RESEARCH TRACK

EMMETT: Extreme Meta-Classification for Large-Scale Zero-Shot Retrieval

Sachin Yadav*, Deepak Saini*, Anirudh Buvanesh*, Bhawna Paliwal, Kunal Dahiya, Siddarth Asokan, Yashoteja Prabhu, Jian Jiao, Manik Varma

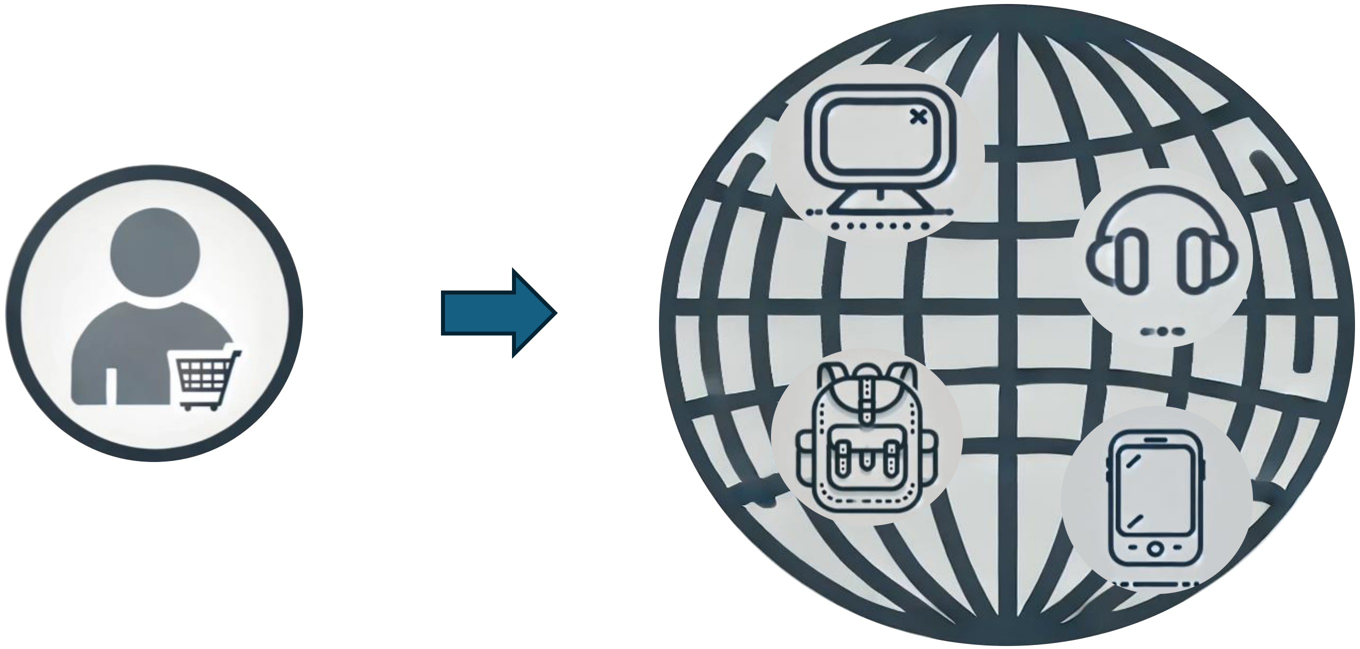
Microsoft Research India

Microsoft

KDD2024
BARCELONA, SPAIN

Large-Scale Retrieval

- Retrieval of items relevant to a user query from a pool of hundreds of millions.

