Improving Pinterest Search Relevance Using Large Language Models

Pinterest Engineering · Following

Published in Pinterest Engineering Blog

7 min read · Apr 4, 2025

Han Wang | Machine Learning Engineer, Relevance & Query Understanding; Mukuntha Narayanan | Machine Learning Engineer, Relevance & Query Understanding; Onur Gungor | (former) Machine Learning Engineer, Relevance & Query Understanding; Jinfeng Rao | Machine Learning Engineer, Pinner Discovery

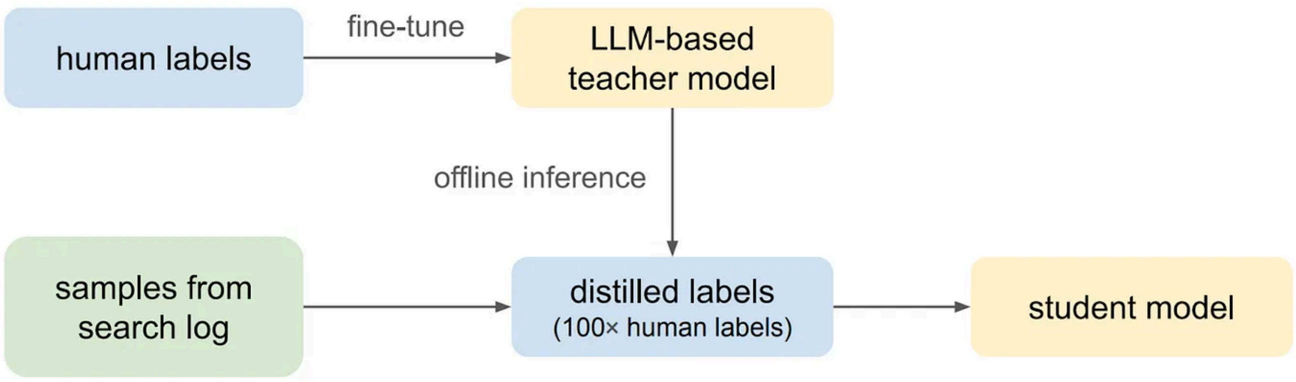
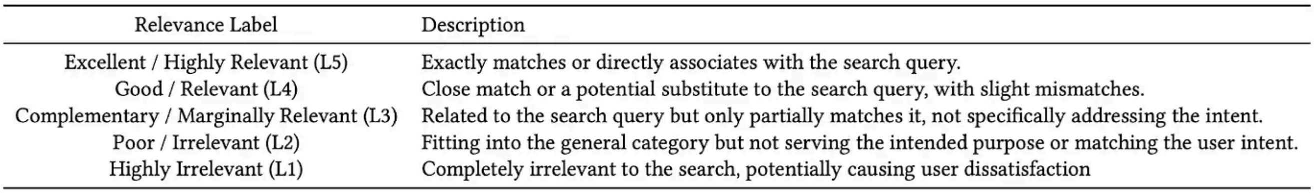


Figure: Illustration of the search relevance system at Pinterest.

Background

Pinterest Search is one of the key surfaces on Pinterest where users can discover inspiring content that aligns with their information needs. Search relevance measures how well the search results aligned with the search query. Using a relevance objective allows the search engine to ensure that the content displayed to users is genuinely pertinent to their information needs, rather than overly relying on factors like past user engagement.

In this work, we focus on improving the search relevance model. To measure the relevance between queries and Pins, we use a 5-level guideline (see Table 1).



Relevance Label	Description
Excellent / Highly Relevant (L5)	Exactly matches or directly associates with the search query.
Good / Relevant (L4)	Close match or a potential substitute to the search query, with slight mismatches.
Complementary / Marginally Relevant (L3)	Related to the search query but only partially matches it, not specifically addressing the intent.
Poor / Irrelevant (L2)	Fitting into the general category but not serving the intended purpose or matching the user intent.
Highly Irrelevant (L1)	Completely irrelevant to the search, potentially causing user dissatisfaction

Table 1: 5-scale Pin relevance guidelines.

