E-Commerce Product Categorization with LLM-based Dual-Expert Classification Paradigm

Zhu Cheng* **Wen Zhang*** **Chih-Chi Chou** **You-Yi Jau**
Archita Pathak **Peng Gao** **Umit Batur**
Amazon, Seattle, WA, USA
{zzcheng, wenzhaw, jimmchou, jauyouyi, arcpatha, gaope, baturab}@amazon.com

Abstract

Accurate product categorization in e-commerce is critical for delivering a satisfactory online shopping experience to customers. With the vast number of available products and the numerous potential categories, it becomes crucial to develop a classification system capable of assigning products to their correct categories with high accuracy. We present a dual-expert classification system that utilizes the power of large language models (LLMs). This framework integrates domain-specific knowledge and pre-trained LLM's general knowledge through effective model fine-tuning and prompting techniques. First, the fine-tuned domain-specific expert recommends top K candidate categories for a given input product. Then, the more general LLM-based expert, through prompting techniques, analyzes the nuanced differences between candidate categories and selects the most suitable target category. We introduce a new in-context learning approach that utilizes LLM self-generated summarization to provide clearer instructions and enhance its performance. Experiments on e-commerce datasets demonstrate the effectiveness of our LLM-based Dual-Expert classification system.

1 Introduction

Accurate product categorization on e-commerce sites is the foundation of a structured catalog system to better meet customer needs. A catalog with accurate categorization helps fuel the search engine, which scopes and ranks the search results from queries efficiently. The buyers can find relevant products through the query or browse directly from the targeted categories. The customer behavior can further enhance the downstream personalized tasks like advertisement and item recommendations. Eventually the accurate catalog leads to customer satisfaction as well as the revenue.

Assigning the category for every single product in the world is far from simple. The problem is to map the product description to the label under a well-defined category taxonomy, which includes over thousands of labels. The category selected by the sellers can be noisy due to the vast number of labels and different interpretation of the categories. Reviewing and fixing the wrongly assigned items manually is not feasible. Therefore, the catalog relies on a categorization system, which utilizes a classification model with high accuracy and coverage to improve the catalog quality.

Although the classification problems have been researched for years, e-commerce product categorization differs from classical ones. This is due to the vast volume of products with noisy and incomplete signals in both product description and categorical labels. Besides, subjective customer opinions about multi-functional products add the complexity, as these opinions can influence product descriptions and optimal category assignment. It is non-trivial to train machine learning models to discern consistent categorical patterns that meet customer expectations for a large population of catalog.

We approach product categorization as a text classification problem, since most product items in e-commerce platform are represented through structured or unstructured textual features. Recently, pre-trained models (PTMs) have shown substantial benefits in capturing universal language representations and strong reasoning capability with RLHF (Ziegler et al., 2019; Lampinen et al., 2022). Two prominent PTM frameworks are: 1) discriminative models with the encoder structure, like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018; Conneau and Lample, 2019; Conneau et al., 2019), and 2) generative models with the decoder structure, like OpenAI's GPT series (Generative Pre-trained Transformer) (Radford et al., 2018, 2019; Brown

*Equal contribution.

et al., 2020; Ouyang et al., 2022). Though some efforts have unified discriminative and generative tasks within a single framework, discriminative language models are generally preferred for sentence understanding, while generative language models are more commonly used for text generation and reasoning. With the increasing parameter sizes and extensive pre-training on vast datasets and various learning tasks, these language models have consistently attained state-of-the-art (SOTA) performance across numerous NLP benchmarks. Given the overlap between pretrained knowledge and e-commerce catalog, we believe that PTMs possess the domain knowledge that is necessary to differentiate the nuances between categories.

In this study, we introduce an innovative dual-expert framework that integrates both discriminative and generative large language models (LLM) in a cascading approach to achieve precise product categorization. Initially, the discriminative language model is fine-tuned with domain data, acting as a domain expert to recommend the top K candidate classes for each product. Subsequently, an off-the-shelf LLM selects the optimal target from the top K suggestions based on certain criteria via prompting. The LLM in our framework serves as the general expert due to its capability acquired through pre-training on a large corpus of general data and well alignment with human instructions. The major contribution of this study can be summarized in 3 folds:

- 1) We propose a novel LLM-based dual-expert categorization system, which is designed to achieve accurate product classification in e-commerce and output reasons for hard cases.
- 2) We introduce the key components of domain-specific and general experts, and describe the strategies to inject domain knowledge into the decision-making process of each expert.
- 3) We compare the performance of this dual-expert framework against the popular text classification models as well as the SOTA model in two e-commerce catalog datasets, proved its superiority on e-commerce categorization.