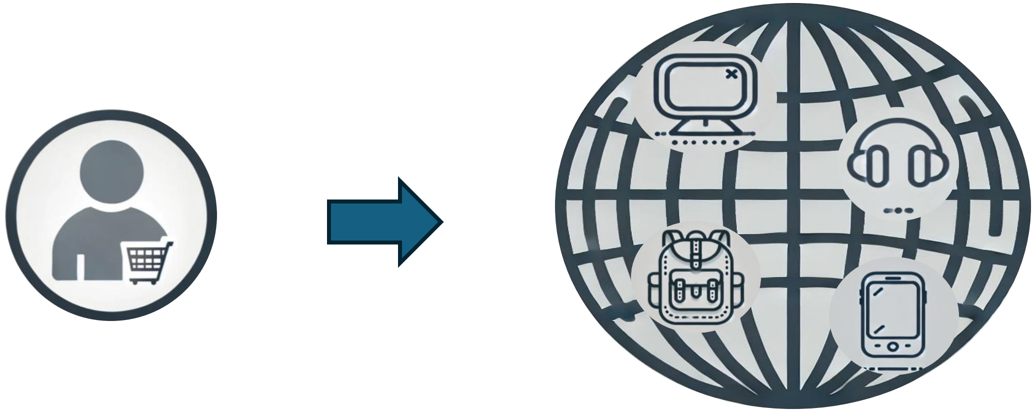


user query

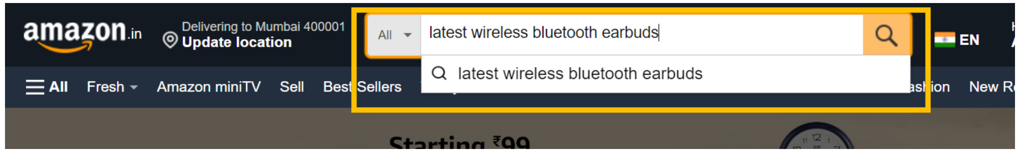
set of items

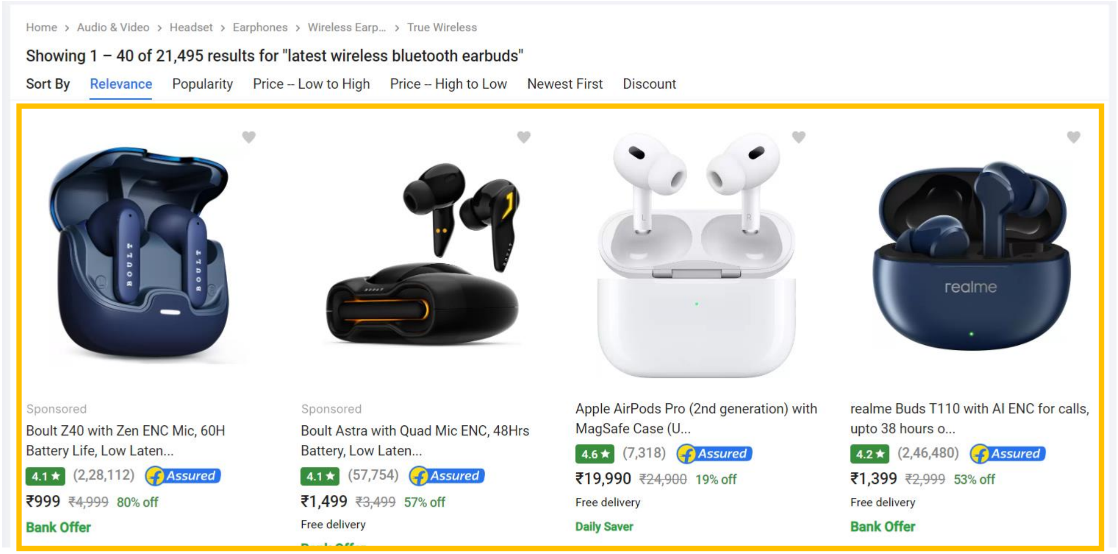
Microsoft

Product Recommendation



"latest wireless bluetooth earbuds"





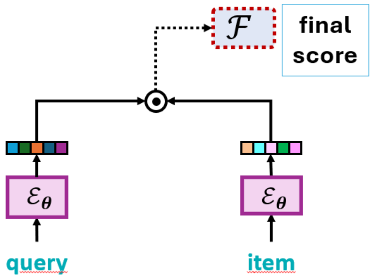
Novel (Zero-Shot) Items

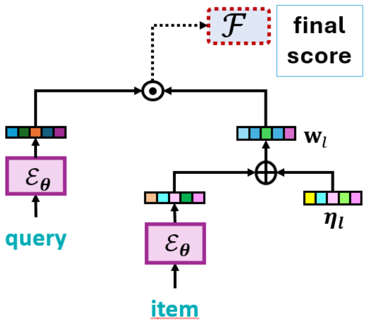
- **Goal:** Efficient and Accurate Solutions for handling zero-shot items.

