Label-Free Distinctiveness: Building a Continuous Trademark Scale via Synthetic Anchors

Huihui Xu, Kevin Ashley
Learning Research and Development Center, School of Computing and Information
University of Pittsburgh
Pittsburgh, PA, USA
{huihui.xu, ashley}@pitt.edu

Abstract

Trademark law protects distinctive marks that are able to identify and distinguish goods or services. The *Abercrombie* spectrum classifies marks from generic to fanciful based on distinctiveness. The Abercrombie spectrum employs hard buckets while the real world of branding rarely falls into neat bins: marks often hover at the blurry border between "descriptive" and "suggestive" for example. By requiring trademark examiners or researchers to pick one of the five buckets, one loses useful information where the lines get blurry. So hard boundaries obscure valuable gradations of meaning. In this work, we explore creating a continuous ruler of distinctiveness as a complementary diagnostic tool to the original buckets. The result is a label-free ladder, where every mark, real or synthetic, gets a real-valued score. These continuous scores reveal subtle distinctions among marks and provide interpretable visualizations that help practitioners understand where a mark falls relative to established anchors. Testing with 95 expert-classified trademark examples achieves a Spearman's $\rho = 0.718$ and Pearson's $r = 0.724$ against human labels, while offering intuitive visualizations on the continuous spectrum. A demo can be found at https://distinctiveness-ruler-demo.streamlit.app/.

1 Introduction

A trademark is a word, symbol, or other identifier that distinguishes a company's goods or services from those of others (Landes and Posner, 1987). For example, "Coca-Cola" serves as a distinctive mark that identifies beverages produced by The Coca-Cola Company, setting them apart from other competing products. As a form of intellectual property, the distinctive nature of an owner's trademark enables consumers to identify the source of goods or services and establishes economic values through brand recognition, customer loyalty, and market differentiation (Landes and Posner, 1987; Dogan and Lemley, 2006).

Assessing distinctiveness is a fundamental task in trademark law. The more distinctive a trademark is, the stronger its legal protection and the greater its potential for granting economic benefits. A trademark's distinctiveness, its ability to signal source and stand apart from other marks, needs to be assessed.

In U.S. trademark law, the assessment is guided by the *Abercrombie* spectrum, a framework introduced in *Abercrombie & Fitch Co. v. Hunting World Inc.*, 537 F.2d 4 (2nd Cir. 1976), categorizing trademarks into varies degrees of protection: generic, descriptive, suggestive, arbitrary, and fanciful. **Generic** terms employ the common name of the product and receive no protection. **Descriptive** marks describe a product feature and require secondary meaning[1] to qualify. **Suggestive** marks imply qualities and are inherently distinctive. **Arbitrary** marks use common words in unrelated contexts and are strongly protected. **Fanciful** marks are invented terms with the highest level of protection.

While the Abercrombie spectrum provides a conceptual framework for assessing a trademark's distinctiveness, it poses hard categorical boundaries on what is inherently a context-dependent, continuous property. In practice, some marks might fall into gray areas between categories, and human judgment can vary (Ouellette, 2014). This presents a challenge for assessing consistently, especially with respect to edge cases.

This challenge motivates the need for a more continuous, interpretable scale of distinctiveness. Instead of hard labels, we propose to leverage synthetic anchors to build a spectrum using a **Bradley-Terry** (BT) model. Real marks can be place along

[1] Secondary meaning is a connection in the public's mind between a mark and a source of goods caused by extensive use and promotion.

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 113–124

the spectrum, potentially enabling a more consistent and data-driven assessment that amplifies the nuance nature of distinctiveness.

Our contributions are three-fold:

1. Continuous Distinctiveness Scaler: We propose a method to model trademark distinctiveness as a continuous spectrum rather than in terms of discrete buckets.

2. Label-Free Ranking via Bradley-Terry: We apply a Bradley-Terry (BT) model to derive distinctiveness scores for real marks without requiring human labeling.

3. Interpretability and Robustness: We show that the resulting scale is interpretable, and robust across different metrics.

2 Related Work

2.1 Trademark Classification

Prior research has approached trademark distinctiveness as a multiclass classification task aligned with the Abercrombie spectrum. Goodhue and Wei (2023) explored whether a large language model like GPT-3.5 can effectively classify trademarks along the spectrum. Guha et al. (2023) (LegalBench) introduced series of benchmark tasks for evaluating legal reasoning, including a dataset on trademark distinctiveness based on the Abercrombie spectrum. Adarsh et al. (2024) used the USPTO Trademark Case Files Dataset (Graham et al., 2013) as a major resource and trained BERT-based models to predict distinctiveness outcomes. Previously mentioned works mostly rely on supervised learning with labeled trademark corpora, but high-quality annotated datasets are scarce due to legal ambiguity and the difficulty of drawing clear boundaries between distinctiveness categories. In contrast, our method avoids manual labeling by leveraging synthetic anchors and pairwise comparisons. It offers an alternative way of measuring legal concepts that lack clear categorical boundaries.