In this blog, we will go through the technical design and share some offline and online results for our LLM-based search relevance pipeline. More details can be found in our full paper.

Technical Design

LLM as Relevance Model

Model Architecture

We use a cross-encoder language model to predict a Pin's relevance to a query, along with Pin text, as shown in Figure 1. The task is formulated as a multiclass classification problem. We fine-tune the models using human-annotated data, minimizing cross-entropy loss.

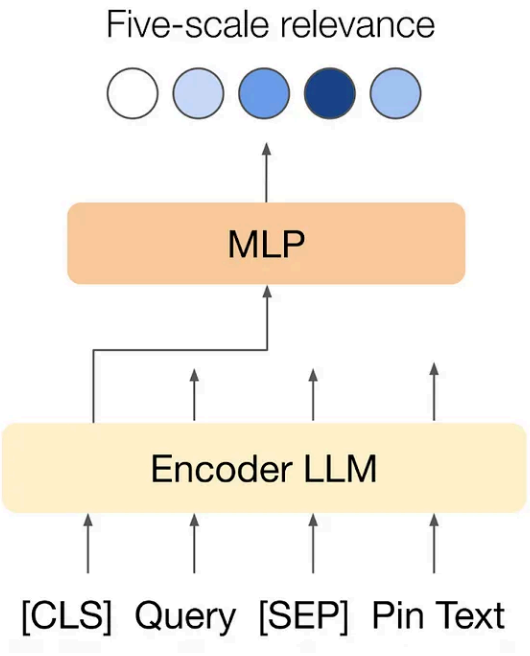


Figure 1: The cross-encoder architecture in the relevance teacher model. Take the encoder language models (e.g., BERT-based models) for illustration.

Pin Text Representations

Pins on Pinterest are rich multimedia entities that feature images, videos, and other contents, often linked to external webpages or blogs. To represent each Pin, we use the following varied set of text features derived from metadata, the image itself, as well as user-curated data. These features are designed with a focus on providing reliable high-quality representations, while retaining high coverage across Pins on Pinterest Search.

- **Pin titles and descriptions:** the titles and the descriptions assigned by the user who created the Pin.
- **Synthetic image captions**: synthetic image descriptions generated by Bootstrapping Language-Image Pre-training (BLIP), an off-the-shelf image captioning model.
- **High-engagement query tokens**: unique queries with the highest engagement with this Pin on the search surface over the past two years.
- **User-curated board titles**: titles of user-curated boards where the Pin has been saved.
- **Link titles and descriptions:** titles and descriptions from the linked external webpages.

Distill LLM into Servable Student Model

Model Architecture

Our cross-encoder LLM-based classifier is hard to scale for Pinterest Search due to real-time latency and cost considerations. Therefore, we use knowledge distillation to distill the LLM-based teacher model into a lightweight student relevance model. The student model served online uses the following features:

- Query-level features: query interest features, shopping interest features, and SearchSAGE query embeddings
- Pin-level features: PinSAGE embedding, visual embedding for the image, and SearchSAGE Pin embeddings

- Query-Pin interaction features: BM25 and text match scores for different text fields, historical engagement rates between the Pin and query, etc.

These features are embedded and passed through a feed-forward network to predict 5-scale relevance scores, as shown in Figure 2.