What we want in an ideal solution:

1. [Accuracy] 🎯 High Accuracy on both novel and observed items.
2. [Efficiency] 🚀 Low Retrieval Time
3. [Efficiency] 🚀 Low Representation time for new items

Previous Approaches





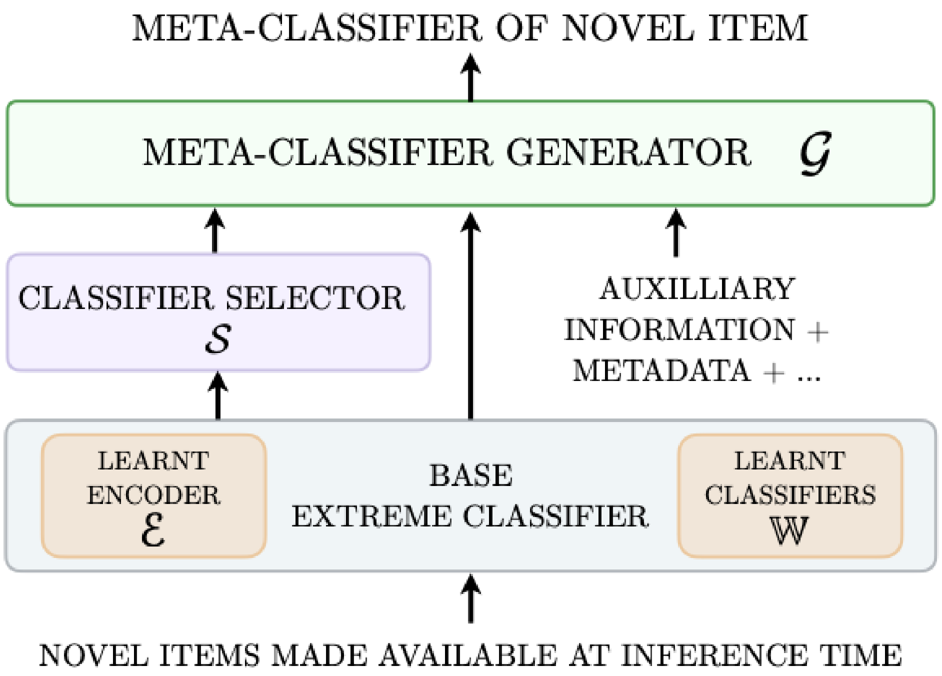
?

	Dense Retrieval	Extreme Classification	What we want!
Accuracy for Observed Items	✓	✓✓	✓✓
Accuracy for Novel Items	✓	✗	✓✓
Retrieval Time	✓✓	✓✓	✓✓
Representation Time	✓✓	✗	✓✓

