Information Processing and Management 63 (2026) 104476

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Classifying illicit dark web content through zero-shot prompting: An empirical study with GPT models

Adrián Domínguez-Díaz*, Luis de-Marcos, Víctor-Pablo Prado-Sánchez, Daniel Rodriguez, José-Javier Martínez-Herráiz

Department of Computer Science, Polytechnic School, University of Alcalá, Campus Universitario, Ctra. Madrid-Barcelona km, 33, 600, Alcalá de Henares, 28805, Madrid, Spain

ARTICLE INFO

Keywords:
Dark web
Zero-shot
GPT
Text classification
Illicit activities

ABSTRACT

This study evaluates the classification performance of four GPT-based models (GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, and o4-mini) under zero-shot prompting conditions on the complete, multilingual CoDA dataset of Dark Web content, comprising 10 illicit activity categories. The models GPT-4.1, GPT-4.1-mini, and o4-mini achieve a weighted F1 score of 0.885, surpassing prior zero-shot baselines on this dataset. Stability analysis using *TARa@10* demonstrates high output consistency for GPT-4.1 (0.964) and GPT-4.1-mini (0.970), indicating their reliability for operational use. Multilingual evaluation reveals only a modest English vs. non-English performance gap for GPT-4.1 (0.031), while other models perform comparably across languages. The strongest results appear in *Drugs*, *Gambling*, and *Porn* (F1 > 0.9), whereas lower scores are observed in ambiguous or overlapping categories like *Violence* (F1 ≤ 0.76) or *Crypto* (F1 ≤ 0.84). A qualitative review of misclassifications suggests that some model predictions align with reasonable semantic interpretations, potentially highlighting annotation inconsistencies. This work establishes a performance baseline for GPT-based models in zero-shot classification of multilingual Dark Web content and underscores the importance of clear category definitions for effective deployment.

1. Introduction

The Dark Web offers an unprecedented degree of anonymity to its users, facilitating both legitimate and illicit activities (Cilleruelo et al., 2021). As criminal operations have increasingly migrated to online platforms, the ability to automatically classify Dark Web content has become a critical need for cybersecurity analysts and law enforcement agencies. However, this task remains highly challenging due to the lack of structured data, the prevalence of obfuscation strategies, and the linguistic heterogeneity of Dark Web platforms (Al-Nabki et al., 2017; Avarikioti et al., 2018).

Recent advances in supervised learning have demonstrated strong performance on curated datasets. For instance, DarkBERT has achieved classification accuracies as high as 0.945 on English-language data (Jin et al., 2023), while more recent architectures such as TextCNN combined with topic modeling have reached up to 0.962 (Shin et al., 2024). Nonetheless, such approaches depend heavily on large quantities of labeled data (Chen et al., 2024) whose collection from the Dark Web may raise significant technical, ethical, and legal challenges. Moreover, existing supervised models are typically trained and evaluated exclusively on English

* Corresponding author.
E-mail addresses: adrian.dominguez@uah.es (A. Domínguez-Díaz), luis.demarcos@uah.es (L. de-Marcos), victor.prado@uah.es (V.-P. Prado-Sánchez), daniel.rodriguezg@uah.es (D. Rodriguez), josej.martinez@uah.es (J.-J. Martínez-Herráiz).

https://doi.org/10.1016/j.ipm.2025.104476
Received 10 April 2025; Received in revised form 5 November 2025; Accepted 5 November 2025
Available online 7 November 2025

content, overlooking the multilingual nature of real-world Dark Web forums. These limitations have sparked growing interest in applying zero-shot classification techniques using large language models (LLMs) such as GPT, which promise high versatility without requiring task-specific training. In a zero-shot setting, models are directly applied to a task without exposure to task-specific labeled examples during training, making this approach particularly relevant when annotated data is scarce or unavailable (Brown et al., 2020; Chae & Davidson, 2025).

Zero-shot learning offers key strengths for content moderation tasks, including the ability to generalize to new domains without fine-tuning, as demonstrated in hate speech detection where GPT models outperform supervised approaches when training data is misaligned with test sets (Bauer et al., 2024). It also reduces ethical risks associated with handling sensitive data during annotation (Kocoń et al., 2023). However, limitations include sensitivity to prompt formulation, which often leads to inconsistent performance, and reduced accuracy on subjective or pragmatically complex tasks compared to supervised state-of-the-art (SOTA) models, often resulting in a 25% quality drop (Kocoń et al., 2023). In content moderation specifically, zero-shot GPT models achieve competitive results (e.g., ≈ 0.800 accuracy) but may confuse related categories like hatefulness and offensiveness, treating them as subsets of toxicity, and exhibit biases from training guidelines (Li et al., 2024).

We focus on GPT models for this study because they are among the most widely used LLMs in research and practice, with their availability through commercial APIs offering a practical baseline for evaluating zero-shot classification in the Dark Web domain. Despite the potential of other LLM families, GPT's established versatility across tasks makes it a suitable starting point. However, the effectiveness of GPT in the Dark Web context remains largely untested, particularly given its unique challenges — intentional obfuscation, specialized criminal jargon, and multilingual usage — which pose significant obstacles for zero-shot models (Chen et al., 2024; Prado-Sánchez et al., 2024). Additionally, the inherent non-determinism of LLMs, where identical inputs may yield different outputs due to sampling randomness (Atil et al., 2025), raises concerns about prediction stability and reproducibility in operational settings.

To address these open issues, this study conducts the first systematic evaluation of GPT models (GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, and o4-mini) in the task of zero-shot classification of Dark Web content, using the CoDA dataset. This zero-shot focus is motivated by the domain's data scarcity and ethical constraints, allowing us to leverage GPT's generalization strengths while probing its limitations in a challenging, multilingual setting. We aim to answer three research questions: (1) How well do GPT models perform in zero-shot classification of Dark Web texts? (2) How stable are GPT predictions across repeated runs? (3) What types of classification errors are most common across different illicit activity categories?