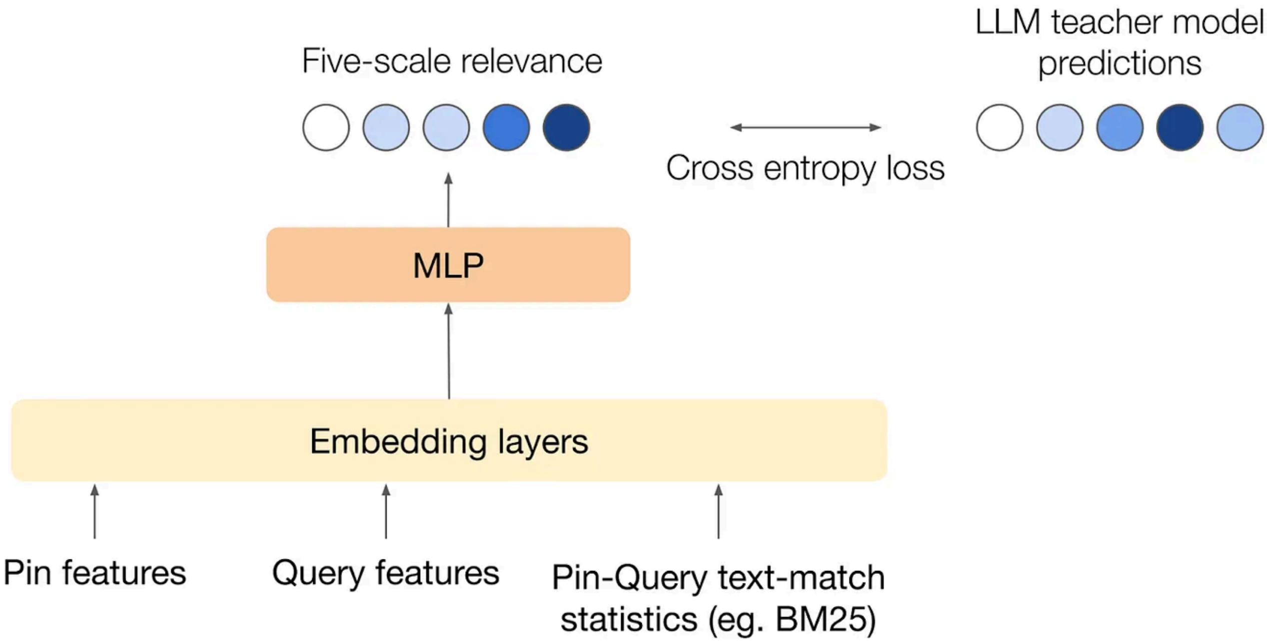


Figure 2: The architecture of the online-served student model, which is trained via distillation from the LLM-based teacher model.

Knowledge Distillation and Semi-Supervised Learning

To train our student relevance model, we employ the LLM-based teacher model to generate 5-scale relevance labels on a daily logged large search engagement and impression dataset with billions of rows. This labeled dataset is subsequently used to train the much smaller student model. A diagram of the search relevance system at Pinterest is shown in Figure 3. The relevance scores generated by the student model are then utilized alongside engagement predictions to determine the final ranking of search results.

This blend of knowledge distillation and semi-supervised learning not only makes effective use of vast amounts of initially unlabeled data, but also expands the data to a wide range of languages from around the world and new concepts not encountered in our human-labeled data owing to the seasonality in Pinterest Search. By using a multilingual LLM-based teacher model, we are able to successfully generalize from human-labeled data focused on US queries to unseen languages and countries.