2 E-commerce Product Categorization

In e-commerce, product categorization involves assigning one or more optimal categories from thousands of labels based on product features. This task is challenging due to noisy and incomplete catalog data. E-commerce sites generally define a taxonomy (a hierarchical structure) as the target label space for categorization. As this taxonomy becomes more granular, categories can become very similar, with only subtle differences distinguishing them.

Output Label Space Online e-commerce sites pre-define the semantic structure of item categories (known as taxonomy) according to business purpose. This taxonomy serves as the target label space for categorization, and is constructed as hierarchical trees. As the taxonomy tree becomes fine and granular, the categories may appear similar to each other, with only subtle differences separating them. Extreme multi-label text classification aims to identify relevant labels from an extremely large set of labels, making it a challenging task (Zhang et al., 2021a; Chang et al., 2020). Accuracy of classification models can vary depending on the complexity and dimensionality of the label space. Additionally, catalog data inherently suffers from label imbalance, which is widely known as the long tail issue. Classification models may struggle to learn patterns for the underrepresented, smaller categories in the skewed distribution.

Catalog Noise and Incompleteness The training data for our ML-based categorization model is mainly derived from samples of catalog data, which often includes noisy labels and incomplete information. A key challenge for e-commerce categorization systems is to extract meaningful signals about customer preferences from this low-quality data. We classify the quality of the model training data into two types:

- **Noise Signals.** Item features and labels often contain noise, leading to unstable learning. This noise can be soft (exaggerated properties) or hard (misleading/irrelevant descriptions) and is common in popular categories. Meanwhile, label assignments can be noisy due to outdated categorization systems, internal biased corrections, and incorrect label suggestions from sellers. These are the major sources of label noise.
- **Incomplete Information.** Incomplete information often arises from the subjective opinions of sellers and customers. For instance, sellers in an automobile store might omit

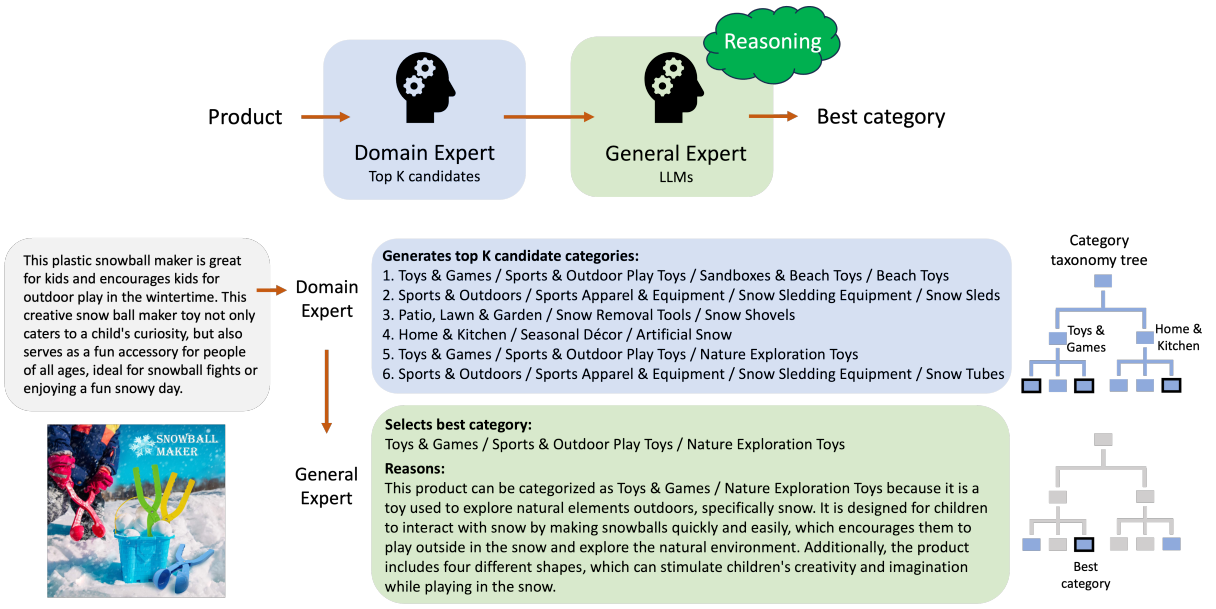

Figure 1: LLM-based Dual-Expert e-commerce categorization framework. The system comprises two key components that operate sequentially: a Domain Expert that identifies the top K categories, followed by a General Expert that decides the optimal category from the top candidates by applying reasoning. We inject the domain knowledge to each expert through model fine-tuning (domain expert) and prompting (LLM-based general expert).

keywords and only provide brand and series numbers, resulting in very brief item descriptions. This limited information confuses general buyers. Additionally, catalog labels are incomplete because selling items may be multi-functional, yet sellers typically provide only a single label which may not align with how different buyers perceive or intend to use the product. In this scenario, our task is to find the most favored category, even when multiple options are acceptable.

