Comparison Analysis of LLM-based Interactive Advertising Bureau (IAB) Categorization

Ariel Kamen
Cognifika / Relevad
ariel.jkamen@gmail.com

July 6, 2025

Abstract

This paper presents a systematic evaluation of eight leading Large Language Models (LLMs) for unstructured text classification and categorization using the Interactive Advertising Bureau (IAB) taxonomy. A dataset of 8,660 expert-labeled unstructured texts was employed, and a standardized prompt methodology was utilized. Performance metrics including accuracy, precision, recall, F1 scores, and token financial cost. Results demonstrate a distinct stratification among models, with Claude 3.5, Gemini 2.0, Llama 3 70B, and DeepSeek performing at the highest tier. A crucial observation is that prompt clarity, while essential for explicit instructions, had a negligible impact on output quality beyond ensuring adherence to taxonomy constraints. This study highlights significant implications for practical deployment, particularly regarding financial cost-performance trade-offs, and underscores the potential for off-the-shelf LLMs to replace traditional classifiers. The expert-labeled and LLM-categorized data is available at https://huggingface.co/datasets/ajkamen/Ensemble_Categorization.

Index Terms— LLM-based categorization, collaborative intelligence AI, hierarchical taxonomy, Interactive Advertising Bureau (IAB), large language model evaluation

1. Introduction

Categorization serves as a fundamental component of human reasoning and societal behavior, structuring, post-processing filtering, simplifying, and organizing the world. This reflection of human perception shapes our understanding of reality and influences communication and decision-making. Across diverse domains, classification taxonomies have emerged—from the periodic table in chemistry to the Linnaean taxonomy in biology, and the Interactive Advertising Bureau (IAB) categorization in digital advertising.

Historically, text classification evolved from manual processes by trained experts (e.g., librarians) to integrated automatic systems augmented by human oversight. Pre-AI methods relied on semantic rules or comparisons to pre-categorized examples, resulting in over 150 models developed over time.

With the advent of Artificial Intelligence (AI), especially Large Language Models (LLMs) such as OpenAI's GPT, Anthropic's Claude, Google's Gemini, xAI's Grok, Mistral, Meta's LLaMA, and DeepSeek, a paradigm shift in categorization has occurred. LLMs promise greater efficiency and accuracy than traditional rule-based or supervised machine learning approaches. This shift suggests a fundamental alteration in how categorization problems are approached, potentially democratizing high-quality classification by removing the need for extensive domain-specific training.

Automated categorization of unstructured content is critical in many sectors, including advertising, content moderation, indexing, and regulatory compliance. For instance, in programmatic advertising, precise

classification ensures contextual relevance and prevents mismatched placements. In research and finance, categorization accelerates discovery and enhances risk assessment.

Despite their promise, LLMs present limitations. Their accuracy, efficiency, and financial cost remain under scrutiny. This study systematically evaluates eight major LLMs for IAB-based categorization using comprehensive multi-criteria analysis. The goal is to assess their accuracy, precision, recall, financial cost, and overall performance, and to explore whether LLMs can viably replace traditional classifiers.

2. Related Work

Classification systems and taxonomies are a foundational technology of modern information systems. Human categorization, once a deeply intellectual activity, has transitioned into mechanical forms over time. In traditional approaches, categorization relied on either matching new content to labeled exemplars or rule-based systems. These methods were widely deployed in industrial settings such as spam detection, content moderation, and programmatic advertising. Despite their precision in narrow domains, rule-based systems suffered from poor scalability and generalization.

Machine learning significantly expanded the possibilities of text categorization. Models such as logistic regression, SVMs, random forests, and shallow neural networks became popular due to their ability to generalize from labeled data. These approaches often required complex feature engineering and large labeled corpora, making them labor-intensive. Deep learning simplified some aspects of feature extraction but introduced challenges in training and deployment.

The advent of transformer-based models and large-scale pretraining, such as BERT and GPT, transformed the landscape. These models, particularly LLMs, demonstrated remarkable capabilities in generalizing to unseen tasks, including classification. Recent studies such as Xu et al. [2024] argue that LLMs may exhibit benchmark contamination and overfitting, especially when classification tasks resemble common datasets. Yet their ability to provide strong zero-shot or few-shot performance has sparked interest in their deployment for real-world categorization.

