orange jelly candy	List of chocolate bar brands, Stick candy, Candy bar, Dragon's beard candy, Orange jelly candy

THANK YOU