Our contributions are threefold. First, we establish the first benchmark of GPT models for multilingual Dark Web classification, directly addressing RQ1. Second, we provide the first empirical assessment of prediction stability in this domain, addressing RQ2. Third, we conduct a detailed error analysis that uncovers systematic misclassification patterns tied to lexical ambiguity and domain-specific obfuscation tactics, thereby addressing RQ3. Together, these contributions offer practical guidance for researchers and practitioners interested in using zero-shot LLMs to augment or complement traditional supervised approaches in cybersecurity settings.

The remainder of this paper is organized as follows. Section 2 reviews related work on Dark Web content classification and zero-shot learning with LLMs. Section 3 describes the dataset, experimental setup, evaluation metrics, and statistical analyses. Section 4 presents our quantitative findings, error analyses, and stability results. Section 5 discusses the broader implications and limitations of our study. Finally, Section 6 concludes the paper and outlines future directions.

2. Literature review

The Dark Web refers to a collection of websites that are publicly available but intentionally hidden from search engines. Accessing these sites typically requires specific software, configurations, or authorization. Anonymization tools like Tor provide users with a high degree of privacy, making the Dark Web an attractive platform for a wide range of activities—both legal and illegal (Cilleruelo et al., 2021). A significant concern is the proliferation of illicit activities, including the sale of drugs, weapons, and counterfeit products, as well as hacking services. Classifying the content of Dark Web documents is crucial for law enforcement and cybersecurity professionals seeking to identify, track, and mitigate such activities (He et al., 2019). However, this task is complicated by the linguistic diversity, lack of structured data, and rapidly evolving nature of the content (Al-Nabki et al., 2017; Avarikioti et al., 2018).

In recent years, machine learning—particularly natural language processing (NLP)—has played a key role in automating Dark Web content classification. A growing body of literature shows the potential of large language models (LLMs) across various domains, including medicine, engineering, social sciences, and the humanities (Fan et al., 2024). With increasingly powerful pre-trained models, researchers have applied a range of classification techniques to detect illegal activities on the Dark Web (Jin et al., 2022). These include bidirectional LSTMs (Elma et al., 2024), support vector machines (SVMs) (Alaidi et al., 2022; Elma et al., 2024), and pre-trained transformers like BERT, RoBERTa, Mistral, and LLaMA (Chen et al., 2024; Jin et al., 2023). Most of these approaches have achieved high classification accuracy, with reported F1 scores ranging from 0.810 to 0.960, depending on the category and model. However, all supervised learning studies to date have focused exclusively on classifying English-language documents from the CoDA dataset, limiting their applicability to multilingual Dark Web content.

Early approaches relied heavily on supervised learning, using manually labeled datasets. One of the earliest contributions was the Darknet Usage Text Addresses (DUTA) dataset by Al-Nabki et al. (2017), which includes 6831 documents across 26 categories such as drugs, weapons, and counterfeit goods. Supervised models trained on this dataset, such as logistic regression with TF-IDF features, achieved an F1 score of up to 0.970 in a subset of nine categories. This dataset was later expanded to DUTA-10K with

Table 1
Classification performance (F1 score) of prior studies using LLMs and supervised models for Dark Web content on the CoDA dataset, including details of models, dataset, and experimental setup.

Study	Models	Setting/Dataset	F1
Jin et al. (2023)	DarkBERT	Supervised, CoDA (English only)	0.945
Chen et al. (2024)	LLaMA-3, Mistral, RoBERTa, DarkBERT	Zero-shot prompting vs. supervised, CoDA (English only)	0.748–0.945
Prado-Sánchez et al. (2024)	GPT-3.5-turbo	Zero-shot prompting, CoDA (*Porn* excluded, max 16,385 tokens)	0.805
Shin et al. (2024)	TextCNN + topic modeling	Supervised, CoDA (English only)	~0.962

10,367 documents (Al-Nabki et al., 2019). However, DUTA and its extended version suffer from important limitations, including severe class imbalance, high document similarity, and a large number of empty or extremely short documents (Al-Nabki et al., 2017; Jin et al., 2022). These issues raise concerns about inflated performance metrics due to train-test leakage, making these datasets unsuitable for robust model evaluation.

To address these limitations, Jin et al. (2022) introduced the Comprehensive Darkweb Annotations (CoDA) dataset, comprising 10,000 documents distributed across balanced and well-defined illicit activity categories. Using fine-tuned BERT models, they achieved a weighted F1 score of 0.925 when classifying English-language documents across 10 illicit activity categories. This was later improved by DarkBERT, a domain-specific model trained on a large Dark Web corpus, which reached 0.945 (Jin et al., 2023). These results highlight the importance of domain adaptation when applying LLMs to specialized content.

Building on this, Chen et al. (2024) carried out an extensive evaluation of both supervised and zero-shot classifiers on CoDA. In the supervised setting, they compared a range of transformer architectures, including RoBERTa, DarkBert, as well as fine-tuned LLaMa-3 and Mistral models, reporting weighted F1 scores of above 0.940 on English-language documents. Their study confirmed that domain adaptation (as in DarkBERT) and fine-tuning remain the most reliable strategies for maximizing classification accuracy, but also emphasized their dependency on large volumes of high-quality labeled data.

More recently, Shin et al. (2024) pushed supervised approaches further by introducing a hybrid architecture that combines TextCNN with topic modeling to capture both local lexical patterns and broader semantic features in Dark Web texts. Their experiments on CoDA (restricted to English content) achieved a weighted F1 score of 0.962, the highest reported to date for supervised classifiers. This progression — from BERT-based fine-tuning to more specialized hybrid models — demonstrates the upper bound of what can be achieved with supervised learning under controlled, monolingual settings.

Despite these advances, training and evaluating fine-tuned models requires large, up-to-date datasets, which are costly, risky, and often infeasible to collect in the Dark Web context. Moreover, the quality of these datasets, including issues like annotation errors, class imbalance, or inconsistent labeling, can significantly impact the performance of supervised classifiers. For example, Cascavilla et al. (2022) reported a weighted F1 score of 0.800 on a diverse multi-source Dark Web dataset using a fine-tuned BERT model, significantly below results from other studies based on the CoDA dataset. Finally, it is important to note that all supervised experiments were computed on English-language documents, excluding non-English documents from evaluation.