2.2 Legal Synthetic Data

Due to the scarcity of annotated legal data and the high cost of expert labeling, several attempts have been made to generate synthetic data for legal NLP tasks. Perçin et al. (2022) proposes a method of substituting phrases through WordNet and word embeddings. Ghosh et al. (2023) presents a framework that uses selective masking strategies tailored to legal documents' structured language to produce diverse and coherent synthetic samples. Xu and Ashley (2023) generated legal question-answer pairs using LLMs based on human-written summaries for evaluating the quality of machine-generated summaries. Zhou et al. (2025) introduces a knowledge-guided approach for legal question-answer generation. The synthetic data was used to train a legal LLM, which achieve comparable performance to proprietary LLMs. Whereas most synthetic data generation methods aim to replicate labels for classification tasks, we instead use synthetic data to define a ranking structure that enables label-free inference through pairwise comparisons.

2.3 Pairwise Ranking with the Bradley–Terry Model

The Bradley-Terry (BT) Model is a probability model that is frequently used for determining the relative "strength" of an object via pairwise comparisons (Bradley and Terry, 1952). The model estimates the probability that the pairwise comparison of a pair of items i and j draws from some distribution. It can be represented as

$$Pr(i > j) = \frac{p_i}{p_i + p_j} \tag{1}$$

where p_i represents the underlying strength score of item i, and $\Pr(i > j)$ denotes the probability that i is preferred over j. One of the common score functions is defined as $p_i = e^{\theta_i}$ and the Equation 1 can be parameterized as

$$Pr(i > j) = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}} \tag{2}$$

where θ can be estimated by maximizing the likelihood of oberved comparison outcomes.

The BT model has been used in several NLP tasks. Luo et al. (2022) develop a method of evaluating summary quality by leveraging a BT model to turn pairwise preferences into a continuous quality score. In the RLHF setting, methods like Direct Preference Optimization (DPO) (Rafailov et al., 2023) adopt the BT setting to human preferences when learning a reward function from pairwise comparisons.

Compared to these approaches, our work uses BT not to train a model but to construct an interpretable continuous "distinctiveness" ruler from synthetic data. This enables interpreting real-world trademarks' distinctiveness without hard labeling. In trademark disputes, courts often rely on survey evidence to assess public perception of a mark's

distinctiveness (Ouellette, 2014). However, such surveys are costly, subjective, and not easily reproducible. Our method could provide a scalable and reproducible alternative by simulating comparative judgments and fitting them into a BT-based scoring framework.

3 Problem Formulation

3.1 Task Definition

Prior work relies on hard labels to classify trademarks under the Abercrombie spectrum. We aim to construct a continuous measure of distinctiveness.

Let

$$\mathcal{R} = \{(r_1, d_1), (r_2, d_2), \ldots, (r_N, d_N)\}$$

denote the set of N **real trademarks**, where r_i is the *mark text* (e.g., "Salt") and d_i is its *real-world product or service description*.

$$\mathcal{S} = \{(s_1, \tilde{d}_1, c_1), (s_2, \tilde{d}_2, c_2), \ldots, (s_M, \tilde{d}_M, c_M)\}$$

denote the set of M **synthetic anchors**, where each s_j is a *mark text* (taken from r_i), $\tilde{d}_j$ is a *synthetic product or service description* generated to represent a different level of distinctiveness, and $c_j \in \{\text{generic}, \text{descriptive}, \text{suggestive}, \text{arbitrary}, \text{fanciful}\}$ is the known *Abercrombie spectrum*.

Each real trademark (r_i, d_i) is compared against a subset of synthetic anchors $(s_j, \tilde{d}_j, c_j)$ to simulate pairwise judgments. We record a binary outcome:

$$y_{ij} = \begin{cases} 1, & \text{if } (r_i, d_i) \text{ is more distinctive than,} \\ & (s_j, \tilde{d}_j) \\ 0, & \text{otherwise.} \end{cases}$$

In Equation 2, θ_j represents BT values for synthetic anchors and θ_i is learned from the comparisons for a real mark. The resulting θ_i provides a continuous measurement of where each real trademark stands on the distinctiveness spectrum.

3.2 Synthetic Anchor Generation

We first use regular expressions to extract the mark name and goods/services from each real trademark example. To construct synthetic anchors $\mathcal{S}$, we fix the extracted mark name and prompt an LLM to generate alternative goods/services descriptions that systematically cover all five Abercrombie distinctiveness categories. This ensures the pairwise comparisons span all the spectrum. One example is listed below:

Consider a real trademark $(r_i, d_i) =$ (*Salt*, "packages of sodium chloride."). Synthetic anchors $(s_j, \tilde{d}_j, c_j)$ are generated by fixing the mark *Salt* and varying the goods/services description to target each Abercrombie category, e.g.:

- ***Generic:*** (*Salt*, "a brand of table salt substitutes")
- ***Descriptive:*** (*Salt*, "a skincare line emphasizing natural ingredients")
- ***Suggestive:*** (*Salt*, "a brand of ocean-themed clothing")
- ***Arbitrary:*** (*Salt*, "a tech startup offering cloud storage solutions")
- ***Fanciful:*** (*Salt*, "a line of energy drinks")

4 Setup

We conduct experiments to assess whether synthetic anchor comparisons can produce a meaningful distinctiveness spectrum for real trademarks. Our evaluation focuses on (1) the correlation between model-derived scores and expected trademark distinctiveness, and (2) the robustness and monotonicity of the resulting spectrum derived from pairwise comparisons.

