

Order from Chaos: Comparative Study of Ten Leading LLMs on Unstructured Data Categorization

Ariel Kamen
RingCentral, UC Davis
ariel.kamen@ringcentral.com

Sep. 15, 2025

Abstract

This study presents a comparative evaluation of ten state-of-the-art large language models (LLMs) applied to unstructured text categorization using the Interactive Advertising Bureau (IAB) 2.2 hierarchical taxonomy. The analysis employed a uniform dataset of 8,660 human-annotated samples and identical zero-shot prompts to ensure methodological consistency across all models. Evaluation metrics included four classic measures—accuracy, precision, recall, and F1-score—and three LLM-specific indicators: hallucination ratio, inflation ratio, and categorization cost.

Results show that, despite their rapid advancement, contemporary LLMs achieve only moderate classic performance, with average scores of 34% accuracy, 42% precision, 45% recall, and 41% F1-score. Hallucination and inflation ratios reveal that models frequently overproduce categories relative to human annotators. Among the evaluated systems, Gemini 1.5/2.0 Flash and GPT 20B/120B offered the most favorable cost-to-performance balance, while GPT 120B demonstrated the lowest hallucination ratio. The findings suggest that scaling and architectural improvements alone do not ensure better categorization accuracy, as the task requires compressing rich unstructured text into a limited taxonomy—a process that challenges current model architectures.

To address these limitations, a separate ensemble-based approach was developed and tested. The ensemble method, in which multiple LLMs act as independent experts, substantially improved accuracy, reduced inflation, and completely eliminated hallucinations. These results indicate that coordinated orchestration of models—rather than sheer scale—may represent the most effective path toward achieving or surpassing human-expert performance in large-scale text categorization.

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

1. Introduction

Text categorization is a core task in natural language processing (NLP), supporting applications such as spam filtering, sentiment analysis, document retrieval, and content moderation. Early approaches relied on manual expert annotation, rule-based heuristics, or traditional machine learning models that required domain-specific training and extensive feature engineering. While effective, these methods often demanded significant resources and lacked scalability.

Recent advances in large language models (LLMs) such as OpenAI’s GPT, Google’s Gemini, Anthropic’s Claude, xAI’s Grok, Meta’s LLaMA, Mistral, and DeepSeek have introduced strong zero-shot classification capabilities, lowering the barrier to deploying text categorization systems. Yet their application raises key questions: how do LLMs compare to traditional models, what hidden costs and limitations

emerge, and which models provide the best trade-off between quality and efficiency? To address these questions, we benchmark ten major LLMs on an 8,660-document corpus annotated with the Interactive Advertising Bureau (IAB 2.2) taxonomy, evaluating both standard metrics (accuracy, precision, recall, F1-score) and three LLM-specific measures: hallucination ratio, category inflation ratio, and token-processing cost. We further examine the impact of prompt design and API-level hyperparameter variations (temperature, top-k, and maximum tokens), offering a comprehensive assessment of the strengths, limitations, and practical trade-offs of LLMs in large-scale unstructured text categorization.

The next section reviews related work in traditional text classification methods and recent studies of LLMs applied to categorization tasks.

2. Related Work

Classification systems and taxonomies are a cornerstone of modern information systems. Human categorization, once an intellectual and interpretive activity, has gradually transitioned into mechanical and computational forms. Traditional approaches relied on matching new content to labeled exemplars or on manually constructed rule-based systems. These methods were widely deployed in industrial contexts such as spam detection, content moderation, and programmatic advertising. While precise in narrow domains, rule-based systems were difficult to scale, struggled to generalize to unstructured documents, and remained heavily language dependent.

The rise of machine learning significantly broadened the scope of text categorization. Models such as logistic regression, support vector machines (SVMs) [Joachims, 1998], random forests [Breiman, 2001], and shallow neural networks gained popularity due to their ability to generalize from labeled data. Comprehensive surveys such as Sebastiani [2002] document the rapid development of automated text categorization during this period. However, these methods typically required complex feature engineering and large annotated corpora, making them resource intensive. Deep learning approaches [LeCun et al., 2015] reduced the burden of manual feature design but introduced new challenges in training, computation, and deployment.