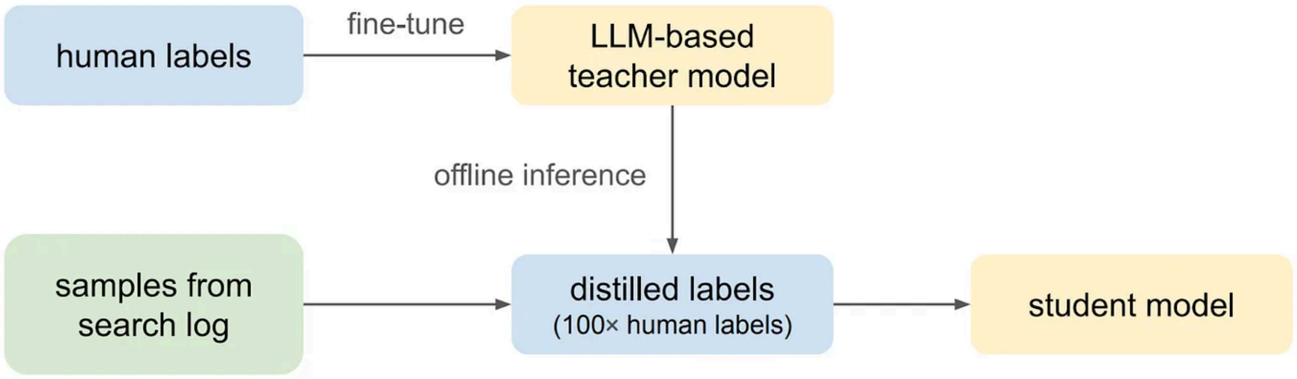


Figure 3: Diagram of the proposed search relevance system at Pinterest.

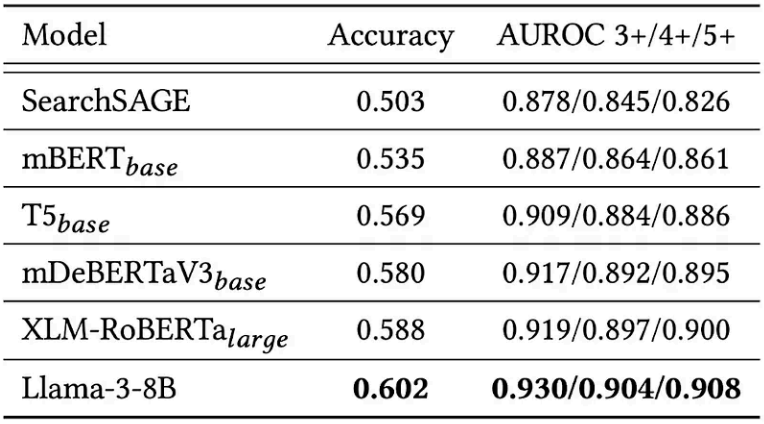
Offline Experiments

We now present offline experiments to demonstrate the effectiveness of each modeling decision. The teacher model is trained and evaluated using human-annotated relevance labels. In all offline experiments, we report the accuracy of 5-scale relevance predictions and the AUROC metrics for binarized labels with thresholds at 3, 4, and 5, since correctly identifying highly relevant content is more important for search ranking.

Comparison of Language Models

In this experiment, we evaluate the following pre-trained language models: multilingual BERT-base, T5-base, mDeBERTa-V3-base, XLM-RoBERTa-large, and Llama-3–8B. These models are initialized from Hugging Face checkpoints and fine-tuned using our in-house search relevance training data. For larger language models such as Llama, we first load quantized model weights and then apply qLoRA for fine-tuning. Additionally, we incorporate gradient checkpointing and mixed precision techniques to further improve training efficiency and memory usage.

Table 3 shows the performance of different language models. As a baseline, we include a model that relies solely on the SearchSAGE embeddings. In this comparison, we keep the text features for each Pin and the maximum text length fixed, varying only the language models. The results in Table 3 clearly demonstrate that the language models offer additional improvements over our in-house content and query embedding. Furthermore, more sophisticated language models and larger model sizes consistently enhance the relevance prediction performance. Specifically, the Llama-3–8B outperforms the multilingual BERT-base model by 12.5% and the baseline model by 19.7% in terms of 5-scale accuracy.