4.1 Datasets

4.1.1 Real Trademark Dataset

We use a small set of 100 real trademarks from (Guha et al., 2023). This dataset consists of 100 mark–product description pairs, and each was labeled with one of the five Abercrombie distinctiveness categories: generic, descriptive, suggestive, arbitrary, or fanciful. The samples were carefully curated by legal experts and derived from textbook-style exercises.

For our setup, we adopt the same split used by LegalBench: one sample per category is selected as the example set (5 total) for synthetic anchor generation, while the remaining 95 examples are used as real mark candidates in our pairwise comparisons. These real samples' labels remain hidden in our framework, and their distinctiveness is inferred via comparisons to the labeled anchors.

4.1.2 Synthetic Anchor Dataset

For each real trademark (r_i, d_i) in our candidate set, we generate a set of five synthetic anchors $\mathcal{S}$ by fixing the mark name r_i and varying the

goods/services description $\tilde{d}_j$ to target each of the five Abercrombie distinctiveness categories.

This generation is performed using GPT-4o with a temperature setting of 0.3, guided by a structured LLM prompt (Appendix A.1) designed to avoid explicitly revealing the legal category in the text while ensuring coverage across the spectrum. An illustrative example is provided in Section 3.2.

The result is a synthetic dataset of $95 \times 5 = 475$ anchors, where 95 is the number of real marks. These anchors serve as labeled reference points ("anchors") in our pairwise comparison framework. They are used to position real trademarks on a continuous distinctiveness scale.

4.2 Pairwise Comparison Procedure

We design two complementary comparison types to estimate continuous distinctiveness scores for real trademarks: within-group comparison and inter-group (bridge) comparison.

For within-group comparisons, we compare only the synthetic anchors of the same mark. For each mark m, let $A_{m,k} = \{(s_m, \tilde{d}_{m,k}, c_k)\}$ span the available Abercrombie labels $k \in \{0, .., 4\}$. After sorting by label (0 - 4), we add all directed "higher-beats-lower" edges: if $k_j > k_i$, record $A_{m,k_j} > A_{m,k_i}$. This deterministic construction imposes a strict ordinal priority within a mark and does not use LLM judgments. Inter-group comparisons are used to ensure score comparability across different marks by introducing cross-mark matchups. We perform inter-group comparisons using LLMs. Without these bridge cases, the BT model will only estimate a separate scale for each mark, making the scores incomparable across marks. By paring a synthetic anchor with another mark randomly, we create a connected comparison network where all real marks can share a common distinctive scale. Together, the two types of comparisons allow BT to learn the ranking information globally while preserving the relative order within a local community (same mark).

We experimented with two types of strategies to construct inter-group comparisons: random comparison strategy and "chain-link" strategy. Random comparison focuses on randomly comparing the synthetic anchors of different trademarks. The "chain-link" strategy guarantees that all 95 trademarks can be compared on a unified scale by first creating a loop that connects them in a circle. In this setting, each mark gets compared to its neighbors. After establishing the foundational connection for all the marks, the remaining budget (105 comparisons if using 200[2] total) is spent on random pairs that create "shortcuts" across the circle.

We employ both GPT-4o and GPT-5 to conduct the inter-group pairwise comparisons. GPT-4o is used for its proven stability and controllable temperature setting (temperature = 0.3). We set temperature to 0.3 to balance determinism and response flexibility. A temperature of 0 would yield deterministic responses, which can sometimes cause models to be overly sensitive to prompt wording. GPT-5 is included to evaluate whether the latest generation model can provide improved alignment with human-perceived distinctiveness despite its fixed default temperature setting (temperature = 1)[3]. Both models receive identical system and user prompts as shown in Appendix A.2.

5 Experimentation

5.1 Constructing the Anchor Ruler

The goal is to estimate the global BT score θ for synthetic anchors to construct a ruler[4].

5.1.1 Synthetic Anchor Dataset

To initialize the distinctiveness ruler, we generated a balanced set of synthetic anchors spanning the five Abercrombie categories. For each category, an LLM produced 95 anchor marks with short product descriptions. In total, the synthetic dataset contained 475 unique anchors.

5.1.2 Bradley–Terry Model Fitting

We constructed pairwise comparisons to evaluate relative distinctiveness among anchors. As mentioned before, we design two complementary comparison types for estimating distinctiveness cores: within- and inter-group comparisons.

In the within-group comparison setting, anchor pairs were compared directly based on their synthetic labels, without involving LLM judgments. As expected, this procedure will produce consistent local subgraphs with no transitivity violations. We established 987 voting pairs in this setting. For the inter-group comparisons, we introduced cross-mark matchups judged by LLMs to enable global

[2]We limit the total number of inter-group comparisons to balance practical cost constraints while ensuring effective coverage of all 95 trademarks.

[3]For GPT-5, OpenAI fixes the temperature parameter at 1.0, and it cannot be modified by the user.

[4]The code and data are available at https://github.com/JoyceXu02/bt_ruler

comparability across marks. To compare construction strategies, we generated 200 inter-group comparisons under each of the two strategies, random and chain-link construction, resulting in two alternative graphs for performance evaluation.These comparison results are used for BT score estimation.

5.2 Real Trademark Projection on the BT Ruler

After fitting the BT model on synthetic anchors, we estimated distinctiveness scores for real trademarks by aligning them with the synthetic anchor scale. Each real mark was paired with the five synthetic variants of the same mark, covering all Abercrombie categories, and compared in turn by LLMs. This produced a total of 475 comparisons.