An alternative approach that avoids the need for labeled data and domain-specific fine-tuning is zero-shot or few-shot prompting. These paradigms use pre-trained LLMs to classify text by simply describing the task and categories in natural language. Zero-shot models receive only the task description and input text, while few-shot models include a few labeled examples in the prompt. This represents a shift away from supervised learning by leveraging general-purpose language understanding.

In addition to their supervised experiments, Chen et al. (2024) conducted one of the first systematic evaluations of zero-shot LLM-based classifiers on a subset of the CoDA. Their experiments tested LLaMA-3 and Mistral models under a zero-shot approach, comparing them with their respective fine-tuned versions as well as supervised BERT variants. While supervised models such as DarkBERT or fine-tuned LLaMa-3 achieved weighted F1 scores above 0.940, zero-shot models lagged behind; for example, LLaMA-3 under a zero-shot approach reached 0.748 F1 score. They also reported large performance gaps across categories, with technical domains such as malware showing the highest error rates. Importantly, their evaluation was restricted to the English portion of the CoDA dataset, excluding multilingual content. When extending the analysis to other multilingual datasets that combined English and Chinese documents, they found that BERT-based models performed substantially worse, whereas fine-tuned LLaMA and Mistral models were comparatively more robust.

Building on this line of work, Prado-Sánchez et al. (2024) evaluated GPT-3.5-turbo in a zero-shot setting on part of the CoDA dataset. Their study used carefully crafted prompts describing each of the ten illicit activity categories, and compared model predictions against human annotations. They reported a weighted F1 score of 0.805, notably below fine-tuned approaches but within the range of intercoder agreement among human annotators. However, their experiments excluded the *Porn* category and constrained the input context to 16,385 tokens, which is smaller than the size of some documents in the dataset. These factors may have limited the comparability of their evaluation with respect to both the dataset and other LLM-based approaches.

As shown in Table 1, prior research on the CoDA dataset has explored both supervised and zero-shot approaches, with supervised fine-tuned models generally outperforming LLM-based prompting. However, most supervised studies have been restricted to English content and rely on resource-intensive annotation pipelines, limiting their scalability and applicability in real-world, multilingual Dark Web contexts. Likewise, existing zero-shot evaluations have been either partial (restricted to subsets of CoDA) or limited to specific model families, leaving important questions unanswered regarding the performance, robustness, and error patterns of GPT models—the most widely used LLMs in practice. By systematically benchmarking multiple GPT variants in a fully multilingual setting, and by examining prediction stability alongside detailed error analyses, our study extends prior work and addresses critical gaps at the intersection of Dark Web classification, zero-shot learning, and applied LLM evaluation.

Further evidence from other domains reinforces the potential of zero-shot prompting. GPT models have demonstrated the ability to perform text classification in multiple domains by relying on a combination of general knowledge and task-specific instructions provided in natural language prompts (Kalyan, 2024), although its performance is generally below state-of-the-art techniques for each problem (Kocoń et al., 2023). Studies in financial document classification show that GPT models can perform competitively with minimal prompting (Loukas et al., 2023). GPT-based models have also been effective in classifying hateful, toxic, or offensive content in social media, often achieving around 0.800 accuracy using few-shot prompts (Bauer et al., 2024; Gupta, 2022; Li et al., 2024). GPT models have also demonstrated their effectiveness in classifying names by gender (Domínguez-Díaz et al., 2024; Goyanes et al., 2024) and sentiment analysis (Belal et al., 2023; Mathebula et al., 2024).

Nonetheless, zero-shot prompting presents clear challenges. Reiss (2023) emphasize this non-determinism, which complicates deployment in sensitive or production contexts. While general-purpose LLMs cover a broad range of knowledge, they often struggle with domain-specific terminology, particularly in fields like law (Padiu et al., 2024), medicine (Tariq et al., 2024) or the Dark Web, which features distinctive jargon, multilingual content, and unconventional phrasing. Moreover, the rapid pace at which LLMs evolve—with frequent releases of increasingly capable models—creates a moving target for researchers and practitioners alike, making continuous evaluation essential to understand current capabilities and limitations.

Our work contributes to the current literature by offering the first comprehensive evaluation of the most recent GPT-based models (GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, and o4-mini) for zero-shot classification on the full CoDA dataset, including multilingual texts. Unlike prior studies that focused on partial datasets or older GPT versions, this study examines the classification performance and stability of these models across all 10 illicit activity categories in CoDA, establishing a baseline and providing insights into their applicability for Dark Web content classification.

3. Methodology

3.1. Data sources

This research utilized the Comprehensive Darkweb Annotations (CoDA) dataset, developed to support the analysis of linguistic features in Dark Web content (Jin et al., 2022). CoDA consists of 10,000 unique documents collected from various Dark Web marketplaces and forums, each manually labeled into one of ten thematic categories. The class distribution is not perfectly balanced: *Others* (2919; 29.2%), *Porn* (1205; 12.1%), *Drugs* (1166; 11.7%), *Financial* (1003; 10.0%), *Gambling* (787; 7.9%), *Crypto* (761; 7.6%), *Hacking* (649; 6.5%), *Arms* (599; 6.0%), *Violence* (485; 4.9%), and *Electronic* (426; 4.3%). This imbalance motivates the use of weighted metrics and per-class reporting in our evaluation (see Table 6).

Notably, the dataset includes documents in multiple languages, primarily English (8855 documents), with the remaining 1145 in languages such as Russian, German, and French, reflecting the multilingual nature of Dark Web content. While its class distribution is not perfectly uniform, CoDA is substantially better balanced than earlier corpora such as DUTA and DUTA-10K, which suffered from severe class imbalance and redundancy. As a result, CoDA has become one of the most widely used benchmarks for evaluating automated classification techniques in this domain Chen et al. (2024), Jin et al. (2023), Sennad et al. (2025), Shin et al. (2024).