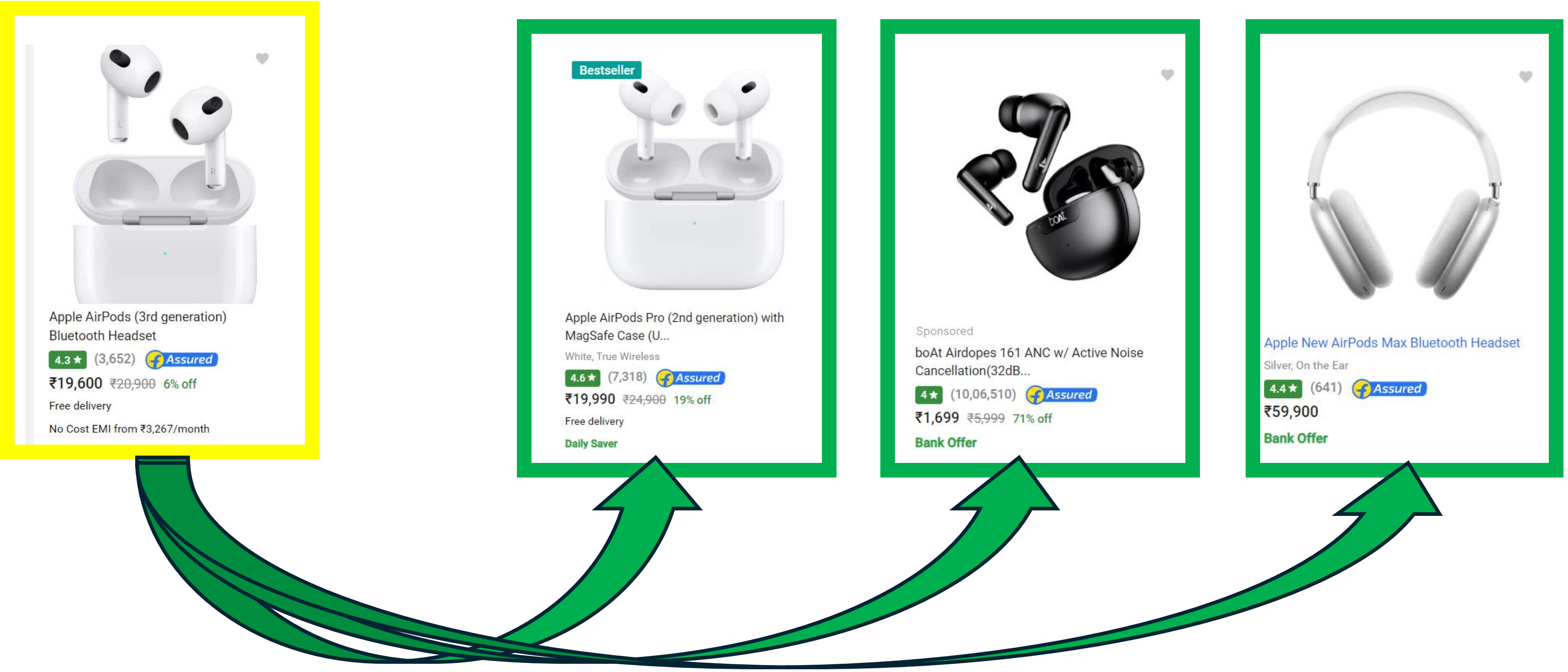
EMMETT: ExtreMe MEta-ClassificaTion

A pipeline built with two modules:

- **Classifier Selector ($\mathcal{S}$):** Takes a novel-item as input and shortlists a few observed item classifiers most informative for it.
- **Meta-classifier Generator ($\mathcal{G}$):** Combines these shortlisted classifiers and auxiliary information about the novel item to synthesize a novel item's classifier.