For each mark, we separated the anchor scores into wins and losses relative to the real mark. The real mark's score was then placed on the BT scale by bracketing: if it defeated all anchors, it was assigned just above the strongest win with a small margin; if it lost to all, just below the weakest loss with a small margin. Otherwise, it was positioned at the midpoint between the strongest win and weakest loss. This approach integrates the LLM comparison outcomes into the BT ruler without refitting the full model, yielding a consistent estimate of real distinctiveness on the same continuous scale as the synthetic anchors. The projection process is illustrate in Algorithm 1.

Algorithm 1: Project Real Trademark onto BT Ruler

Input : Group for one mark: 1 real trademark and 5 synthetic trademarks with known anchor scores $\{\theta_{m,k}\}$;
LLM outcomes $y_{m,k} \in \{0, 1\}$;
Margin δ (small buffer, e.g. 0.1).

Output : Estimated real-mark score $\hat{\theta}_m^{\text{real}}$.

$W \leftarrow \{\theta_{m,k} \mid y_{m,k} = 1\}$ `// anchors the real mark beat`

$L \leftarrow \{\theta_{m,k} \mid y_{m,k} = 0\}$ `// anchors the real mark lost to`

if $W = \varnothing$ ***and*** $L = \varnothing$ **then**

  `// No evidence`

return NaN

if $L = \varnothing$ **then**

  `// beat all anchors`

  return $\max(W) + \delta$

if $W = \varnothing$ **then**

  `// lost to all anchors`

  return $\min(L) - \delta$

return $\frac{\max(W)+\min(L)}{2}$ `// midpoint between strongest win and weakest loss`

6 Results

6.1 Bradley-Terry Model Fitting

We first evaluated the fitted BT scores for synthetic anchors across all the categories. Table 1 reports summary statistics of the anchor scores by label, split across inter-group construction strategies (random vs. chain-link) and judgment models (GPT-4o vs. GPT-5). The slight variation in N across labels arises because not every mark produced a complete set of synthetic anchors spanning all five Abercrombie categories. In total, 2 out of 95 marks were missing one or more category anchors (see Table 6 in Appendix).

The BT model recovered the expected ordering. Generic anchors received the lowest scores (-0.38), followed by descriptive (≈ -0.17), suggestive (≈ 0.02), arbitrary (≈ 0.19), and fanciful (≈ 0.33). The monotone increase in both mean and median scores across labels confirms that the learned latent scale is aligned with the Abercrombie spectrum.

Results were highly consistent across models and strategies. GPT-4o and GPT-5 produced nearly identical score distributions with differences in mean scores never exceeding 0.003. Random and chain-link constructions yielded similar statistics with slight differences in IQRs.

To evaluate whether the fitted BT scores respect the expected monotone ordering across adjacent categories, we examined boundary-wise violations. Overall, violations were rare, typically under 4% for any given boundary and condition (see Table 7 in Appendix). Most violations occurred at the $0 \rightarrow 1$ and $2 \rightarrow 3$.

Overall, these results demonstrate that the BT model produces a stable and well-ordered continuous scale from synthetic anchors and largely invariant to model choice or inter-group construction strategy. This synthetic anchor scale serves as the distinctiveness ruler onto which real trademarks can be mapped in the next stage of analysis.

Table 1: BT score (θ) statistics by Abercrombie label, split by inter-group strategy and model. Labels: 0=Generic, 1=Descriptive, 2=Suggestive, 3=Arbitrary, 4=Fanciful.

Strategy	Model	Label	N	Mean	SD	Median	IQR
Random	GPT-4o	0	93	−0.377	0.119	−0.363	0.009
		1	94	−0.173	0.060	−0.154	0.008
		2	96	0.015	0.057	0.025	0.007
		3	96	0.190	0.067	0.185	0.007
		4	96	0.330	0.053	0.325	0.007
Random	GPT-5	0	93	−0.380	0.120	−0.362	0.006
		1	94	−0.174	0.068	−0.153	0.005
		2	96	0.015	0.051	0.027	0.006
		3	96	0.193	0.049	0.185	0.004
		4	96	0.330	0.048	0.325	0.004
Chain-link	GPT-4o	0	93	−0.387	0.105	−0.363	0.013
		1	94	−0.178	0.079	−0.153	0.013
		2	96	0.024	0.066	0.027	0.011
		3	96	0.189	0.044	0.186	0.009
		4	96	0.336	0.045	0.326	0.009
Chain-link	GPT-5	0	93	−0.387	0.107	−0.362	0.015
		1	94	−0.180	0.081	−0.152	0.016
		2	96	0.026	0.067	0.027	0.011
		3	96	0.191	0.038	0.186	0.008
		4	96	0.334	0.044	0.326	0.009

6.2 Mapping Real Trademarks

To evaluate how the BT ruler generalizes to real-world cases, we mapped 95 real trademarks onto the synthetic anchor scale. For each mark, we constructed a set of five synthetic anchors spanning the Abercrombie categories and generated pairwise outcomes from LLMs comparing the real mark against each anchor. This produced a total of 475 LLM-based comparisons. Using these outcomes, we estimated a BT score for each real mark by bracketing it between the strongest anchor it defeated and the weakest anchor it lost to with a small margin adjustment at the extremes. The resulting scores place real marks directly onto the continuous distinctiveness scale defined by the synthetic anchors.