The use of CoDA in this study facilitates meaningful comparisons with prior work on zero-shot and supervised classification approaches, ensuring compatibility with existing benchmarks and methodological standards. Furthermore, no additional preprocessing or filtering was applied to the dataset apart from its conversion to UTF-8 encoding, which was necessary to prevent potential character decoding issues during inference. All document texts, including those in non-English languages, were used in their original form to preserve the authenticity of the linguistic content.

3.2. Classification procedure

In our experiments, we assessed four open-weight large language models—GPT-4.1-nano, GPT-4.1-mini, GPT-4.1, and o4-mini—under zero-shot classification conditions through the OpenAI API. The zero-shot setting was selected to evaluate model capabilities in realistic low-resource scenarios where labeled data may be unavailable or difficult to obtain, as is often the case in the Dark Web domain. This approach also facilitates direct comparison with prior zero-shot evaluations reported in the literature (Chen et al., 2024).

The four models evaluated in this work represent different configurations of the GPT-4 family, varying mainly in parameter scale and computational cost. GPT-4.1-nano and GPT-4.1-mini are lightweight models optimized for efficiency, while GPT-4.1 is a larger model designed for higher accuracy. o4-mini is part of the o-series optimized for cost-effective reasoning. Although exact parameter counts are not publicly disclosed, these variants are known to differ in capacity, context length, and inference speed, which directly impact classification performance. All models expose a maximum context window of 1,048,576 tokens, except o4-mini, which is limited to 200,000 tokens. These context windows are enough to process any chosen document of the CoDA dataset in a single query to the OpenAI API.

To minimize output variability and promote response stability across repeated inferences, all models were configured with a `temperature` value of 0, reducing sampling randomness as much as possible. This setting was applied to GPT-4.1-nano, GPT-4.1-mini, and GPT-4.1. However, the o4-mini model does not expose a `temperature` parameter due to its architectural design.

Instead, o4-mini was configured with its `reasoning_effort` parameter set to `low`, corresponding to the least exploratory mode available for this model.

Each model was prompted with the same English-language instruction template, which included a task description, the full list of category definitions, and a request to classify the content of a given darknet document into one of the ten predefined CoDA classes. The descriptions used for each category were quoted verbatim from the original annotation guidelines published by Jin et al. (2022), ensuring full alignment with the human classification protocol described in their study. The complete prompt is provided in Appendix to support reproducibility.

To facilitate automated parsing and ensure consistency across model outputs, each prompt explicitly instructed the model to return its classification decision in a structured JSON format, using a single key named "`category`" and a value corresponding to one of the ten predefined class labels. This constraint minimized ambiguity in the output format and simplified downstream evaluation by reducing the need for additional text normalization or pattern matching.

All classification inferences were performed document by document, without batching, to avoid interactions between inputs and to ensure consistency across models. This strategy also allowed us to precisely monitor token usage and model refusals on a per-document basis. While batching can improve computational efficiency, it was avoided here to maintain strict control over inference conditions and comparability with prior zero-shot evaluations (Chen et al., 2024; Prado-Sánchez et al., 2024).

Following each classification test, we applied two post-processing steps to validate outputs and handle moderation-related exceptions. First, all model responses were checked against the predefined list of valid CoDA labels. Any output that did not exactly match one of the ten expected categories — including lexical variants (e.g., "`Electronics`" instead of *Electronic*) or unrelated terms — was classified as invalid and counted as a misclassification in the evaluation. Second, we recorded instances in which models refused to provide a classification due to content moderation policies, particularly in sensitive categories such as *Porn* or *Violence*. These refusals were likewise treated as misclassifications, ensuring that all types of classification failure were reflected in the reported metrics.

3.3. Evaluation metrics

We adopt weighted Precision, Recall, and F1 score as our primary evaluation metrics, consistent with prior studies on CoDA and Dark Web text classification (Chen et al., 2024; Jin et al., 2023, 2022; Prado-Sánchez et al., 2024). These metrics provide a transparent way to capture both the ability to correctly identify positive instances (Recall) and to avoid false alarms (Precision), which is especially relevant in high-risk applications such as cybersecurity and law enforcement. The F1 score balances these two aspects and is therefore widely regarded as the most informative single measure of classifier performance in this context.

Interpreting these metrics requires considering both the evaluation setting and prior benchmarks. In supervised settings, state-of-the-art models such as DarkBERT or TextCNN+topic modeling have achieved weighted F1 scores above 0.940 (Jin et al., 2023; Shin et al., 2024), establishing an approximate upper bound under favorable conditions. In contrast, zero-shot models typically report lower performance, with weighted F1 scores around 0.750–0.800 (Chen et al., 2024; Prado-Sánchez et al., 2024), making improvements within this range meaningful despite being below supervised standards. Moreover, variation across categories is critical: for example, misclassifying *Drugs* as *Financial* may be less consequential than failing to detect *Violence*, underscoring the need to analyze per-class results in addition to global metrics. Our evaluation protocol aligns with emerging best practices in evidence-centered benchmark design for NLP (Liu et al., 2024), by explicitly reporting per-class metrics and discussing their implications rather than relying solely on aggregate scores.

We computed weighted Precision, Recall, and F1 score, along with per-category F1 scores (Sokolova & Lapalme, 2009). To account for class imbalance and ensure that global performance metrics reflected the relative frequency of each category, we employed a weighted averaging scheme. Let C be the set of classes, s_c the number of instances in class c, and $S = \sum_{c \in C} s_c$ the total number of instances.