To overcome the issues, we propose a novel LLM based Dual-Expert approach for product classification.

3 LLM Based Dual-Expert System

LLM-based multi-agent systems have emerged as a novel technology with advanced capabilities. These systems specialize LLMs into various distinct agents, each with different expertise (Wu et al., 2023; Qian et al., 2024; Yue et al., 2024). Our domain-specific and general expert system has two language models cooperating with each other and each has a specialty. Specifically, we have designed two expert models that work sequentially to assign the optimal category to a given product. The whole pipeline is shown in Figure 1. First, a discriminative model work as the domain expert to find top K candidate categories for the selling product given its item data. Then, an off-the-shelf LLM serves as the general expert, evaluating which categories from the top K candidates are most suitable and accurate for the selling product in question. The LLM outputs its decision and the reasoning behind its selection.

3.1 Domain Expert

The primary objective of the domain expert is to identify the top K most relevant leaf categories for a given product, with relevance determined by similar patterns observed in the training data. Simultaneously, the domain expert ensures a highly accurate top 1 prediction to support the online inference pipeline. The backbone of the domain expert is XLM-R (Conneau et al., 2019), a Transformer model that is pre-trained on monolingual data using the multilingual masked language modeling (MLM) objective.

3.1.1 Label Semantic Capture via Label Augmentation

Discriminative models face a limitation in explicitly lacking semantic knowledge about the labels. Our in-depth study observed a high frequency of label-related keywords in the item data written by sellers, indicating that keyword matching could benefit semantic understanding in our domain tasks. Therefore, we strategically expose the label names to the model, aiding its few-shot and zero-shot

learning capabilities. To enhance the training data with label names, we use the full path of labels, i.e., a path in a taxonomy tree. We randomly mask the branch along this path and replace the title or description of sampled training data with the masked path (Figure 2). These synthetic training samples are then added to the original data.

3.1.2 Two-phase Learning

Learning from large, noisy catalog data is difficult due to label imbalance and errors in signals. To tackle this, we split model training into two phases. In the first phase, the domain expert reviews challenging cases and uses focal loss to handle imbalance. In the second phase, the model focuses on major patterns, reinforcing the initial phase with bootstrap loss. Further details are in the following sections.

Phase 1: Exploration of Category Relationship The catalog data inherently suffers from label imbalance, commonly referred to as the long tail issue. To address this, we incorporate focal loss (Lin et al., 2017) into our objectives as a dynamic learning approach to better capture challenging cases in smaller categories. The mathematics definition of focal loss for classification can be defined as:

$$L_{FL} = -\sum_{k=1}^{N} \alpha_k \left(1 - q_k\right)^{\gamma} \log\left(q_k\right), \qquad (1)$$

where q_k is the predicted probability of the true label k by model. α_k is the corresponding class weight of the true label. It is predefined based on the desired label distribution, e.g., popularity score of the product in catalog. γ is a hyperparameter controlling the learning weight of hard examples. The higher the value of γ, the lower the loss for well-classified examples.

Phase 2: Self-Exploitary The second phase of training employs a self-justifying learning mechanism that accounts for knowledge consistency during training (Reed et al., 2014). It augments the usual prediction objective with a notion of perceptual consistency, which allows the model to disagree with a perceptually-inconsistent training label and effectively relabel the data while training. The assumption behind this idea is that incorrect labels are likely to be eventually highly inconsistent with other data points predicted to the same label by the model. Therefore, it acts in a manner of self label clean-up and bootstraps itself until convergence to stable knowledge. Here, we incorporate this idea into the cross-entropy training loss:

$$L_{BT\_BCE}(p, q) = -\sum_{k=1}^{N} \beta p_k log(q_k) + \beta(1 - p_k) log(1 - q_k) + \sum_{k=1}^{N} (1 - \beta) q_k log(q_k), \qquad (2)$$

where p_k, q_K are ground truth label and model prediction, respectively. N is the size of target labels. Parameter $0 \leq \beta \leq 1$ balances bootstrap learning and supervised classification. It is empirically set in the range $[0.8, 0.95]$. Due to the large batch training steps (t_{batch}), we can set a delta activation $\hat{\beta}$ that adaptively turns on/off the bootstrap loss at a given global step T_{gate}:

$$\hat{\beta} = \begin{cases} 1, & \text{if } t_{batch} < T_{gate} \\ \beta, & \text{if } t_{batch} \geq T_{gate} \end{cases} \qquad (3)$$