The introduction of transformer-based architectures and large-scale pretraining, exemplified by BERT [Devlin et al., 2019] and GPT [Radford et al., 2019], transformed the field. These models, and more recently large language models (LLMs), demonstrated unprecedented ability to generalize to unseen tasks, including text classification. Nevertheless, concerns remain: Xu et al. [2024] note that LLMs may suffer from benchmark contamination and overfitting, particularly when evaluation datasets overlap with pretraining distributions. Despite these limitations, their strong zero-shot and few-shot performance has spurred widespread interest in applying LLMs to real-world categorization.

To address inherent weaknesses, several frameworks have proposed structured prompting and model decomposition. For example, the CARP framework Sun et al. [2023] decomposes classification into simpler subtasks to improve performance in hierarchical taxonomies. SPIN Jiao et al. [2024] prunes internal neurons to emphasize task-relevant features, while Edwards and Camacho-Collados [2024] investigate in-context learning for text classification, finding that results vary substantially with taxonomy depth and complexity.

Although most prior work has focused on accuracy and overall classification performance, relatively few studies have examined zero-shot LLMs as a dedicated solution for categorization and thus have not addressed LLM-specific factors such as computational cost, hallucination, or category inflation. To our knowledge, no prior study has applied LLMs to the Interactive Advertising Bureau (IAB) taxonomy, despite its status as one of the most widely used categorization frameworks in internet marketing and advertising. This study addresses that gap by systematically evaluating ten major LLMs on an IAB-based categorization task, incorporating both traditional metrics and novel measures of real-world relevance.

3. Methodology

3.1. Dataset Compilation

The dataset consists of 8,660 unstructured textual samples drawn from diverse topical domains. The texts were sourced from open news corpora and manually categorized by expert annotators using the 690 general-purpose categories of the IAB 2.2 taxonomy [Kamen, 2025]. Each sample was assigned one or more categories judged by human experts to best represent its content. Owing to the hierarchical structure of the IAB taxonomy, these categories may be drawn from different tiers of the taxonomy.

3.2. LLM Selection and Configuration

We evaluated ten publicly available and widely used LLMs: Anthropic’s Claude 3.5; Google’s Gemini 1.5 and Gemini 2.0 Flash; Meta’s LLaMA 3.3 70B and LLaMA 3 8B; Mistral’s Mistral-Large-Latest Nano; xAI’s Grok; Groq-hosted DeepSeek’s DeepSeek; and Groq-hosted GPT OSS-20B and GPT OSS-120B. All models were accessed through their official APIs directly or using Groq hosting services. To ensure fair comparison, evaluations were conducted on the same dataset using a same prompting strategy.

3.3. Categorization Schema

Categorization follows the Interactive Advertising Bureau (IAB) taxonomy, version 2.2 [Lab, 2022]. The IAB framework functions as a de-facto industry standard in online advertising and internet marketing and is used by hundreds of internet publishers, advertisers, and technology providers. The taxonomy contains 690 general-purpose categories and subcategories organized in a four-tier hierarchy. We restrict our experiments to the general-purpose portion of taxonomy. Figure 1 illustrates the IAB hierarchical structure.

3.4. Prompting Procedure

We employ an iterative, hierarchy-aware prompting strategy that mirrors human taxonomy navigation. The model first selects a Tier-1 category, then progressively refines its choice through Tier-2, Tier-3, and Tier-4 prompts. To minimize hallucinations and ensure validity, taxonomy constraints such as allowable children, canonical identifiers, and formatting requirements are embedded directly into the prompts. All outputs are normalized to canonical IAB labels before scoring. The exact prompt template used in the experiments is provided below.