For each class $c \in C$, define:

$$\mathrm{TP}_c = \text{true positives}, \quad \mathrm{FP}_c = \text{false positives}, \quad \mathrm{FN}_c = \text{false negatives}$$

Then, per-class Precision, Recall, and F1 score are:

$$\mathrm{Precision}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c} \quad \text{if } \mathrm{TP}_c + \mathrm{FP}_c > 0, \text{ else } 0 \tag{1}$$

$$\mathrm{Recall}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c} \quad \text{if } \mathrm{TP}_c + \mathrm{FN}_c > 0, \text{ else } 0 \tag{2}$$

$$\mathrm{F1}_c = \frac{2 \cdot \mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c} \quad \text{if } \mathrm{Precision}_c + \mathrm{Recall}_c > 0, \text{ else } 0 \tag{3}$$

The weighted Precision, Recall, and F1 score are defined as:

$$\mathrm{Precision}_{\mathrm{weighted}} = \sum_{c \in C} \frac{s_c}{S} \cdot \mathrm{Precision}_c \tag{4}$$

$$\mathrm{Recall}_{\mathrm{weighted}} = \sum_{c \in C} \frac{s_c}{S} \cdot \mathrm{Recall}_c \tag{5}$$

Table 2
Comparison of class distributions between the full dataset of 10,000 documents and the 500-document sample. Values are ordered by frequency in the full dataset.

Category	Full dataset (n)	Full dataset (%)	Sample (n)	Sample (%)
Others	2919	29.2%	151	30.2%
Porn	1205	12.1%	73	14.6%
Drugs	1166	11.7%	52	10.4%
Financial	1003	10.0%	47	9.4%
Gambling	787	7.9%	38	7.6%
Crypto	761	7.6%	31	6.2%
Hacking	649	6.5%	37	7.4%
Arms	599	6.0%	31	6.2%
Violence	485	4.9%	26	5.2%
Electronic	426	4.3%	14	2.8%

$$\mathrm{F1}_{\mathrm{weighted}} = \sum_{c \in C} \frac{s_c}{S} \cdot \mathrm{F1}_c \tag{6}$$

These formulas ensure that each class contributes proportionally to the overall metric, providing a more representative and balanced assessment of model performance across all categories.

To specifically address potential differences in model performance across languages, we also computed weighted F1 scores separately for documents written in English and for those written in other languages (e.g., Russian, German, French). This stratified evaluation allowed us to directly compare classification accuracy across linguistic groups while controlling for the imbalance in language distribution within the CoDA dataset (8855 English vs. 1145 non-English documents). This procedure ensures that our analysis of multilingual capability is empirically grounded and that performance disparities between languages are explicitly quantified.

To assess model stability, we adopted the *TARa@10* metric (Total Agreement Rate over Parsed-Out Answers, with 10 repetitions), originally proposed by Atil et al. (2025). To the best of our knowledge, this is one of the first metrics specifically designed to measure the stability of LLM outputs under repeated queries with a fixed prompt. *TARa@10* quantifies the proportion of identical outputs obtained across ten repeated inferences. This zero-shot variance was evaluated by sampling 500 random CoDA documents, querying the models ten times for each document, and computing the output agreement rate. The choice of ten repetitions and a sample size of 500 was made to ensure sufficient precision in the estimation of agreement rates while maintaining computational feasibility, given the repeated inferences required per document. Table 2 shows the category distribution of the 500-document sample compared to the full dataset. A post-hoc power analysis was conducted to evaluate whether the sample size was sufficient to detect statistically significant differences in output agreement between models. The results indicated that the design had adequate power (> 0.8) to detect small effect sizes (greater than 0.1, in terms of Cohen's d), thereby supporting the reliability of our comparative findings.

In addition, confusion matrices were constructed for each model using the entire test set. The counts in these matrices were row-normalized to represent the proportion of true instances for each class that were classified into each predicted category, facilitating the identification of misclassifications patterns across classes. These normalized counts were estimated through bootstrap resampling to provide confidence intervals around the observed proportions, with values rounded to three decimal places for clarity. Analyzing row-normalized confusion matrices with bootstrap confidence intervals enabled us to pinpoint specific categories that models tended to confuse, revealing semantic weaknesses under zero-shot classification.

To gain further insight into the underlying causes of the identified misclassification patterns, we conducted a qualitative analysis. A random sample comprising 20% of the misclassified documents was manually reviewed to explore common linguistic or semantic features that may have contributed to the errors. This analysis helped contextualize quantitative trends and supported a deeper interpretation of model limitations in handling ambiguous or overlapping categories.

3.4. Statistical analysis

All reported metrics were accompanied by 95% confidence intervals estimated via nonparametric bootstrap resampling. Specifically, each metric was recalculated across 1000 bootstrap samples drawn with replacement from the evaluation set, following the bias-corrected and accelerated (BCa) method, implemented via the `bootstrap` function in the Python `scipy` library.

In addition to constructing confidence intervals for individual metrics, we applied bootstrap testing to compare differences between models. Pairwise comparisons of global, per-category, and per language F1 scores, as well as *TARa@10* stability metrics, were performed by generating the bootstrap distribution of score differences across 1000 resamples.

The choice of 1000 bootstrap resamples was made to balance statistical precision and computational feasibility, as this number is widely regarded as sufficient for stable estimation of 95% confidence intervals and p-values in most practical settings (Davison & Hinkley, 1997; Efron & Tibshirani, 1993).

The bias-corrected and accelerated (BCa) method was employed due to its established robustness in the statistical literature, as it corrects for bias and skewness in the sampling distribution, providing more accurate confidence intervals compared to the standard percentile method, particularly for non-normal or complex data distributions (Davison & Hinkley, 1997; Efron, 1987).

Table 3
Classification stability (*TARa@10*) and performance (F1 score, Precision, Recall) with 95% confidence intervals (CIs) for evaluated models.

Model	Stability (*TARa@10*)	F1	Precision	Recall
GPT-4.1-nano	0.928 [0.904, 0.948]	0.793 [0.785, 0.802]	0.829 [0.821, 0.836]	0.792 [0.783, 0.800]
GPT-4.1-mini	0.970 [0.952, 0.982]	0.886 [0.880, 0.892]	0.890 [0.885, 0.897]	0.886 [0.880, 0.892]
GPT-4.1	0.964 [0.944, 0.978]	0.885 [0.879, 0.891]	0.889 [0.883, 0.895]	0.886 [0.879, 0.892]
o4-mini	0.902 [0.874, 0.926]	0.886 [0.880, 0.892]	0.891 [0.885, 0.897]	0.886 [0.880, 0.892]

Table 4
Pairwise comparisons of model stability (*TARa@10*) with 95% confidence intervals (CIs), Cohen's d, effect size interpretation, achieved power, and BH-adjusted p-value.