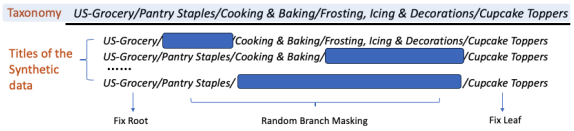

Figure 2: Example of synthetic data for capturing label semantics.

3.2 General Expert

After the Domain Expert produces top K candidate categories, the LLM-based General Expert then reasons about the top candidate categories via proper prompting strategies, and selects optimal category among the candidates.

3.2.1 Zero-shot

Product category names often carry rich semantic meaning. For instance, hierarchical path of category "Toys and Games / Sports & Outdoor Play Toys / Sandboxes & Beach Toys / Beach Toys" self explains that "Beach Toys" is for outdoor play and is under Toys and Games department. Thus, we directly prompt the LLM-based General Expert with product item data and candidate categories' path names.

3.2.2 In-context learning via LLM self-generated summarization

LLMs demonstrate remarkable capabilities in in-context learning (ICL), they can learn to do a specific task by conditioning on a prompt consisting of

input-output examples (Brown et al., 2020). LLMs can generalize to previously unseen data by using few-shot examples provided in the prompt, without explicit pre-training for the specific task (Xie et al., 2021). ICL are recently used in text classification(Milios et al., 2023; Zhu and Zamani, 2023; Simig et al., 2022; D'Oosterlinck et al., 2024a).

In e-commerce product categorization task, there are a vast number of different categories in the taxonomy tree, each with numerous products associated with it. In the traditional approach of few-shot in-context learning, we need to select example products for each candidate category in the prompt. However, the selected products may contain information irrelevant to the candidate category, and may not adequately represent the candidate category.

To address these issues, we propose a novel in-context learning approach. Rather than providing a few products and their associated categories as few-shot examples in the prompts, we provide clear definitions of the candidate categories to the LLM-based General Expert, where the category definition is self-generated by LLMs. The self-generation process is as follows. For each category, we curated a collection of data points that have been previously labeled as belonging to that particular category, then LLMs were instructed to summarize from the pool of data and generate a clear definition for the category based on the provided data. To ensure diversity in the summarizing samples, we include multi-source data from both popular selling products and catalog representatives of each category via unsupervised learning. Consequently, a summarized definition of each category was self-generated by LLMs. We then feed these LLM-generated category definitions to the LLM-based General Expert, aiding in more accurate category selection (Figure 3b, Figure 6).

(a) *Zero-shot:*
You will try to classify a product in the catalog of an e-commerce site to a product category. Below is the data for the product: {product information}
Can this product be categorized to any of the categories listed below?
Category 0: …, Category 1: …, Category 2: …
If this product belongs to none of them, answer none. If the product information is not provided enough for you to make judgement, answer "cannot decide". If this product is multi-functional, answer the categories with its primary function.

(b) *Representation of the categories:*
- Leaf node only: Category 0: Shorts, …
- Full path: Category 0: Fashion / Women / Clothing / Active / Base Layers & Compression / Shorts, …
- Descriptive name: Category 0: Women's Base layers & Compression Shorts, …
- Descriptive name with In-Context Learning via LLM self-generated summarization/definition: Category 0: Women's Base layers & Compression Shorts, Women's Base Layer Shorts are designed for various activities such as gym workouts, cycling, running. They are typically made of stretchy, breathable, and moisture-wicking materials that provide support, comfort, and flexibility. They are shorts or tights that cover the upper part of the legs …

(c) *Enhanced reasoning via ranking:* This product can be possibly categorized to the categories listed below. Can you rank the relevant categories for this product in order of likelihood, starting with the most probable and ending with the least probable.

(d) *Enhanced reasoning via CoT and ranking:* Let's think step by step. First, find clues in the product data sources and candidate category data, then think of who may be the users of this product and under what scenarios will this product be used, finally rank the candidate categories. Your thinking process should be in the <reasons> section.