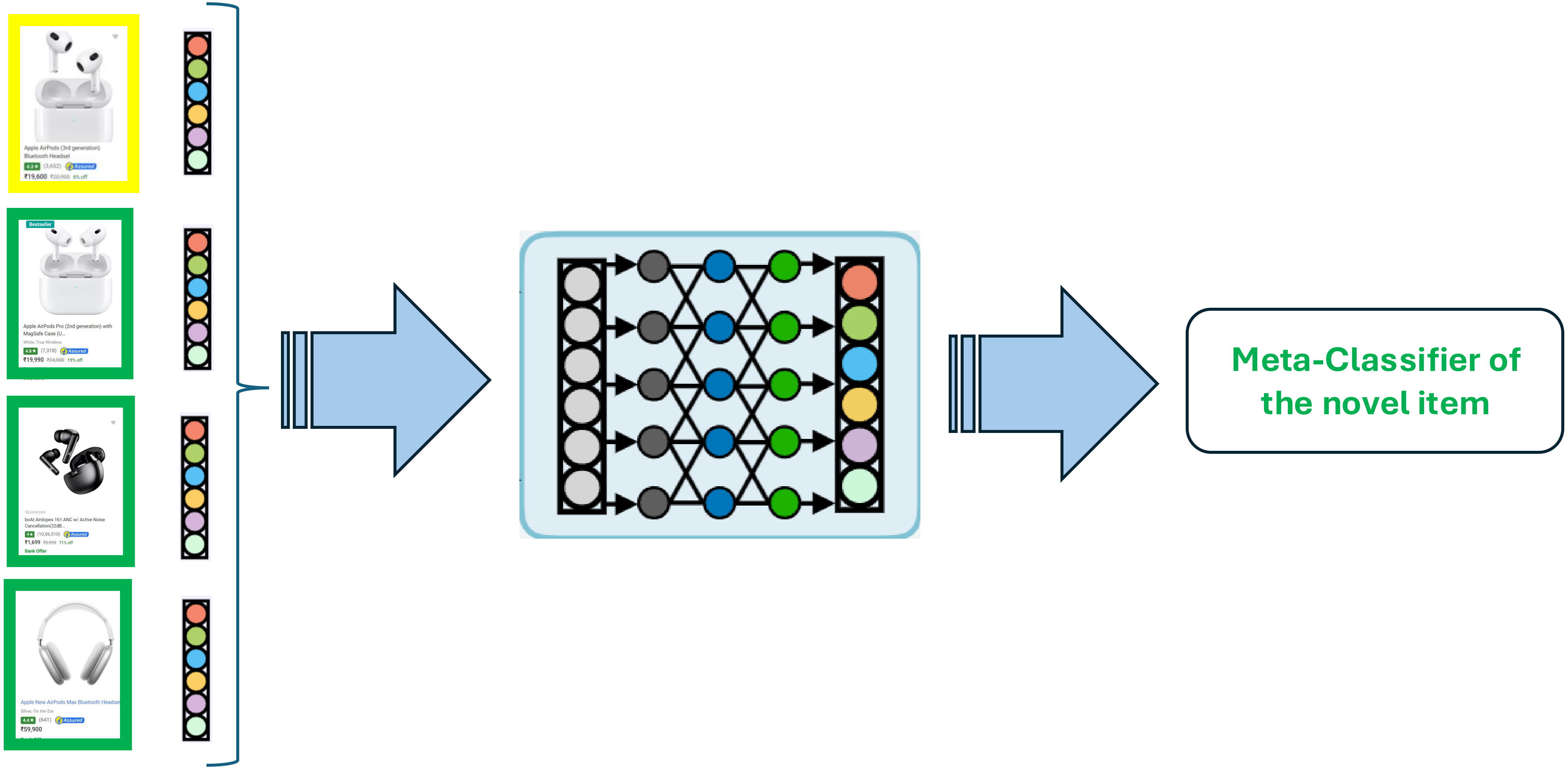


Classifier Selector



Selected classifiers of seen items

Meta-classifier Generator



Generalization Performance of EMMETT

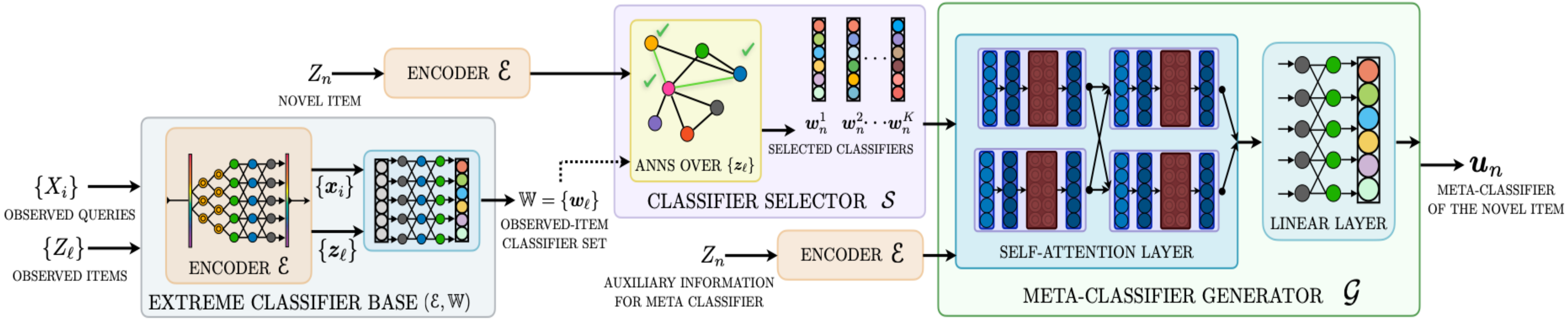
Theorem: *Let R and $\hat{R}$ be the true and empirical risk, and $\hat{\mathcal{R}}_s$ be the empirical Rademacher complexity over the set $\mathcal{S}$ of query-item pairs. Let $p \ll 1$ be the probability of a positively related query-item pair, and q be the probability that $\mathcal{S}$ $(|\mathcal{S}| = M)$ has at most κ positive pairs. Then, $w.p.(1-\delta)$:*

$$R \leq \hat{R} + \hat{\mathcal{R}}_s + 3\left(q + \frac{\sqrt{\ln\left(\frac{2}{\delta - 2q}\right)}}{2M}\right), \text{where } q = \exp\left\{-2M\left(1 - p - \frac{\kappa}{M}\right)\right\}$$

- Insights:
  - The generalization gap is inversely related to the dataset size (*i.e.*, # of items).
  - The large-scale setting improves zero-shot learning.
  - Simpler meta-classifiers (smaller $\hat{\mathcal{R}}_s$) yield superior generalization.

IRENE Algorithm

- **IRENE:** Instance of EMMETT, designed for large-scale zero-shot performance.
  - **ANNS-based Classifier Selector ($\mathcal{S}$):** Uses an Approximate Nearest Neighbor Search (ANNS) index built atop the item encoder-representations.
  - **Transformer-based Meta-classifier Generator ($\mathcal{G}$):** The module combines a shortlist of item classifiers to generate the meta-classifier.