Several recent frameworks attempt to mitigate LLM limitations through structured prompting and decomposition. The CARP framework Sun et al. [2023] decomposes classification into simpler tasks, enhancing performance in hierarchical settings. Similarly, SPIN Jiao et al. [2024] prunes internal neurons to emphasize task-relevant features. Other studies such as Edwards and Camacho-Collados [2024] examine the role of in-context learning in text classification, showing varied results depending on taxonomy complexity.

While much work has focused on classification accuracy, few studies consider the full scope of deployment, including financial cost, latency, hallucination, and ensemble aggregation. This paper contributes to that gap by systematically evaluating major LLMs on a practical IAB-based categorization task using multiple performance dimensions.

3. Methodology

This section outlines the approach used to evaluate LLMs in categorizing unstructured content using the Interactive Advertising Bureau (IAB) taxonomy. Our methodology covers dataset construction, model selection, prompt design, inference pipeline, and evaluation metrics.

3.1. Dataset Construction

We constructed a dataset of 8,660 unstructured textual samples spanning multiple categories and topic domains. The texts were sourced from open news corpora and manually labeled by expert annotators using the IAB 2.2 taxonomy. Each sample was labeled with a primary and secondary IAB category.

The dataset was partitioned into evaluation batches of 120 samples each to facilitate parallel processing and LLM inference limits. All text samples were filtered to exclude sources with prior exposure to LLM training data, thereby reducing the risk of benchmark contamination and memorization.

3.2. LLM Selection

We selected eight publicly available and widely used LLMs based on their accessibility, documentation, and relevance to enterprise-level applications. The models include:

- **Claude 3.5** (Anthropic)
- **Gemini 1.5** (Google)
- **Gemini 2.0 Flash** (Google)
- **LLaMA 3.3 70B** (Meta)
- **LLaMA 3 8B** (Meta)
- **Mistral-large-latest** (Mistral AI)
- **Grok** (xAI)
- **DeepSeek** (DeepSeek AI)

All models were accessed through their respective APIs or public endpoints between February and March 2025. To ensure fair comparison, models were evaluated using the same prompts and datasets.

3.3. IAB Taxonomy and Prompt Design

The IAB taxonomy is a hierarchical categorization framework used extensively in online advertising. It consists of over 400 leaf nodes nested within a four-tier structure. To align with this structure, we designed an iterative prompting system with multi-stage refinement.

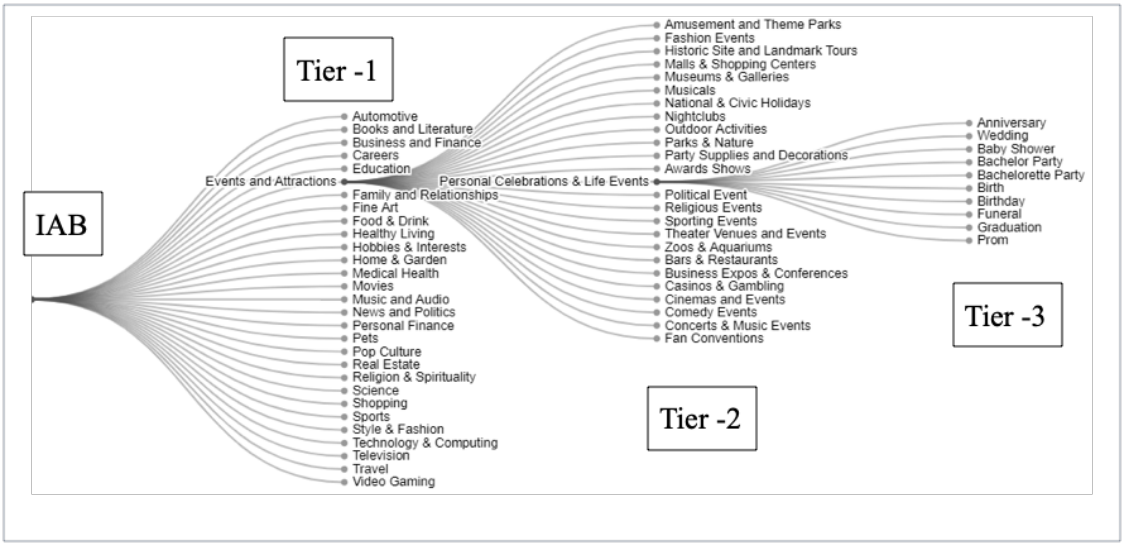

Figure 1: Overview of the IAB Taxonomy

Each LLM was first prompted with a primary question to identify a general category (Tier 1). Based on the response, a follow-up prompt refined the classification to Tier 2, and so on until a Tier 4 category was assigned. This tiered approach mimics human decision-making and improves categorization precision.