Your job is to categorize unstructured text according to the following list of categories. You will be given a certain amount of text from the text. Your response should contain only the categories, with no other text. If the text fits multiple categories, output them separated by a comma and a space. Categories are separated via comma. You may not output categories not in the list. If no categories fit the text, output 'None'. Categories: categories"

3.5. Hyperparameter Sweeps

To test sensitivity to decoding settings, we implemented a parameter-optimization harness over temperature, top-k, and maximum tokens. We conducted thousands of runs across models and tiers. Consistent with the short, schema-constrained nature of the outputs, hyperparameter variation produced negligible changes in categorization quality, while occasionally affecting verbosity or formatting. We therefore report main results under standardized defaults.

3.6. Evaluation Protocol (Overview)

Performance is measured on the 8,660-document benchmark using accuracy, precision, recall, and F1-score. We further compute three LLM-specific measures introduced in this work: hallucination ratio, category inflation ratio, and token-processing cost. Formal definitions and matching rules are detailed in Section 4 (Categorization Models and Evaluation Criteria). All invalid or out-of-taxonomy predictions are counted toward hallucination; multi-label emissions where a single label is required contribute to inflation.

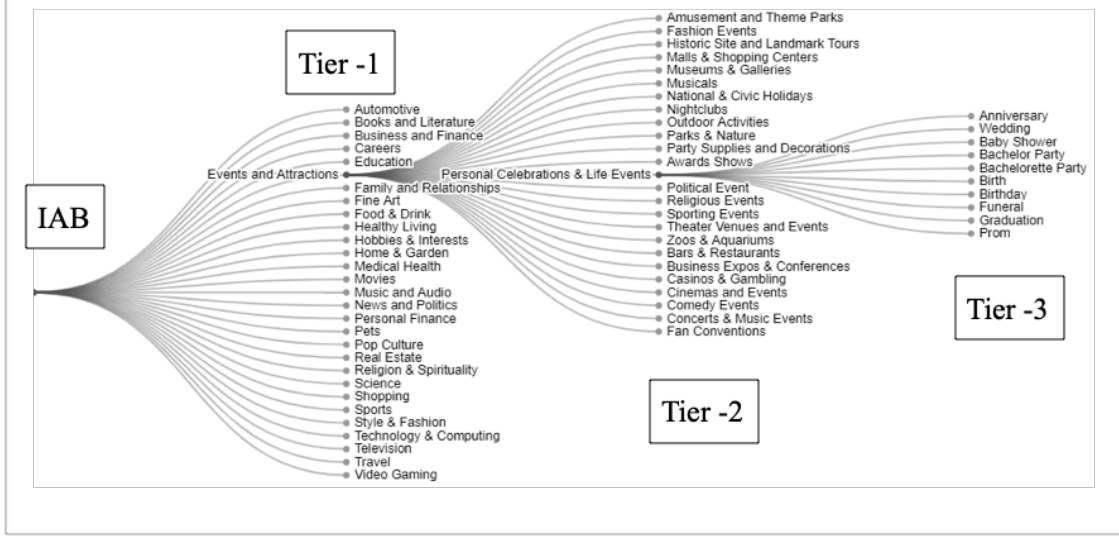


Figure 1: Overview of the IAB Taxonomy

4. Categorization Models and Evaluation Criteria

Let T denote a hierarchical taxonomy of categories.

The categorization of a text x is defined as generating a set of categories that describe its content:

$$\text{LLM}_{T,x} = \{t_1, t_2, \dots, t_n\} \quad (4.1)$$

Each predicted category may either belong to the taxonomy T or represent a hallucination (i.e., a category outside T).

For evaluation, we assume access to an expert (human or other reliable ground truth source) that categorizes the same text:

$$E(T, x) = \{e_1, e_2, \dots, e_m\}, \quad e_j \in T \quad (4.2)$$

We define:

- True Positives (TP): categories assigned by both the LLM and the expert.
- False Positives (FP): categories predicted by the LLM but absent in the expert set.
- False Negatives (FN): categories assigned by the expert but missed by the LLM.