Figure 3: Prompting strategies. (a) LLM is prompted to directly select an optimal category. (b) Categories are represented by various levels of information, including in-context learning via summarization. (c) LLM is enforced to rank in order to reason. (d) LLM is encouraged to execute CoT before ranking.

3.2.3 Enhanced reasoning

To boost LLM's decision-making capabilities, we employed prompts that are designed to enhance the reasoning processes within LLMs. We instructed LLMs to identify the categories that match the main functionality or intended usage of the product (Figure 3a). A product category consists of a root level node (typically a Department) and intermediate nodes, followed by a fine-grained leaf node. We experiment with prompts containing various levels of information from the categories (Figure 3b).

Think step by step enables LLMs to generate task reasoning processes (Kojima et al., 2022). Chain-of-thought (CoT) prompting significantly enhances reasoning abilities of LLMs through chained reasoning steps (Wei et al., 2022, 2021). CoT prompting, which involves the presentation of intermediate reasoning steps, has proven effective in zero-shot (Kojima et al., 2022) and in-context learning (Wei et al., 2022) settings. To enhance LLM's reasoning capability on product classification task, we instructed LLMs to rank the relevant candidate categories from the most likely to the least likely for a given product (Figure 3c). Furthermore, LLM is encouraged to find clues in the product item data, think of a potential user and a use case for the product, then finally proceed to perform the ranking task (Figure 3d).

4 Experiments

4.1 Dataset

We evaluate our Dual-Expert framework on two benchmark datasets.

RetailProducts2023. This dataset contains 95,526 products that potentially belong to 2,214 categories from an E-commerce site. The dataset contains categories that have limited number of data entries. Each category has at least 10 associated data points to guarantee sufficient data for training and testing.

E-commerceCatalog. For curating this data, we select the e-commerce catalog data of 3 locales in different languages to assess the robustness of our

Table 1: Model performance on RetailProducts2023 dataset.

	Precision*	Recall*	F1 score*	F1 score (macro)
fastText	0.857	0.837	0.836	0.716
BERT	0.901	0.890	0.891	0.779
XLM-R	0.902	0.910	0.899	0.782
Domain Expert alone	0.925	0.929	0.921	0.825
Dual-Expert	**0.972**	**0.969**	**0.968**	**0.925**

*Weighted average.

Table 2: Classification accuracy on the E-commerceCatalog dataset.

	Locale 1	Locale 2	Locale 3
DHPC (Zhang et al., 2021b) (baseline)	+0%	+0%	+0%
Domain Expert alone	+1.01%	+1.33%	+1.57%
Domain Expert w/ XLM-R Selector	+1.12%	+1.05%	+1.31%
Dual-Expert	**+3.81%**	**+4.01%**	**+3.14%**

dual-expert approaches. In each locale, we collect an evaluation dataset of 10K products. This dataset was curated through multiple iterations of human review to provide a fair evaluation of all models compared. The Domain Expert is fine-tuned on millions of sampled catalog data per locale and we pick $K = 10$ as the number of suggested candidate categories for the LLM-based General Expert. The SOTA model Deep Hierarchical Product Classifier (DHPC) (Zhang et al., 2021b) is used as the baseline for comparison.

We leverage Mixtral from mistral.ai, a high-quality sparse mixture of experts model (SMoE) as the General Expert. Unless otherwise stated, we perform experiments with a temperature of 0.

4.2 Results

4.2.1 Dual-Expert model achieves better classification performance compared to the baseline

The results indicate that Dual-Expert model achieves higher classification performance consistently across RetailProducts2023 and E-commerceCatalog datasets compared to baseline models (Tables 1 and 2). On the RetailProducts2023 dataset, many categories have limited number of data points, consequently, vanilla XLM-R models exhibit poor performance on these minority classes, as evidenced by the significantly lower macro F1 score of 0.782, when compared to our Dual-Expert model (0.925). Similarly, fastText (Joulin et al., 2017) and BERT models exhibit relatively poor performance (Table 1). The Domain Expert model, which is a specialized version of XLM-R, has improved classification performance, although it requires relatively large amount of training data to accurately learn and distinguish between different categories. The Dual-Expert model demonstrates generalization capabilities on minority classes, showcasing its remarkable zero-shot and few-shot capabilities (Table 1). This is powered by the extensive knowledge gained during pretraining and alignment stages of the LLMs.