Rademacher Complexity of IRENE

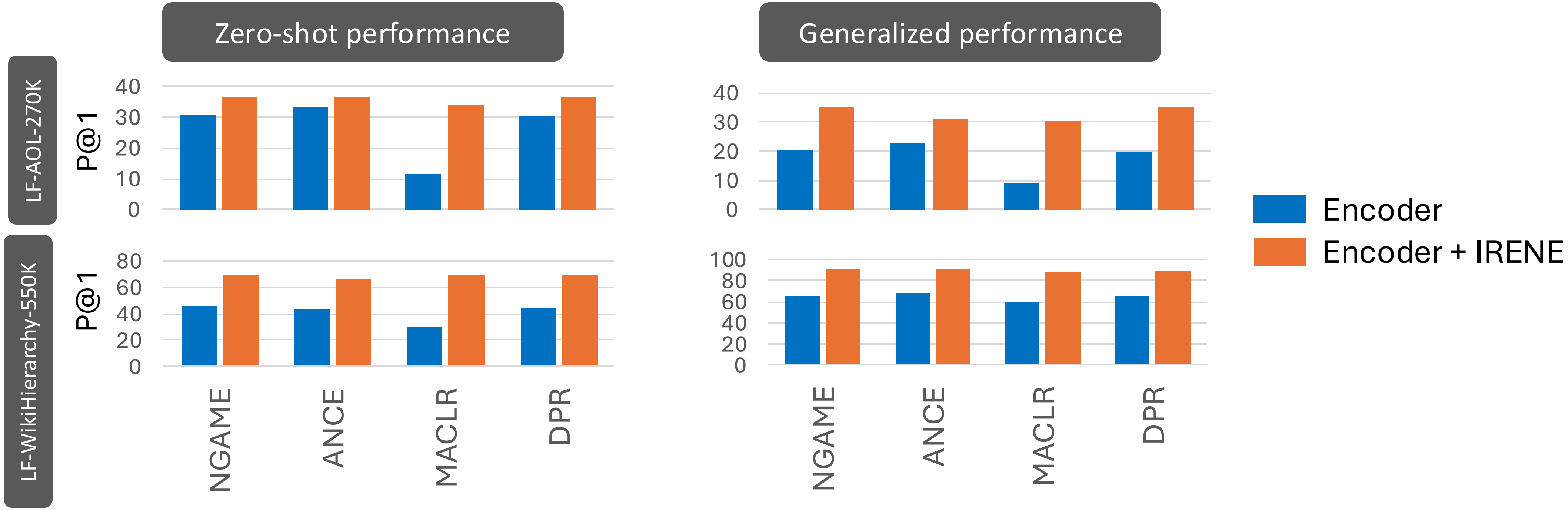
Lemma: *Let $\mathcal{F}$ be the class of functions defined in the IRENE algorithm, comprising pre-determined encoder representations and classifiers, a given classifier selector that outputs K classifiers, and G, the meta-classifier generator. Let $\boldsymbol{M}$ be the weight matrix of the linear layer and $\max_{\boldsymbol{x}}\left(||\boldsymbol{x}||_2\right) = B$. Then, the Rademacher complexity of $\mathcal{F}$ can be bounded as follow:*

$$\hat{\mathcal{R}}_s(\mathcal{F}) \leq \mathcal{O}\left(\mathrm{B}||\boldsymbol{M}||_2\sqrt{d\ \ln(K+1)}\right)$$

- Rademacher complexity of IRENE meta-classifier improves with smaller K:
    - Combining $\mathrm{K} \approx 3$ seen classifiers yields better generalization!
- Analysis for XC classifiers and IRENE with trainable classifier in the paper!