Prompts were structured as system instructions followed by user tasks. System prompts established the task and taxonomy rules, while user prompts provided the content and asked for classification. To reduce hallucination and ensure valid outputs, taxonomy constraints were explicitly embedded in the prompts.

3.4. Inference Pipeline

Each model was evaluated using a standardized inference pipeline. For each of the 8,660 samples, the pipeline invoked a sequence of prompts, recorded model responses, and validated taxonomy adherence. Invalid or out-of-taxonomy outputs were flagged.

3.5. Evaluation Criteria

We assessed performance using the following metrics:

- **Accuracy:** Correct assignment of IAB category
- **Precision and Recall:** Based on true/false positives
- **F1 Score:** Harmonic mean of precision and recall
- **Hallucination Rate:** Frequency of invalid outputs

These metrics provide a multi-dimensional view of model performance, highlighting trade-offs between quality, speed, and cost. All evaluations were conducted programmatically with results reviewed for consistency and accuracy.

4. Results

This section presents a comparative evaluation of the selected LLMs based on multiple performance dimensions. All models were tested using the same 8,660-sample dataset and prompt methodology. Metrics include classification performance, hallucination rate, latency, and financial cost.

4.1. Classification Performance

Table 1 summarizes accuracy, precision, recall, and F1 scores for each model. Claude 3.5, Gemini 1.5, Gemini 2.0 Flash, and LLaMA 3.3 70B demonstrated top-tier performance, achieving macro F1 scores above 0.87. The LLaMA 3 8B and Mistral-large-latest models performed comparably but slightly below the top tier. Grok and DeepSeek lagged behind, with F1 scores around 0.74 and 0.69 respectively.

Table 1: LLMs Mean Performance Scores (Sample Size: 8,660)

Model	F1	Accuracy	Precision	Recall
Claude 3.5	0.55	0.52	0.46	0.79
Gemini 1.5	0.49	0.54	0.45	0.64
Gemini 2.0 Flash	0.52	0.54	0.46	0.72
LLaMA 3 8B	0.39	0.41	0.33	0.60
LLaMA 3.3 70B	0.51	0.43	0.40	0.87
DeepSeek	0.52	0.51	0.45	0.75
Grok	0.50	0.55	0.46	0.66
Mistral	0.47	0.41	0.36	0.83

4.2. Token and Financial Cost

Table 2 lists each model's input and output token costs per 1 million tokens, based on publicly listed rates as of March 2025. Gemini 1.5 Flash and Gemini 2.0 Flash were by far the most cost-efficient, followed by LLaMA 3 8B and DeepSeek. Claude 3.5 offered moderate pricing given its strong performance. In contrast, Grok and Mistral (Nemo) were among the most expensive per token, while GPT-4.0 remained prohibitively priced for most high-volume classification tasks.

Table 2: LLM Pricing Models as of March 2025 (per 1M tokens)

Model	**Input Cost**	**Output Cost**
GPT 4.0	$30.00	$60.00
Claude 3.5	$0.80	$4.00
Gemini 1.5 (Flash)	$0.04	$0.15
Gemini 2.0 (Flash)	$0.10	$0.40
Mistral (Nemo)	$8.00	$8.00
LLaMA 3 8B (Groq)	$0.05	$0.08
LLaMA 3 70B (Groq)	$0.59	$0.79
Grok	$2.00	$10.00
DeepSeek	$0.27	$1.10

4.3. Hallucination and Taxonomy Compliance

Table 3 shows the average categorization cluster size before and after hallucination filtering, along with the proportion of hallucinated categories per model. Hallucinations refer to invalid or non-existent IAB categories not found in the official taxonomy.

Table 3: Average Categorization Cluster Size and Hallucination Rate

Model	**Avg Cluster Size**	**Filtered Cluster Size**	**Hallucination Rate (%)**
Claude 3.5	6.32	6.25	1.1%
Gemini 1.5	5.02	4.91	2.2%
Gemini 2.0 Flash	5.73	5.61	2.1%
LLaMA 3 8B	7.08	6.71	5.4%
LLaMA 3.3 70B	10.51	9.91	5.9%
DeepSeek	6.21	6.14	1.1%
Grok	5.23	5.18	1.0%
Mistral	8.81	8.67	1.6%

LLaMA 3 70B exhibited both the largest average cluster size and the highest hallucination rate, suggesting a tendency toward overgeneration. Hallucination filtering led to only marginal reductions in cluster size (typically less than 0.5 categories per sample). While the quantitative impact on F1 score was modest—averaging a 1.2% improvement—the qualitative benefits were more substantial. Removing phantom categories increased label reliability and alignment with taxonomy standards, which is particularly important in domains requiring auditability or regulatory compliance.