On the E-commerceCatalog dataset (Table 2), Dual-Expert model demonstrates significant accuracy improvement in 3 locales compared to the baseline SOTA model DHPC and Domain-specific Expert alone (Table 2). These results demonstrate that collaboration between the two experts, where the Domain Expert provides relevant categories and the LLM-based General Expert applies its reasoning capability to distinguish among categories and select the optimal one, leads to increased classification performance. Of note, we trained a XLM-R based binary classification model that makes binary predictions for (product, category) pairs. We used this model as a selector, substituting the General Expert. The overall accuracy was comparable or inferior to Domain Expert, suggesting these models likely learned the same noise in the training data.

Dual-Expert achieves higher classification accuracy partially due to its ability to address noisy mislabeled data in the training set. Consider the product shown in Figure 1, there are snowball clippers that are incorrectly labeled as beach toys in training data, a BERT-based discriminative model would learn this inaccurate classification during fine-tuning. In contrast, LLMs have the extensive general knowledge to recognize that such product is not a beach toy, but rather a snow exploration toy. Consequently, this approach effectively mitigates the issue of incorrect labeling in training data.

4.2.2 Impact of domain expert training strategies

We conduct ablation study to assess the impact of removing the proposed components of domain expert's training strategies. As shown in Table 3, removing any of these strategies causes performance drop. The bootstrap learning in phase 2 has the most significant impact on the accuracy of domain-expert's top1 prediction, as it stabilizes the later stages of model training and prevent over-fitting. For the entire dual-expert system, label augmentation and phase 1 training play a more crucial role than phase 2 since they enhance model's learning from the few-shot knowledge and improves topK

retrieval performance of the domain expert.

4.2.3 Clear category definitions through LLM self-generated summarization enhance Dual-Expert's decision-making capabilities

Table 4 summarizes Dual-Expert's performance when using prompts that provide clear category definitions and enhance its reasoning capabilities. We observed that the prompts employing short phrases to represent categories achieved relatively low classification accuracy (Table 4, with ambiguous category definition). This is expected, as short phrases encode limited category information. For example, 'accessory' as a category name is ambiguous, therefore LLM misunderstands the category and makes errors.

To make the category definitions more clear, we propose a novel in-context learning approach via LLM self-generated summarization. For each category, we first instructed LLMs to summarize from the pool of data and generate a clear definition for the category based on the provided data. Then, instead of providing products and their associated categories as few-shot examples directly in the prompts, we provide the LLM with self-generated category summary, and instruct the LLM to select the most appropriate category among the candidates. As a result, the Dual-Expert model achieves the highest classification accuracy improvement of 3.8%, 4.0%, 3.1% for the 3 locales, respectively (Table 4, descriptive category name with ICL summarization). The findings suggest that LLMs excel at summarizing the core characteristics of a particular category. By leveraging the summarizations of categories generated by LLMs themselves, the models are equipped with more precise and well-defined descriptions of the categories, enabling them to make more accurate classification predictions (Figure 6).

4.2.4 Classification accuracy of the LLM-based Dual-Expert improves via enhanced reasoning

Our baseline prompting strategy involves instructing the LLM to directly choose optimal category from candidate classes (Table 4, zero-shot). LLM often states that "category A is correct" and that "categories B, C, and D are incorrect" without further explanations and reasoning. LLMs likely did not engage in extensive reasoning, classification accuracy was relatively low. When prompted to rank all relevant candidate categories in descending order, from the most likely to the least likely, LLM enhanced its reasoning capabilities. As a result, we observe classification accuracy improvement by 3.85%, 3.55% and 2.26% in the 3 locales, respectively (Table 4).

Table 3: Ablation Study: Impact of training strategies on Domain Expert's classification accuracy, i.e. Label Augmentation (LA), Phase 1&2 training.

Training Method	Domain Expert Acc (top 1)	Dual-Expert Acc (top k -> 1)
Domain Expert w/o LA	-0.7%	-1.5%
Domain Expert w/o Phase1	-0.2%	-0.5%
Domain Expert w/o Phase2	-1.4%	-0.25%

Table 4: Comparison of LLM prompting strategies on Dual-Expert's classification accuracy. Baseline is DHPC (Zhang et al., 2021b)

Prompt Strategy	Locale 1	Locale 2	Locale 3
with ambiguous category definition	+0.85%	+0.53%	-0.68%
descriptive category name	+1.23%	+0.80%	+2.06%
descriptive category name with ICL summarization*	**+3.81%**	**+4.01%**	**+3.14%**
enhanced reasoning via rank	**+3.85%**	+3.55%	+2.26%
enhanced reasoning via CoT and rank	+3.32%	**+3.59%**	**+2.86%**