Using these, the classic evaluation criteria are:

4.1. Classic Criteria

- Accuracy: overall proportion of correct predictions.

$$\text{Accuracy}_x = \frac{TP_x}{TP_x + FP_x + FN_x} \quad (4.3)$$

- Precision: fraction of predicted categories that are correct.

$$\text{Precision}_x = \frac{TP_x}{TP_x + FP_x} \quad (4.4)$$

- Recall: fraction of expert categories recovered by the model.

$$\text{Recall}_x = \frac{TP_x}{TP_x + FN_x} \quad (4.5)$$

- F1-score: harmonic mean of Precision and Recall, balancing both metrics into a single measure.

$$\text{F1}_x = \frac{2 \text{Precision}_x \text{Recall}_x}{\text{Precision}_x + \text{Recall}_x} \quad (4.6)$$

4.2. Additional Criteria for LLM-Based Categorization

To capture behaviors specific to LLMs, we introduce three additional metrics:

- Hallucination Ratio (HR): measures how often the LLM produces categories outside the taxonomy.

$$h_x = \{ t_i \in \text{LLM}_{T,x} : t_i \notin T \} \quad (4.7)$$

$$\text{HR}_x = \frac{|h_x|}{|\text{LLM}_{T,x}|} \quad (4.8)$$

- Category Inflation Ratio (IR): compares the number of predicted categories with the number assigned by the expert.

$$\text{IR}_x = \frac{|\text{LLM}_{T,x}|}{|E(T, x)|} \quad (4.9)$$

To avoid redundancy in hierarchical taxonomies, we apply the Parent Exclusion Rule (PER), which removes a parent if its child is present. The reduced inflation ratio is:

$$\text{IR}_x^* = \frac{|\text{PER}(\text{LLM}_{T,x})|}{|E(T, x)|} \quad (4.10)$$

- Price of Computation (Cost): quantifies monetary expense based on token usage.

$$\text{Cost}_x = c_{\text{in}} N_{\text{in}}^{(x)} + c_{\text{out}} N_{\text{out}}^{(x)} \quad (4.11)$$

where c_{in} and c_{out} are the prices per input and output token, and $N_{\text{in}}^{(x)}$, $N_{\text{out}}^{(x)}$ are the respective input and output token counts.

4.3. Corpus-level Aggregation

For the corpus of texts $D = \{x_1, \dots, x_k\}$, metrics are aggregated as:

$$\text{Metric}^{\text{macro}}(D) = \frac{1}{k} \sum_{i=1}^k \text{Metric}(x_i) \quad (4.12)$$

The total computation cost is:

$$\text{Cost}_D = \sum_{i=1}^k \text{Cost}(x_i) \quad (4.13)$$

Traditional metrics such as accuracy, precision, recall, and F1-score provide a well-established way to measure correctness. However, LLMs introduce unique challenges: they may generate categories not present in the taxonomy (hallucinations), assign too many or too few labels (inflation or under-assignment), and incur significant computational cost. By incorporating hallucination ratio, inflation ratio, and cost, we capture aspects of performance that are especially relevant to large-scale, real-world deployments of LLM-based categorization. This combination of classical and novel measures allows us to evaluate not only how correct the predictions are, but also how reliable, efficient, and usable the categorization is in practice.

Having established the evaluation framework and introduced both classical and LLM-specific performance criteria, we now turn to a comparative analysis of ten leading models.

5. Performance Evaluation

We analyzed ten popular zero-shot LLMs using the same 8,660-sample dataset and a unified prompt against a human-annotated benchmark. Comparison metrics include four classic metrics (accuracy, precision, recall, and F1-score) and three LLM-specific metrics (hallucination ratio, inflation ratio, and categorization cost). The Groq API was used to evaluate open-source models LLaMA 3 8B, LLaMA 3 70B, GPT 20B, GPT 120B, and DeepSeek R1. Other models were accessed through their providers’ APIs. Table 1 reports corpus-level aggregated accuracy, precision, recall, and F1-score. Classic metric performance varied as follows: 34% for accuracy; 42% for precision; 45% for recall; and 41% for F1-score. Claude 3.5 and the GPT models delivered the strongest classic-metric performance, whereas LLaMA 3 8B and Mistral lagged behind.