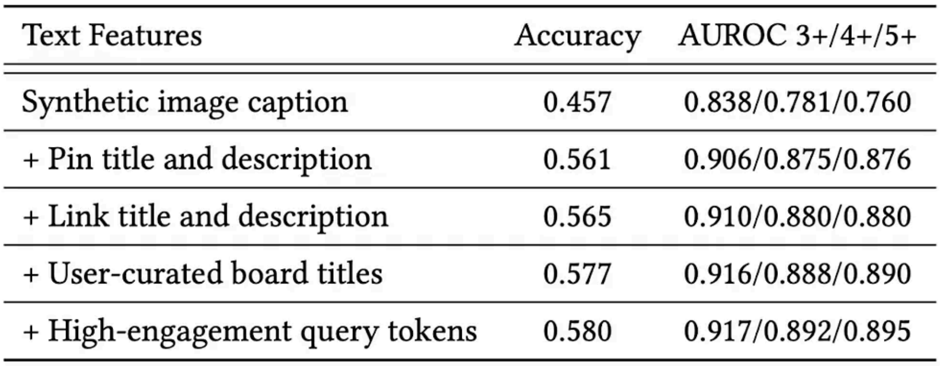


Model	Accuracy	AUROC 3+/4+/5+
SearchSAGE	0.503	0.878/0.845/0.826
mBERT$_{base}$	0.535	0.887/0.864/0.861
T5$_{base}$	0.569	0.909/0.884/0.886
mDeBERTaV3$_{base}$	0.580	0.917/0.892/0.895
XLM-RoBERTa$_{large}$	0.588	0.919/0.897/0.900
Llama-3-8B	**0.602**	**0.930/0.904/0.908**

Table 3: Comparisons of different language models on 5-scale relevance prediction. The AUROC metrics are reported for binarized labels with thresholds 3, 4, and 5.

Importance of Enriching Text Features

To predict the relevance of a Pin to a query using only textual information, we enrich the Pin text representations with several carefully designed text features. We conduct an analysis to assess the impact of each text feature on relevance prediction, using mDeBERTa-V3-base as the language model and fixing the maximum text length to 256. The results, summarized in Table 4, demonstrate that the model's performance consistently improves with the sequential addition of these text features. This indicates that enriched text features and metadata significantly contribute to building a more robust relevance model.



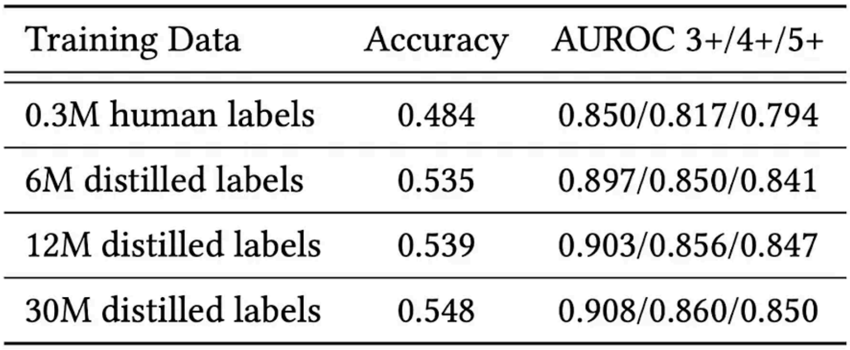
Text Features	Accuracy	AUROC 3+/4+/5+
Synthetic image caption	0.457	0.838/0.781/0.760
+ Pin title and description	0.561	0.906/0.875/0.876
+ Link title and description	0.565	0.910/0.880/0.880
+ User-curated board titles	0.577	0.916/0.888/0.890
+ High-engagement query tokens	0.580	0.917/0.892/0.895

Table 4: Benchmark the improvement with the sequential addition of text features.

Scaling Up Training Labels through Distillation

By using knowledge distillation and semi-supervised learning, we can effectively scale the training data beyond the limited human-annotated data. Table 5 demonstrates the benefits of training on increasing amounts of augmented teacher-generated labels.

Table 5: Comparisons of production model performance when training on different amounts of labels.



Training Data	Accuracy	AUROC 3+/4+/5+
0.3M human labels	0.484	0.850/0.817/0.794
6M distilled labels	0.535	0.897/0.850/0.841
12M distilled labels	0.539	0.903/0.856/0.847
30M distilled labels	0.548	0.908/0.860/0.850

Table 5: Comparisons of production model performance when training on different amounts of labels.

Online Results

Besides offline experiments, we also conducted an online A/B experiment to assess the effectiveness of our new relevance model.