Table 2 summarizes the distribution of BT scores assigned to the 95 real trademarks across different inter-group construction strategies and LLM models.The mean scores range from –0.176 to –0.149. Median values are slightly lower (–0.229 to -0.25). These statistics indicate that, regardless of model or strategy, real trademarks are mapped onto a similar region of the continuous ruler with comparable variability. This stability suggests that the mapping procedure is robust to modeling choices and produces a stable placement of real marks on the synthetic anchor scale.

Figures 1 and 2 present the overall distributions of BT scores assigned to the 95 real trademarks under the random bridge and chain-link bridge strategies using GPT-4o and GPT-5. Across all conditions, we see distinct spikes appear in the -0.3 to -0.4 range. It reflects that clusters of marks are assigned with similar scores. The choice of bridge strategy has subtle effects: for GPT-4o, the chain-link approach produces more pronounced clustering, while GPT-5 yields a smoother spread with the chain-link approach.

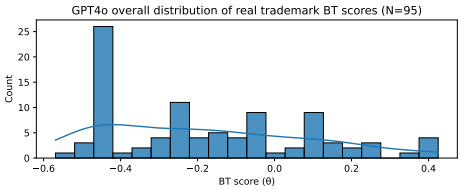
Table 2: Descriptive statistics of real trademark BT scores under different inter-group strategies and LLM models.

Strategy	Model	Mean	SD	Median	IQR
Random	GPT-4o	−0.176	0.266	−0.229	0.488
Random	GPT-5	−0.149	0.298	−0.229	0.556
Chain-link	GPT-4o	−0.171	0.269	−0.250	0.493
Chain-link	GPT-5	−0.149	0.298	−0.251	0.546

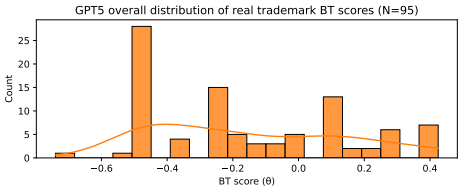
6.3 Validation of Real Trademark Mapping

To assess whether the mapped BT scores for real trademarks align with their gold Abercrombie categories, we conducted both correlation and distributional analyses.

First, we computed rank-order correlations between the continuous BT scores and the categorical labels across different models and inter-group strategies. We report Spearman's ρ, Kendall's τ and Pearson's r. Spearman's ρ measures the monotonic association between two ranked variables, whereas Pearson's r captures linear correlation between them. Kendall's τ quantifies the proportion

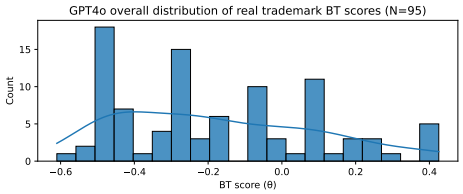


(a) Distribution of real trademarks BT scores using random bridge strategy with GPT-4o.

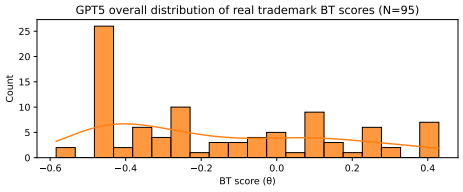


(b) Distribution of real trademarks BT scores using random bridge strategy with GPT-5.

Figure 1: Overall distributions of real trademark BT scores under the random bridge strategy. Subfigure (a) shows results using GPT-4o, while subfigure (b) presents results using GPT-5.



(a) Distribution of real trademarks BT scores using chain-link bridge strategy with GPT-4o.



(b) Distribution of real trademarks BT scores using chain-link bridge strategy with GPT-5.

Figure 2: Overall distributions of real trademark BT scores under the chain-link bridge strategy. Subfigure (a) shows results using GPT-4o, while subfigure (b) presents results using GPT-5.

of concordant versus discordant pairs and provides a more conservative measure of ordinal agreement.

From Table 3, under the random bridge strategy, GPT-4o achieved Spearman's $\rho = 0.471$,Kendall's $\tau = 0.332$ and Pearson's r = 0.506, while GPT-5 improved to $\rho = 0.675$, $\tau = 0.494$ and $r = 0.705$. With the chain-link strategy, GPT-4o reached $\rho = 0.526$, $\tau = 0.381$ and $r = 0.513$, and GPT-5 again achieved the highest alignment with $\rho = 0.718$, $\tau = 0.547$ and $r = 0.724$. These findings demonstrate that the BT ruler preserves the intended ordinal structure of the Abercrombie spectrum, with GPT-5 producing consistently stronger correlations than GPT-4o across both bridge strategies.

We further examined the distributions of real trademark scores within each category using boxplots for the two best-performing configurations: GPT-4o and GPT-5 under the chain-link strategy. Figure 3 show clear separation at the extremes: generic and descriptive marks clustering toward the lower end of the scale and arbitrary and fanciful marks concentrating toward the higher end.