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 Flash	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
GPT-20B	0.52	0.55	0.47	0.71
GPT-120B	0.53	0.55	0.47	0.72

Table 2 lists each model’s input and output token costs per 1 million tokens (public rates as of September 2025). As expected, open-source token pricing was significantly lower than private models. Input costs

ranged from \$0.05 for LLaMA 3 8B to \$0.59 for LLaMA 3 70B per 1 million input tokens. Private LLM input pricing ranged from \$0.80 for Claude 3.5 up to \$8 for Mistral. Gemini 1.5/2.0 Flash, LLaMA 3 8B, and GPT 20B/120B were the most cost-efficient models in this experiment.

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

Model	Input Cost	Output Cost
Claude 3.5	\$0.80	\$4.00
Gemini 1.5 Flash	\$0.075	\$0.30
Gemini 2.0 Flash	\$0.10	\$0.40
Mistral	\$8.00	\$8.00
LLaMA 3 8B	\$0.05	\$0.08
LLaMA 3 70B	\$0.59	\$0.79
Grok	\$2.00	\$10.00
DeepSeek	\$0.27	\$1.10
GPT 20B	\$0.10	\$0.50
GPT 120B	\$0.15	\$0.75

Table 3 summarizes the average categorization cluster size before and after hallucination filtering, the hallucination ratio (HR), and the inflation ratio (IR). Hallucination ratio varies at 843%, and inflation ratio at 209%. GPT 120B demonstrated the lowest hallucination, approximately 40% lower than the next-best Grok. The benchmark dataset averaged 4.01 categories per article, indicating that all LLMs tended to overproduce labels relative to human annotators. Gemini 1.5/2.0, Grok, and GPT 20B/120B showed the lowest inflation. Both LLaMA 3 8B and 70B exhibited high hallucination and inflation, tending to over-generate labels and reduce focus.

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 Flash	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%
GPT 20B	5.41	5.34	1.3%
GPT 120B	5.36	5.32	0.7%

It is worth noting that hallucination filtering led to only marginal reductions in cluster size (typically less than 0.5 categories per sample) and slight improvements in model performance.

What is the best LLM overall? In price/performance terms, the Gemini and GPT families are the clear winners, and GPT 120B is preferred due to its very low hallucination ratio.

These observations motivate the following discussion of why classic metrics remain modest for zero-shot categorization, how task structure drives these outcomes, and where model scaling and orchestration strategies might change the picture.

6. Discussion

In our comparison, classic performance metrics remain modest. Given the rapid advances and increasing power of contemporary LLMs, one might expect substantially higher accuracy, precision, recall, and F1-scores. Instead, the zero-shot setup produced only moderate results, revealing a clear gap between general generative competence and targeted classification of arbitrary unstructured text within a complex hierarchical taxonomy.

Task specificity represents the primary bottleneck. Categorization requires compressing arbitrary, unstructured texts into a sparse, predefined label space. Although the IAB taxonomy is a well-designed and widely adopted framework, it includes only 690 generic categories—insufficient to reflect the full complexity and diversity of real-world content. This limitation becomes more apparent when compared to the DMOZ taxonomy (the Open Directory Project), which contains over 750,000 nodes spanning numerous hierarchical levels [Kosmopoulos, 2015, Hoek, 2021, ResearchGate, 2010]. Achieving human-expert-level IAB categorization therefore demands not only linguistic fluency but also a stable, taxonomy-aware world model capable of mapping dispersed semantic evidence into concise, non-overlapping labels. Standard next-token prediction pretraining does not guarantee this structured compression ability, particularly in zero-shot conditions.