*Proposed prompt strategy in Dual-Expert

4.3 Discussion

4.3.1 Inference cost

Inference cost is crucial for the practical application of this work to large-scale e-commerce product categorization. Consider the online/offline traffic of practical categorization system, we utilized the thresholding within the domain expert to regulate the traffic flowing into the general expert. In our practice, this approach reduces total traffic by 80% while maintaining overall accuracy improvements, as the 20% of data that passed through the entire workflow are typically cases the Domain Expert alone struggles to classify correctly (Figure 4). Furthermore, the Dual-Expert system (Table 1), in return, can act as a reliable auditor for determining the appropriate threshold for the Domain Expert model, further dynamically optimizing the trade-off between performance and computational cost.

4.3.2 Probing framework feasibility

From our experiments, we found that for classification tasks with fine-grained categories and limited number of data points per category, LLMs demonstrate robust zero-shot and few-shot capabilities. As shown in Figure 5, when minimum number of data points per category is small, Dual-Expert outperforms the Domain Expert with larger margin. E-commerce categorization task falls under this regime, since catalog data inherently exhibits long

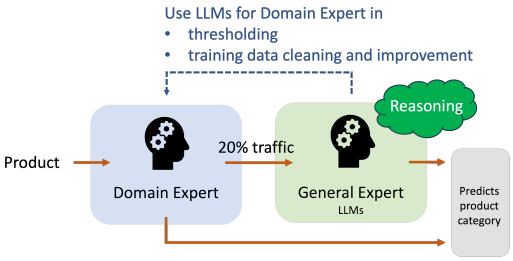

Figure 4: Modified framework that utilizes both Dual-Expert and Domain Expert alone for large scale applicability.

tail distribution, and the categories are fine-grained with subtle differences, such example categories are shown in Figure 6. As the categories become larger with sufficient amount of training data per category, and categories are well-separated with no conceptual overlap or nuanced difference, discriminative classification models tend to provide on-par classification performance compared to the LLM-based Dual-Expert (Figure 5).

5 Related Work

When the label space is vast with thousands of labels, a typical approach towards classification based on ICL is reducing the label space by identifying most relevant candidates for a given input (Milios et al., 2023). In this regards, research community has worked with both generative and non-generative techniques to narrow down to most relevant labels. Simig et al. (2022) explored generating candidate labels in the setting where task involves classification in unseen labels. Zhu and Zamani (2023) uses a set of labels and map the LLM generated candidates to actual labels by using semantic similarity. D'Oosterlinck et al. (2024b) takes a step further and ranks the retrieved labels by using an additional LLM. Semantic similarity works well when there is direct mapping between input and output. In our work, we target e-commerce data where the direct mapping between input to leaf categories does not work because a large number of leaf categories can have semantically similar definition which defeats the purpose of classifying the product in a single leaf category. Further, using multiple LLMs and making several calls to them is expensive. We reduce that cost by using only one LLM that processes the relevant labels selected by a non-generative model. In the non-generative approaches, Jain et al. (2019) considered building an approximate nearest neighbor (ANN) graph as an indexing structure over the labels by relying on sparse features engineered from the text. The relevant labels for a given text were then found quickly from the nearest neighbors of the instance via the ANN graph. With the introduction of PLMs, classification performance on several tasks improved significantly through PLMs' ability of learning better text representation from the raw, unstructured text. In our work, we explore LLM's capability for classification in different situations that occur in e-commerce domain - when product text is noisy, and when classification labels are fine-grained and conceptually overlapping. Each situation has their own challenges. We show that the Dual-Expert paradigm overcomes these challenges and outperforms the discriminative classification model in selecting the optimal category. We also show that enhancing the LLM prompt with self-generated summarization outperforms other prompt-tuning techniques experimented in this paper.

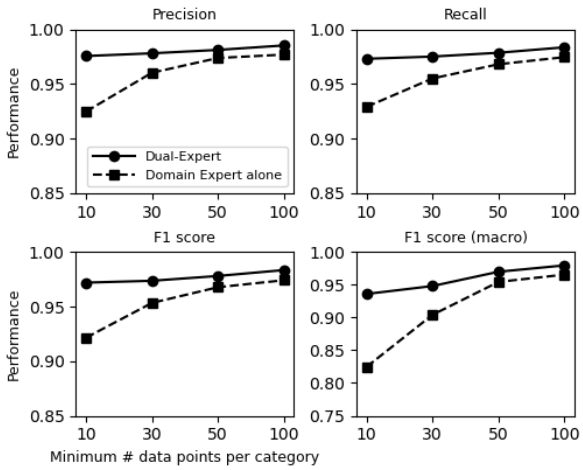

Figure 5: Framework feasibility on RetailProducts2023.