7 Discussion

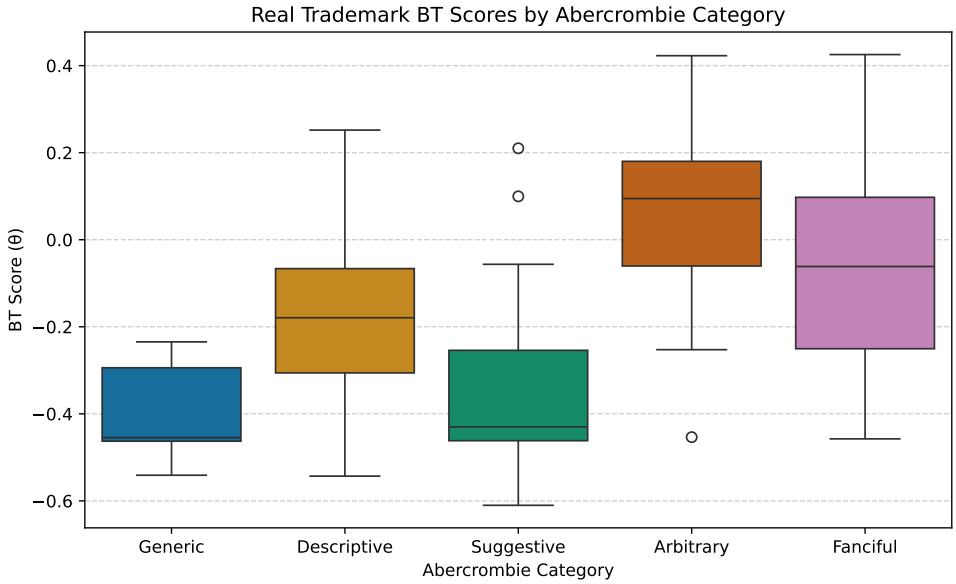
7.1 Synthetic Anchor Fitting

The fitting of synthetic anchors demonstrated that the BT model can successfully recover a continuous distinctiveness scale aligned with the Abercrombie spectrum. The model produced a clear monotonic progression from generic to fanciful, with consistent score distributions across both models and inter-group construction strategies.

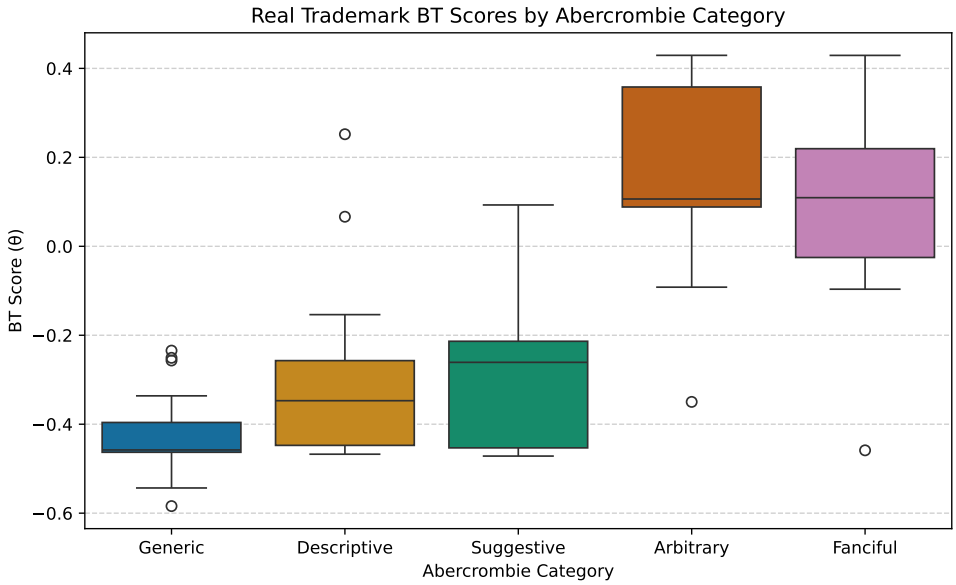
Table 3: Rank-order correlation between real trademark BT scores and gold Abercrombie categories, reported as Spearman's ρ, Kendall's τ and Pearson's r under different inter-group strategies and LLM models.

Strategy	Model	Spearman ρ	Kendall τ	Pearson r
Random	GPT-4o	0.471	0.332	0.506
Random	GPT-5	0.675	0.494	0.705
Chain-link	GPT-4o	0.526	0.381	0.513
Chain-link	GPT-5	0.718	0.547	0.724

At the same time, we observed rare monotonicity violations, which are typically under 5% at any boundary. These flips most often occurred at the edges of the spectrum (like generative vs. descriptive), where the legal distinctions are more ambiguous. For instance, the mark "Pen" was mapped closer to descriptive rather than generic, despite its direct reference to the product ($0 \rightarrow 1$ violation). Similarly, "Cutlery" showed a $3 \rightarrow 4$ violation, suggesting difficulty in separating arbitrary from fanciful uses. These cases demonstrate that flips tend to cluster around boundaries where legal interpretation is already ambiguous. They highlight the gray zones where categorical boundaries are hard to enforce, and a continuous scoring approach can reveal uncertainties.



(a) Mapping of real trademarks on the BT ruler using GPT-4o.



(b) Mapping of real trademarks on the BT ruler using GPT-5.

Figure 3: Comparison of real trademark mappings on the BT ruler across models. Subfigure (a) shows results obtained with GPT-4o, while subfigure (b) presents results from GPT-5. Both visualizations illustrate how real marks are distributed across the continuous scale relative to the Abercrombie categories.