Scaling and architectural improvements alone do not ensure performance gains. Newer or larger models—such as Gemini 1.5/2.0 and GPT 20B/120B—did not consistently outperform their predecessors across all classic metrics. We hypothesize that for categorization tasks, performance improvements diminish beyond a certain scale threshold; architectural refinement, instruction tuning, and prompt design likely play a greater role than raw model size.

Cost also significantly affects the practical usability of LLMs for categorization. Models such as Gemini 1.5/2.0, GPT 20B/120B, and DeepSeek combine low token pricing with competitive accuracy, enabling broader experimentation and deployment. When coupled with a low hallucination ratio, GPT 120B emerges as the strongest overall choice for balanced cost and performance. Hallucination filtering, however, provides only marginal benefits, typically reducing cluster size by less than half a category per sample, indicating that post-processing alone cannot substantially improve precision.

To address these limitations, we explored a different paradigm: LLM ensembles. In this approach, multiple models act as independent experts and make categorization decisions collaboratively. The ensemble framework, tested and optimized in a separate study, substantially improved overall categorization performance, completely eliminated hallucinations, and reduced category inflation. While a detailed analysis of ensemble design lies beyond the scope of this paper, these findings suggest that orchestration—rather than scale alone—offers a promising path toward achieving or surpassing human-expert consistency in large-scale text categorization.

7. Conclusion

Zero-shot LLMs can perform large-scale text categorization with consistent and reproducible behavior; however, their classic performance metrics remain modest, and their outputs tend to inflate category sets relative to human annotations. Cost-efficient models such as Gemini 1.5/2.0 and GPT 20B/120B offer strong price-to-performance profiles, with GPT 120B standing out for its notably low hallucination ratio. The core challenge remains structural: compressing semantically rich text into a sparse taxonomy demands disciplined, taxonomy-aware reasoning that current general-purpose pretraining does not yet guarantee.

Near-term progress is most likely to come from orchestration rather than scale. The collaborative use of multiple LLMs organized in an ensemble has proven to significantly improve categorization quality while maintaining efficiency. The additional computational cost of ensemble-based categorization appears to be

an attractive trade-off for its multiple benefits, complete elimination of hallucinations, reduced category inflation, and a consistent level of performance that may match or even surpass human experts.

References

- Anthropic. Claude AI: Safe and scalable AI by anthropic. <https://www.anthropic.com/claude>, 2025. Accessed: 2025-03-25.
- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.
- DeepSeek AI. Deepseek chat: AI-powered conversational model. <https://chat.deepseek.com>, 2025. Accessed: 2025-03-25.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.
- 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.
- Lisa Hoek. Web classification using DMOZ. Bachelor’s thesis, Radboud University, 2021. URL https://www.cs.ru.nl/bachelors-theses/2021/Lisa_Hoek____1009553____Web_classification_using_DMOZ.pdf. Accessed 2025-10-07.
- 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>.
- Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML)*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142, Berlin, Heidelberg, 1998. Springer. doi: 10.1007/BFb0026683.
- Ariel Kamen. Unstructured text dataset. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/ajkamen/Ensemble_Categorization. Accessed 2025-10-07.
- Aris Kosmopoulos. *Large Scale Hierarchical Text Classification*. PhD thesis, National Centre for Scientific Research “Demokritos”, Athens, Greece, 2015. URL https://www.iit.demokritos.gr/sites/default/files/akosmopoulos_phd_thesis_final.pdf. Accessed 2025-10-07.
- IAB Tech Lab. Content taxonomy guidelines, 2022. URL <https://iabtechlab.com/standards/content-taxonomy>.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed 2025-10-07.

ResearchGate. Point-based visualization of the DMOZ hierarchy (765,328 nodes). Web page, 2010. URL https://www.researchgate.net/figure/Point-based-visualization-of-the-DMOZ-hierarchy-It-contains-765-328-nodes-of-whi-fig1_224197967. Accessed 2025-10-07.

F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

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>.