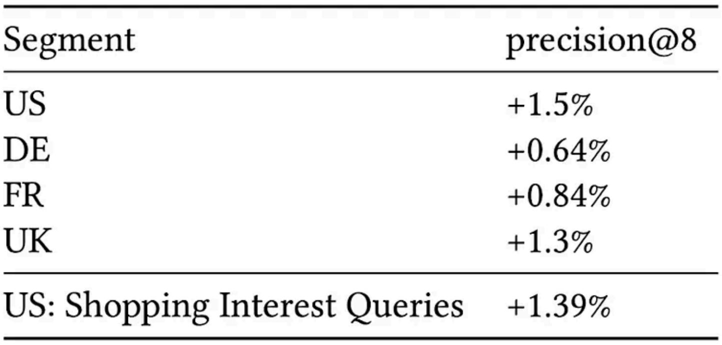
Human Relevance Evaluations

We set up evaluations with human annotators to assess the relevance of the search feeds with and without the new relevance model serving traffic. The nDCG@K here is calculated as follows (more details in our paper):

$$nDCG@K = \frac{\sum_{k=1}^{K} 0.25(L-1)/log_2(1+k)}{\sum_{k=1}^{K} 1/log_2(1+k)},\ L \in \{1,2,3,4,5\}.$$

Our proposed relevance modeling pipeline leads to a **+2.18%** improvement in search feed relevance, as measured by nDCG@20.

The results in Table 6 indicate that the multilingual LLM-based relevance teacher model effectively generalizes across languages not encountered during training (precision@8 used for controlling annotator costs).

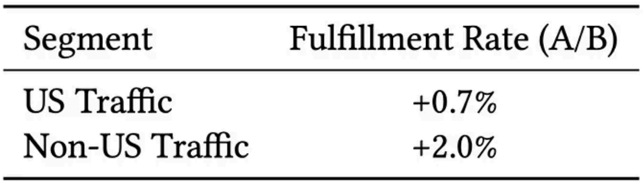


Segment	precision@8
US	+1.5%
DE	+0.64%
FR	+0.84%
UK	+1.3%
US: Shopping Interest Queries	+1.39%

Table 6: Human relevance judgements with search feeds seen across different countries demonstrate generalization across unseen languages.

User Triggered Experiments

In addition to relevance, another primary metric is the search fulfillment rate. This metric is defined as the number of search sessions that result in a high-significance user action. The improvements in relevance also result in increased fulfillment in non-US countries, despite not having annotated data available for those countries during model training (Table 7).



Segment	Fulfillment Rate (A/B)
US Traffic	+0.7%
Non-US Traffic	+2.0%

Table 7: Search Fulfillment Rate increases with the new relevance system show a significant uptick globally.

Summary

In this work, we presented an LLM-based relevance system for Pinterest Search. We thoroughly described each building block of this system, including model architecture, enriched text features, augmented label generation, and online serving. We conducted extensive offline experiments to validate the effectiveness of each modeling decision. Lastly, we presented the results from online A/B experiment, which showed an improvement of >1% in search feed relevance and >1.5% in search fulfillment rates.

Future Work

To further enhance the efficacy of our relevance system, future work will explore the integration of servable LLMs, vision-and-language multimodal models (VLMs), and active learning strategies to dynamically scale and improve the quality of the training data.

Acknowledgement

- Ads Relevance: Helen Xu, Rakesh Chalasani
- Search Leadership: Krishna Kamath, Kurchi Subhra Hazra

Pinterest Engineering Pinner Experience Machine Learning

Recommendations

Follow

Published in Pinterest Engineering Blog

15.7K Followers · Last published Apr 4, 2025

Inventive engineers building the first visual discovery engine, 300 billion ideas and counting.

Following

Written by Pinterest Engineering

58K Followers · 329 Following

https://medium.com/pinterest-engineering | Inventive engineers building the first visual discovery engine https://careers.pinterest.com/

No responses yet

navneet chaudhary

What are your thoughts?