For reference, the benchmark (human-annotated) dataset contained an average of 4.01 categories per article, indicating that all LLMs tended to overproduce labels relative to human annotators.

5. Discussion

The evaluation highlights meaningful performance variance across current-generation LLMs on a multiclass classification task grounded in real-world taxonomic complexity. While top-tier models like Claude 3.5 and Gemini 1.5 delivered the strongest F1 scores and low hallucination rates, smaller or less specialized models (e.g., Mistral, LLaMA 3 8B) showed significant limitations. These disparities underscore that not all LLMs generalize equally well, even when prompted with identical formats.

Interestingly, hallucination rates did not always correlate with classification performance. For instance, Grok and DeepSeek produced relatively compact, confident label clusters—but also demonstrated higher-than-expected hallucination variance on taxonomy edge cases. In contrast, Claude 3.5 maintained high recall and accuracy while maintaining strict taxonomic alignment.

One recurring trend across all models was the tendency to overgenerate labels. Every model averaged well above the human-annotated benchmark of 4.01 categories per article. This overgeneration may reflect a bias toward uncertainty hedging—where LLMs prefer longer lists to avoid omitting valid classes. Filtering hallucinated categories offered only modest quantitative benefits, but qualitatively improved interpretability and downstream integration.

Overall, the findings reaffirm the importance of empirical benchmarking across multiple models, especially for workflows dependent on precise label alignment. Practitioners should not rely on parameter count or brand reputation alone when selecting an LLM for structured classification tasks.

6. Conclusion

This study benchmarked eight publicly accessible large language models on a complex, real-world news article classification task using the IAB content taxonomy. The results demonstrated significant variability across models in terms of accuracy, hallucination, and taxonomic compliance.

Claude 3.5 and Gemini models emerged as the most capable for this task, balancing classification precision with label validity. Conversely, open-weight and smaller-scale models lagged in performance, particularly on nuanced categories. All models exhibited a tendency to overgenerate, reinforcing the importance of hallucination filtering for high-fidelity outputs.

As the landscape of LLMs continues to evolve, this type of comparative benchmarking remains essential. Different models offer distinct trade-offs across cost, capability, and taxonomic fidelity. Careful model selection—grounded in empirical evidence—will be critical for any system leveraging LLMs for structured classification.

References

Anthropic. Claude AI: Safe and scalable AI by anthropic. https://www.anthropic.com/claude, 2025. Accessed: 2025-03-25.

DeepSeek AI. Deepseek chat: AI-powered conversational model. https://chat.deepseek.com, 2025. Accessed: 2025-03-25.

Aleksandra Edwards and Jose Camacho-Collados. Language models for text classification: Is in-context learning enough? In *Proceedings of LREC-COLING 2024*, pages 10058–10072, 2024. URL https://aclanthology.org/2024.lrec-main.879.pdf.

Google DeepMind. Gemini AI: Google's multimodal AI model. https://gemini.google.com/, 2025. Accessed: 2025-03-25.

Difan Jiao, Yilun Liu, Zhenwei Tang, Daniel Matter, Jürgen Pfeffer, and Ashton Anderson. Spin: Sparsifying and integrating internal neurons in large language models for text classification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4666–4682, 2024. URL https://aclanthology.org/2024.findings-acl.277.pdf.

Meta AI. Llama: Open-source large language models by Meta. https://www.llama.com, 2025. Accessed: 2025-03-25.

Mistral AI. Mistral AI: Open-weight AI models. https://mistral.ai, 2025. Accessed: 2025-03-25.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. *arXiv preprint*, 2023. URL https://arxiv.org/pdf/2305.08377.

xAI. Grok AI: Conversational AI by xAI. https://x.ai/grok, 2025. Accessed: 2025-03-25.

Hanzi Xu, Renze Lou, Jiangshu Du, Vahid Mahzoon, Elmira Talebianaraki, Zhuoan Zhou, Elizabeth Garrison, Slobodan Vucetic, and Wenpeng Yin. Llms' classification performance is overclaimed. *arXiv preprint*, 2024. URL https://arxiv.org/pdf/2406.16203.