Model A	Model B	Mean Δ	Cohen's d	Effect Size	Power	p
GPT-4.1-mini	o4-mini	0.068 [0.044, 0.096]	0.28	Small	1.000	<0.001***
GPT-4.1	o4-mini	0.062 [0.040, 0.088]	0.25	Small	1.000	<0.001***
GPT-4.1-nano	GPT-4.1-mini	−0.042 [−0.068, −0.018]	−0.19	Negligible	0.990	0.001**
GPT-4.1-nano	GPT-4.1	−0.036 [−0.062, −0.010]	−0.16	Negligible	0.945	0.012*
GPT-4.1-nano	o4-mini	0.026 [−0.004, 0.056]	0.09	Negligible	0.548	0.152
GPT-4.1-mini	GPT-4.1	0.006 [−0.014, 0.026]	0.03	Negligible	0.116	0.668

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

For each pairwise comparison, statistical significance was tested by calculating raw p-values from the bootstrap distribution of the mean differences for F1 scores and *TARa@10* stability metrics, following the methodology described by Davison and Hinkley (1997). The two-sided p-value is defined as:

$$p_{\text{bootstrap}} = 2 \cdot \min\left(\frac{1}{B}\sum_{i=1}^{B}\mathbb{I}(\delta_i \leq 0), \frac{1}{B}\sum_{i=1}^{B}\mathbb{I}(\delta_i \geq 0)\right) \tag{7}$$

where $B = 1000$ is the number of bootstrap resamples, δ_i is the difference in means for the ith resample, and $\mathbb{I}$ is the indicator function.

To address multiple comparisons, we applied the Benjamini–Hochberg (BH) procedure to control the false discovery rate (FDR) at a significance level of $\alpha = 0.05$. The BH correction was performed on the raw p-values from bootstrap comparisons, using the `multipletests` function from the Python `statsmodels` library. All p-values from global and per-category comparisons were concatenated, corrected using the FDR-BH method, and reassigned to their respective comparisons, with adjusted p-values and rejection decisions saved for analysis (Benjamini & Hochberg, 1995).

Effect sizes for pairwise comparisons were quantified using Cohen's d, adapted for paired samples to account for the repeated measures design, as described by Cohen (1988). Effect sizes were classified as negligible ($|d| < 0.2$), small ($0.2 \leq |d| < 0.5$), medium ($0.5 \leq |d| < 0.8$), or large ($|d| \geq 0.8$).

A post-hoc power analysis was conducted for all pairwise tests of F1 scores and *TARa@10* stability metrics, using the `TTestPower` class from the `statsmodels` library, assuming a two-sided t-test for paired samples with $\alpha = 0.05$. For global F1 score comparisons, the sample size was set to 10,000 (the full CoDA dataset). For stability comparisons, the sample size was 500 (the subset used for *TARa@10*). For per-category F1 scores, sample sizes were determined by the number of documents per category in the CoDA dataset, extracted using `pandas`. The effect size (Cohen's d) for each comparison was used to compute achieved power, ensuring sufficient sample sizes to detect observed differences with high confidence (Cohen, 1988).

4. Results

Table 3 presents the classification performance metrics (F1 score, Precision, and Recall) and stability (measured as *TARa@10*) for all evaluated models, along with 95% confidence intervals obtained via bootstrapping. Overall, GPT-4.1-mini, GPT-4.1, and o4-mini achieved the best classification performance, obtaining nearly identical F1 scores (≈ 0.886) as well as Precision and Recall values. In contrast, GPT-4.1-nano exhibited the lowest performance, with an F1 score of 0.793 and lower Precision and Recall. Despite this, it still maintained a relatively high stability score (0.928). Interestingly, although o4-mini matched the top models in F1 score (0.886), its Stability was notably lower (0.902), suggesting that while its classification accuracy is competitive, its predictions may be less consistent across equal inputs.

Table 4 shows that both GPT-4.1-mini and GPT-4.1 were significantly more stable than o4-mini, with mean differences of +0.068 and +0.062, respectively, and small effect sizes (Cohen's d = 0.280 and 0.250). A post-hoc power analysis (n=500) yielded high power (1.000), confirming sufficient sample size for these comparisons. Similarly, both outperformed GPT-4.1-nano in terms of stability, with statistically significant differences and negligible effect sizes (Cohen's d = −0.192 and −0.160), and power values of 0.990 and 0.945. However, the comparison between GPT-4.1-nano and o4-mini was not statistically significant, with a negligible effect size (Cohen's d = 0.093) and low power (0.548). No significant difference was found between GPT-4.1-mini and GPT-4.1 in stability, with a negligible effect size (Cohen's d = 0.034) and very low power (0.116).

Table 5
Pairwise comparisons of model performance (F1 score) with 95% confidence intervals (CIs), Cohen's d, effect size interpretation, achieved power, and BH-adjusted p-value.

Model A	Model B	Mean Δ	Cohen's d	Effect Size	Power	p
GPT-4.1-nano	GPT-4.1-mini	−0.093 [−0.102, −0.086]	−24.25	Large	1.000	<0.001***
GPT-4.1-nano	GPT-4.1	−0.092 [−0.100, −0.084]	−23.12	Large	1.000	<0.001***
GPT-4.1-nano	o4-mini	−0.093 [−0.101, −0.086]	−23.74	Large	1.000	<0.001***
GPT-4.1-mini	GPT-4.1	0.001 [−0.004, 0.006]	0.46	Small	1.000	0.668
GPT-4.1-mini	o4-mini	0.000 [−0.004, 0.005]	0.14	Negligible	1.000	0.913
GPT-4.1	o4-mini	−0.001 [−0.006, 0.003]	−0.33	Small	1.000	0.809

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6
Classification performance (F1 score) for each CoDA category with 95% confidence intervals (CIs).