7.2 Real Trademark Mapping

The mapping of real trademarks onto the BT ruler provides insight into how categorical distinctiveness judgments translate into a continuous scale. Figure 1 and 2 show that real marks span the full ruler. Spikes in the –0.3 to –0.4 range suggest that multiple marks are consistently assigned to similar borderline positions.

We also examined the distributions of real trademark scores within each category using boxplots. Figure 3 shows that the BT ruler recovers the expected ordinal progression across categories with a misalignment at the higher end. GPT-5 shows more prominent separation at the extremes. Meanwhile, GPT-5 improves this ordering by aligning the medians more appropriately on arbitrary and fanciful marks than GPT-4o. This suggests that GPT-5 produces a more coherent representation of the distinctiveness spectrum particularly at the higher end.

To illustrate how real marks are embedded onto the BT ruler, Figure 4 shows the placement of the mark "Salt" for packages of sodium chloride. The use of Salt (black X) falls away from the cluster of generative marks (black cluster). Circles are 5 synthetic anchors across the Abercrombie spectrum. This positioning highlights how the model interprets Salt as leaning strongly toward generic. The case shows the diagnostic value of the BT ruler, which not only assigns a value but also reveals why certain marks are classified under the framework.

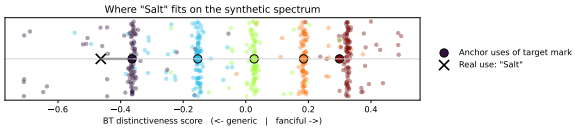


Figure 4: Mapping of the real trademark "Salt" for packages of sodium chloride.

Limitations and Future Work

Although our findings show that the BT ruler provides a robust way to embed trademarks onto a continuous distinctiveness scale, there are several limitations. First, the evaluation was conducted on a relatively small set of 95 real marks. It may not fully capture the heterogeneity of trademark usage in practice. Future work should test the method on larger datasets to evaluate its performance in more borderline or evolving cases.

Second, the current design of synthetic anchors, fixing the mark name while varying product or service descriptions, spans the overall coverage of the Abercrombie spectrum but cannot yield legally valid fanciful marks. True fanciful marks must be invented or linguistically novel, not derived from existing terms. The limitation will be addressed in future work by generating neologisms to strengthen representation at the upper end of the distinctiveness scale.

Third, monotonicity violations and category overlaps show that the BT framework does not eliminate ambiguity. Instead, it expresses uncertainty when forced to assign discrete labels. Interpreting BT scores (e.g., -0.3 vs -0.25) in legal terms will require input from practitioners.

For future work, we can extend this study in several directions. We can incorporate human expert judgment alongside LLM-based comparisons for a deeper validation of the BT ruler. Besides, since our approach provides continuous global distinctiveness scores, it captures the relative positioning of marks across the entire Abercrombie spectrum. This enables quantitative assessment of how a mark's distinctiveness may evolve when new evidence emerges or when evaluated in different contexts. Ultimately, these extensions could transform the BT ruler from a proof of concept into a practical decision-support tool for trademark practitioners when seeking to quantify distinctiveness with transparency.

Acknowledgments

This work was supported by the National Science Foundation (Grant No. 2040490, FAI: Using AI to Increase Fairness by Improving Access to Justice) and by a Pitt Momentum Funds Scaling Grant. This research was supported in part by the University of Pittsburgh Center for Research Computing and Data, RRID:SCR_022735, through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483.

References

Shivam Adarsh, Elliott Ash, Stefan Bechtold, Barton Beebe, and Jeanne Fromer. 2024. Automating abercrombie: Machine-learning trademark distinctiveness. *Journal of Empirical Legal Studies*, 21(4):826–860.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Stacey L Dogan and Mark A Lemley. 2006. Grounding trademark law through trademark use. *Iowa L. Rev.*, 92:1669.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Sakshi Singh, Utkarsh Tyagi, Dinesh Manocha, and 1 others. 2023. Dale: Generative data augmentation for low-resource legal nlp. *CoRR*.

John Goodhue and Yolanda Wei. 2023. Classification of trademark distinctiveness using openai gpt 3.5 model. *Available at SSRN 4351998*.

Stuart JH Graham, Galen Hancock, Alan C Marco, and Amanda Fila Myers. 2013. The uspto trademark case files dataset: Descriptions, lessons, and insights. *Journal of Economics & Management Strategy*, 22(4):669–705.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters,

Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

William M Landes and Richard A Posner. 1987. Trademark law: an economic perspective. *The Journal of Law and Economics*, 30(2):265–309.

Ge Luo, Hebi Li, Youbiao He, and Forrest Sheng Bao. 2022. PrefScore: Pairwise preference learning for reference-free summarization quality assessment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5896–5903, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Lisa Larrimore Ouellette. 2014. The google shortcut to trademark law. *California Law Review*, 102:351.

Sezen Perçin, Andrea Galassi, Francesca Lagioia, Federico Ruggeri, Piera Santin, Giovanni Sartor, and Paolo Torroni. 2022. Combining wordnet and word embeddings in data augmentation for legal texts. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 47–52.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Huihui Xu and Kevin Ashley. 2023. A question-answering approach to evaluating legal summaries. In *Legal Knowledge and Information Systems*, pages 293–298. IOS Press.