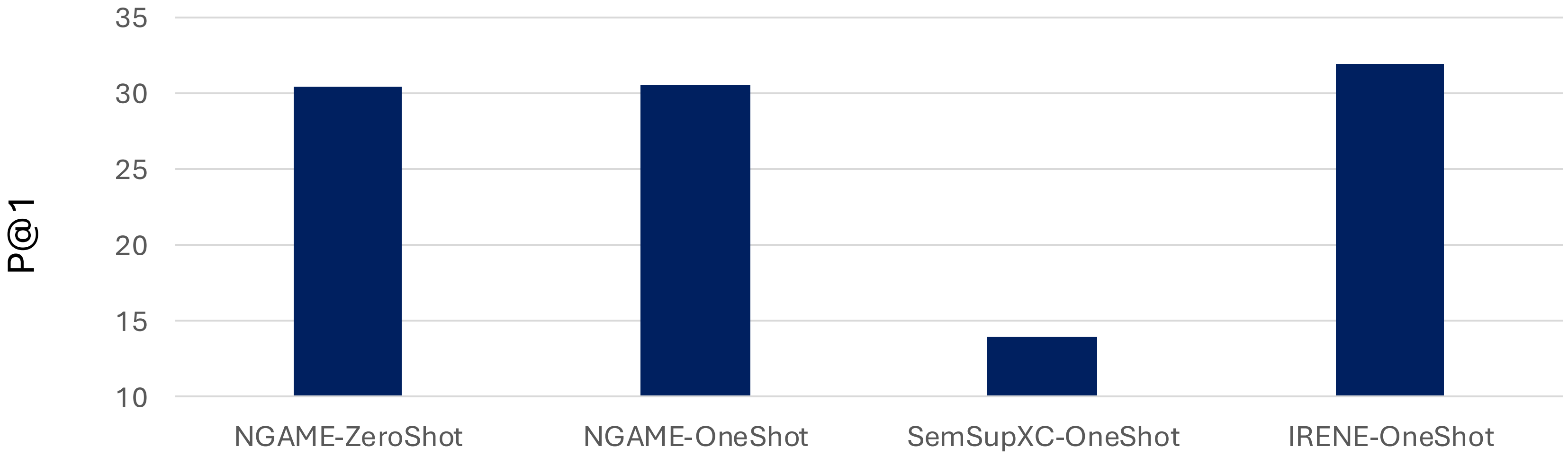
Evaluation on Benchmark Datasets

- Up to 39% improvement in zero-shot and 29% improvement in the generalized setting.



Additional Results

- IRENE can be Adapted seamlessly to one-shot or few-shot.
- Baseline NGAME-OneShot is retrained on the revealed data.
- IRENE out-performs baseline NGAME-OneShot **by 1-2%** without any additional training overhead.



Ablations

- Ablation carried out on
  - The depth D of the generator $\mathcal{G}$.
  - The number of classifiers K shortlisted by the selector $\mathcal{S}$.
  - Architecture of the generator $\mathcal{G}$.
- Setting $D = 1$ and $K = 3$ works well in practice, balancing the performance and computational overhead in learning meta-classier generators.

	Ablations	P@1 ↑	P@5 ↑	R@10 ↑
	IRENE ($D = 1, K = 3$)	69.29	38.81	80.40
Generator ($\mathcal{G}$)	$D = 2,\ K = 3$	70.36	39.06	80.39
	$D = 4,\ K = 3$	70.71	39.11	80.27
	$\mathcal{G}$ as Sum, $K = 3$	45.49	25.46	59.01
	$\mathcal{G}$ as wt. Sum, $K = 3$	46.39	25.88	59.29
Selector ($\mathcal{S}$)	$D = 1,\ K = 1$	68.98	38.56	80.11
	$D = 1,\ K = 2$	69.34	38.74	80.25
	$D = 1,\ K = 6$	69.99	39.12	80.79
	$D = 1,\ K = 20$	69.07	38.57	79.80

Real-world Deployment

- We conducted A/B testing on Microsoft Bing to match user queries with advertisement keywords.
- IRENE boosted the click-through rate by **4.2%** and improved prediction quality by **9%**, according to expert evaluations.
- IRENE encodes a novel keyword in **under 1 millisecond!**

Conclusions

- We studied **large-scale zero-shot retrieval** and developed techniques to efficiently and accurately represent novel items.
- We proposed **EMMETT**, a generic algorithmic framework for learning accurate meta-classifiers for novel items that ellucidates accuracy versus efficiency trade-offs.
- We proposed **IRENE**, a novel, practically deployable algorithm to boost the zero-shot performance of any Siamese encoder.
- IRENE atop leading encoders improves the zero-shot retrieval accuracy by up to 15% points, and improves the ad click-through rate by 4.2%.

RESEARCH TRACK

EMMETT: Extreme Meta-Classification for Large-Scale Zero-Shot Retrieval

Scan for more details!

KDD2024
BARCELONA, SPAIN