Category	GPT-4.1-nano	GPT-4.1-mini	GPT-4.1	o4-mini
Arms	0.810 [0.783, 0.832]	0.957 [0.943, 0.968]	0.933 [0.917, 0.946]	0.942 [0.926, 0.953]
Crypto	0.684 [0.661, 0.708]	0.818 [0.799, 0.838]	0.836 [0.816, 0.854]	0.840 [0.820, 0.858]
Drugs	0.935 [0.924, 0.945]	0.959 [0.950, 0.966]	0.953 [0.944, 0.961]	0.956 [0.946, 0.963]
Electronic	0.276 [0.229, 0.333]	0.863 [0.834, 0.888]	0.912 [0.889, 0.930]	0.904 [0.882, 0.924]
Financial	0.753 [0.730, 0.774]	0.867 [0.851, 0.881]	0.879 [0.864, 0.893]	0.893 [0.879, 0.907]
Gambling	0.908 [0.891, 0.922]	0.985 [0.978, 0.990]	0.976 [0.968, 0.983]	0.969 [0.959, 0.977]
Hacking	0.769 [0.742, 0.792]	0.829 [0.807, 0.851]	0.828 [0.806, 0.846]	0.786 [0.762, 0.808]
Others	0.787 [0.775, 0.800]	0.855 [0.845, 0.865]	0.850 [0.840, 0.859]	0.854 [0.843, 0.863]
Porn	0.937 [0.926, 0.946]	0.943 [0.933, 0.953]	0.943 [0.932, 0.951]	0.947 [0.937, 0.956]
Violence	0.666 [0.633, 0.699]	0.758 [0.722, 0.790]	0.730 [0.697, 0.764]	0.731 [0.697, 0.764]

Regarding F1 score (Table 5), all comparisons involving GPT-4.1-nano showed large and statistically significant differences, with very large effect sizes (Cohen's d = −24.25, −23.12, −23.74), confirming that it performed substantially worse than all other models. A post-hoc power analysis (n=10,000) yielded power of 1.000 for these comparisons. In contrast, the differences between GPT-4.1-mini, GPT-4.1, and o4-mini were minimal and not statistically significant, with small to negligible effect sizes (Cohen's d = 0.462, 0.143, −0.328), and power values of 1.000. The bootstrapped comparisons for Precision and Recall yielded results consistent with those of the F1 score.

Beyond the quantitative performance metrics, several anomalies were observed during the classification tasks, particularly for the GPT-4.1-nano model. In the full dataset evaluation (10,000 documents), a total of 175 responses did not match any of the valid category labels explicitly provided in the prompt. Responses included minor variations such as "Electronics" instead of *Electronic*, "Cards" or "Carding" instead of *Financial*, "Guns" instead of *Arms*, or "Cripto" instead of *Crypto*, reflecting inconsistent adherence to the specified category vocabulary.

For the GPT-4.1-mini model, a different issue emerged: during the full dataset evaluation, four documents from the *Porn* class triggered content policy restrictions, resulting in error messages indicating the model's refusal to classify such content. Interestingly, during the stability testing (which involved repeated classifications over a sample of 500 documents), this issue was observed only once and only affected two of the ten repetitions for a single document—highlighting a lack of consistency in content moderation responses. Notably, the GPT-4.1 and o4-mini models did not return any out-of-range values or policy-related errors in any of the test scenarios.

To better visualize performance differences across models and categories, Fig. 1 displays bootstrapped confidence intervals for per-category F1 scores, excluding GPT-4.1-nano for clarity. In most categories, the three higher-performing models (GPT-4.1-mini, GPT-4.1, and o4-mini) showed relatively close performance, with several cases exhibiting overlapping intervals.

Table 6 presents the F1 scores by category for each model, with 95% confidence intervals. Categories such as *Drugs*, *Gambling*, and *Porn* were classified with high accuracy across all models, achieving F1 scores above 0.94, except for GPT-4.1-nano. In contrast, *Electronic* and *Violence* consistently showed the lowest performance, with F1 scores notably below 0.9 for most models. The GPT-4.1-nano model performed significantly worse across nearly all categories, with particularly poor results in *Electronic* (F1 = 0.276) and *Crypto* (F1 = 0.684).

Table 7 reveals significant variations in per-category F1 scores among GPT-4.1-mini, GPT-4.1, and o4-mini, adjusted for multiple comparisons using the Benjamini–Hochberg procedure at an FDR of 0.05. Notable strengths emerge for GPT-4.1-mini in *Arms*, *Gambling*, and *Hacking*, where it consistently outperforms the other models, particularly against o4-mini. Conversely, GPT-4.1-mini exhibits weaknesses in *Crypto* and *Electronic*, lagging behind both GPT-4.1 and o4-mini, with the largest disparities observed in *Electronic*. The *Financial* category highlights o4-mini's superiority over both GPT models, while *Violence* shows GPT-4.1-mini edging out o4-mini. These differences, supported by large effect sizes and perfect power (1.000), underscore category-specific performance profiles, suggesting potential influences from training data or classification strategies that warrant deeper analysis.