6 Conclusion

In this study, we propose a Dual-Expert classification workflow, which leverages the pre-trained LLMs for accurate e-commerce product categorization. It comprises two experts: a domain-specific expert, trained on a large e-commerce domain data, identifies relevant candidate classes; and a general expert, powered by a LLM with In-Context Learning, that handles nuanced reasoning and decision-making. This dual-expert architecture leverages the complementary strengths of each expert, blending specialized domain knowledge with general reasoning capabilities from pre-training, to achieve high classification accuracy in e-commerce categorization.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3163–3171.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Karel D'Oosterlinck, Omar Khattab, François Remy, Thomas Demeester, Chris Develder, and Christopher Potts. 2024a. In-context learning for extreme multi-label classification. *arXiv preprint arXiv:2401.12178*.

Karel D'Oosterlinck, Omar Khattab, François Remy, Thomas Demeester, Chris Develder, and Christopher Potts. 2024b. In-context learning for extreme multi-label classification. *arXiv preprint arXiv:2401.12178*.

Himanshu Jain, Venkatesh Balasubramanian, Bhanu Chunduri, and Manik Varma. 2019. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 528–536.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Daniel Simig, Fabio Petroni, Pouya Yanki, Kashyap Popat, Christina Du, Sebastian Riedel, and Majid Yazdani. 2022. Open vocabulary extreme classification using generative models. *arXiv preprint arXiv:2205.05812*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,

et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. 2024. Large language model cascades with mixture of thought representations for cost-efficient reasoning.

Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit Dhillon. 2021a. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 34:7267–7280.

Wen Zhang, Yanbin Lu, Bella Dubrov, Zhi Xu, Shang Shang, and Emilio Maldonado. 2021b. Deep hierarchical product classification based on pre-trained multilingual knowledge. *IEEE - The Bulletin of the Technical Committee on Data Engineering*.

Yaxin Zhu and Hamed Zamani. 2023. Icxml: An in-context learning framework for zero-shot extreme multi-label classification. *arXiv preprint arXiv:2311.09649*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Appendix

A.1 Example that showcases the effectiveness of in-context learning via LLM self-generated summarization/definition

LLM self-generated summarization/definition of the categories helps clarify the category definitions to the LLM, which in turn improves LLM's product categorization capability (Figure 6).

Black Glitter Level 16 Unlocked Game Birthday Cake Topper - Boy's 16th Birthday Cake Decorations. The cake toppers are approx 5.9” wide and 5.6” high. This unique decoration is the perfect centerpiece for a truly epic birthday bash. The bottom of the cake topper is made of wooden sticks, please pay attention to your child to avoid accidents.

Candidate categories produced by Domain Expert that have nuanced difference:
- Grocery / Pantry Staples / Cooking & Baking Grocery Supplies / Frosting, Icing & Decorations / Cake Toppers (denoted as cake toppers 1)
- Toys & Games / Kids' Party Supplies / Kids' Party Decorations / Cake & Cupcake Toppers / Cake Toppers (denoted as cake toppers 2)
- ...

LLM self-generated summarization/definition:
- Cake toppers 1: Cake toppers are ***edible*** decorative elements used to decorate cakes. They are typically made from sweet ingredients like sugar, chocolate, or marzipan. These decorative pieces are placed directly on top of the cake.
- Cake toppers 2: Cake toppers are ***non-edible*** decorative elements used to embellish cakes. They are typically made of materials like plastic or cardboard and fashioned into various figures or shapes. These decorative pieces are usually made of plastic or made of cardboard attached to toothpicks.

Optimal category selected by the LLM-based General Expert:
Cake toppers 2.

Reasoning:
Based on the product information, this product is made of glittery card-stock with wooden stick to insert into the cake. The dimensions of 5.9" wide and 5.6" high are suitable for most cake sizes. All these details indicate that this product is meant to decorate cakes. This product is a ***non-edible*** cake topper made of glittery card-stock and wooden sticks. Therefore, it belongs to the category of ***non-edible*** decorative cake toppers.

Figure 6: An example product that showcases the effectiveness of LLM self-generated summarization/definition. Two candidate categories are similar with subtle differences. LLM self-generated summarization identifies the key features that separate them: one is edible cake topper, the other is non-edible.