Zhi Zhou, Kun-Yang Yu, Shi-Yu Tian, Xiao-Wen Yang, Jiang-Xin Shi, Pengxiao Song, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2025. Lawgpt: Knowledge-guided data generation and its application to legal llm. *arXiv preprint arXiv:2502.06572*.

A Appendix

A.1 Prompt Template for Synthetic Anchor Generation

Example Prompt for Synthetic Anchor Generation
Input Mark and Domain:
`Mark: "Salt"`
`Original Domain: Packages of sodium chloride`
System Instruction:
Generate trademark use cases that vary in legal distinctiveness under the Abercrombie spectrum.
The Abercrombie spectrum defines trademark distinctiveness as follows:
- Generic: Common words for a category of goods or services; cannot be protected.
- Descriptive: Directly describes a quality, function, or ingredient of the product.
- Suggestive: Requires imagination or thought to connect to the product.
- Arbitrary: Common words used in an unrelated context.
- Fanciful: Completely invented or meaningless terms.
Task Instruction:
Given the trademark "{mark}" and its current domain "{domain}", generate five alternative product or service domains, each corresponding to a different level of inherent distinctiveness under the Abercrombie spectrum.
- Avoid repeating the original domain or use case.
- Avoid well-known existing marks or legally impossible uses.
- Be specific and legally sound in your reasoning. Do NOT use or paraphrase these words: generic, descriptive, suggestive, arbitrary, fanciful, common, everyday, coined, imaginative, fancifully, arbitrary use, generic term, descriptive term. Focus on the legal facts (consumer perception, inherent meaning, connection to goods).
- Do NOT mention or hint at the legal distinctiveness category in the 'text'. That field should only be a natural description of the trademark in context.
- The distinctiveness label will be provided separately in the "distinctiveness" field.

Table 4: Example prompt template for synthetic anchor generation.

A.2 Prompt Template for Pairwise Comparison

Prompt Template for Pairwise Distinctiveness Comparison
System Instruction:
You are a U.S. trademark examiner.
###Task: Given following two mark descriptions, decide which use of a mark is more inherently distinctive (i.e. easier to protect under the Abercrombie spectrum).
###Rules: - Do NOT reveal or paraphrase the spectrum terms (generic, descriptive, suggestive, arbitrary, fanciful) or synonyms such as "common, everyday, coined, imaginative".
- Base your decision only on how strongly the MARK relates to the GOODS/SERVICES named.
- Only pick one from the given two mark descriptions. - Keep the reason concise and legally relevant.
User Input Example:
Which use is more distinctive?
A: The mark 'Salt' for a tech startup offering cloud storage solutions.
B: The mark "Salt" for packages of sodium chloride. Respond with either A or B with your reason.

Table 5: Prompt template used for pairwise distinctiveness comparisons between two mark–product descriptions. The system instruction remains fixed, while the user input is dynamically populated with the pair being compared.

A.3 Incomplete Synthetic Ladders

Each mark should ideally include one anchor per Abercrombie category (0–4), but both *Gun* and *Telephone* cases exhibit missing or duplicated labels, resulting in incomplete five-level ladders. These irregularities account for minor variations in N across categories reported in Table 1. The full descriptions of the two marks along with their generated anchors and assigned labels are listed in Table 6.

A.4 Monotonicity Violation Counts

Mark	Generated Description	Label
Telephone	The mark ‘Telephone‘ for a brand of high-end fashion clothing.	Arbitrary
	The mark ‘Telephone‘ for a software application for managing digital contacts.	Suggestive
	The mark ‘Telephone‘ for a type of electronic music.	Fanciful
	The mark ‘Telephone‘ for a telecommunications consulting service.	Descriptive
	The mark ‘Telephone‘ for a brand of herbal tea.	Fanciful
Gun	The mark ‘Gun‘ for a brand of energy drinks.	Arbitrary
	The mark ‘Gun‘ for a type of software for data analysis.	Suggestive
	The mark ‘Gun‘ for a line of spicy sauces.	Suggestive
	The mark ‘Gun‘ for a brand of shoes.	Arbitrary
	The mark ‘Gun‘ for a new type of fruit.	Fanciful

Table 6: Examples of incomplete synthetic anchor spans for the marks *Telephone* and *Gun*.

Table 7: Monotonicity violation counts and rates by boundary, for each inter-group construction strategy and judgment model. Boundaries denote adjacent Abercrombie categories (0=Generic, 1=Descriptive, 2=Suggestive, 3=Arbitrary, 4=Fanciful).

Strategy	Model	Boundary	N ladders	Violations	Rate (%)
Random	GPT-4o	0→1	93	2	0.022
		1→2	94	0	0
		2→3	95	2	0.021
		3→4	95	3	0.032
Random	GPT-5	0→1	93	3	0.032
		1→2	94	1	0.011
		2→3	95	0	0
		3→4	95	1	0.011
Chain-link	GPT-4o	0→1	93	3	0.032
		1→2	94	0	0
		2→3	95	4	0.042
		3→4	95	1	0.011
Chain-link	GPT-5	0→1	93	4	0.043
		1→2	94	0	0
		2→3	95	1	0.011
		3→4	95	1	0.011