A language-stratified analysis revealed consistent performance across English and non-English documents, with only limited differences between the two subsets (Table 8). The largest gap was observed for GPT-4.1, which achieved a higher F1 on English documents (Δ = 0.031 [0.006, 0.056]), a statistically significant difference with a large effect size (d = 2.30, p = 0.011; Table 9). In contrast, the other models showed smaller and non-significant differences: GPT-4.1-mini favored English, o4-mini had a mild

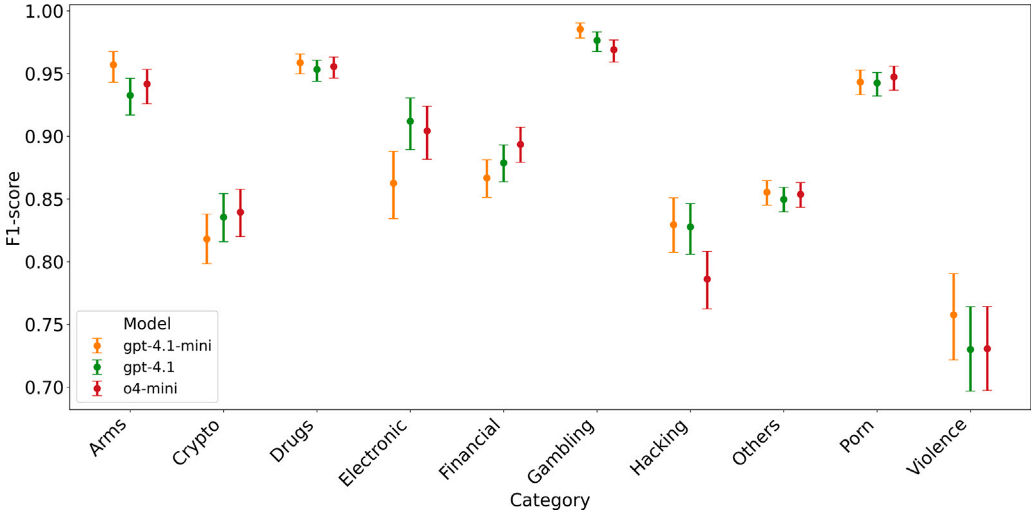

Fig. 1. Bootstrap mean F1 scores with 95% confidence intervals for each category and model, excluding GPT-4.1-nano for clarity. Each point represents the mean F1 score computed via 1000 bootstrap resamples, and error bars indicate the 95% confidence interval. Categories are sorted alphabetically along the x-axis.

Table 7
Statistically significant differences in per-category classification performance (F1 score) with 95% confidence intervals (CIs), Cohen's d, effect size interpretation and BH-Adjusted p-value. Achieved power is 1.000 for all the comparisons. GPT-4.1-nano is excluded for clarity.

Category	Comparison	Mean Δ	Cohen's d	Effect size	p
Arms	GPT-4.1-mini → GPT-4.1	0.025 [0.012, 0.037]	3.80	Large	<0.001***
Arms	GPT-4.1-mini → o4-mini	0.015 [0.004, 0.028]	2.48	Large	0.013*
Crypto	GPT-4.1-mini → GPT-4.1	-0.018 [-0.030, -0.005]	-2.70	Large	0.016*
Crypto	GPT-4.1-mini → o4-mini	-0.022 [-0.036, -0.007]	-2.96	Large	0.004**
Electronic	GPT-4.1-mini → GPT-4.1	-0.050 [-0.070, -0.032]	-5.03	Large	<0.001***
Electronic	GPT-4.1-mini → o4-mini	-0.042 [-0.065, -0.021]	-3.70	Large	<0.001***
Financial	GPT-4.1-mini → o4-mini	-0.027 [-0.039, -0.015]	-4.48	Large	<0.001***
Financial	GPT-4.1 → o4-mini	-0.015 [-0.026, -0.003]	-2.54	Large	0.013*
Gambling	GPT-4.1-mini → o4-mini	0.016 [0.011, 0.024]	4.91	Large	<0.001***
Gambling	GPT-4.1 → o4-mini	0.007 [0.002, 0.014]	2.41	Large	0.021*
Hacking	GPT-4.1-mini → o4-mini	0.043 [0.027, 0.061]	4.96	Large	<0.001***
Violence	GPT-4.1-mini → o4-mini	0.027 [0.003, 0.051]	2.24	Large	0.045*

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 8
Classification performance (F1 score) for English and non-English documents with 95% confidence intervals (CIs).

Model	English F1	Non-English F1
GPT-4.1-nano	0.791 [0.783, 0.800]	0.826 [0.803, 0.846]
GPT-4.1-mini	0.890 [0.884, 0.897]	0.864 [0.844, 0.882]
GPT-4.1	0.891 [0.885, 0.897]	0.852 [0.830, 0.869]
o4-mini	0.890 [0.883, 0.896]	0.868 [0.846, 0.885]

advantage for English, and GPT-4.1-nano performed slightly better on non-English texts, though none of these reached statistical significance ($p > 0.05$). Overall, these results suggest that the evaluated GPT models exhibit broadly robust multilingual performance, with only one model showing a systematic bias toward English content.

To further investigate model behavior, we examined the confusion matrix of GPT-4.1-mini (Fig. 2) and complemented this analysis with a qualitative review of a random 20% sample of misclassified documents. The confusion matrix was generated through bootstrap sampling over predictions on the full dataset. Values represent the mean percentage of documents in each cell along with 95% confidence intervals. The model demonstrated strong performance in categories such as *Arms* (96.5%), *Gambling* (98.9%), and *Drugs* (95.7%), but also revealed systematic misclassification patterns that reflect recurring sources of confusion.

The most frequent misclassification involved documents incorrectly labeled as *Others*, despite belonging to more specific categories. Notably, the model misclassified 27.9% of documents originally labeled as *Violence* into *Others*, along with cases from

Table 9
Classification performance differences (F1 score) between English and non-English language with 95% confidence intervals (CIs), Cohen's d, effect size interpretation and BH-Adjusted p-value.

Model	Δ F1	Cohen's d	Effect Size	Power	p
GPT-4.1-nano	−0.024 [−0.052, 0.005]	−1.59	Large	1.000	0.162
GPT-4.1-mini	0.009 [−0.018, 0.036]	0.69	Medium	1.000	0.531
GPT-4.1	0.031 [0.006, 0.056]	2.30	Large	1.000	0.011*
o4-mini	0.020 [−0.006, 0.045]	1.49	Large	1.000	0.170

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Porn (6.6%), *Hacking* (3.6%), *Financial* (4.4%), and *Drugs* (3.6%). Based on the qualitative review, one possible explanation is the model's conservative handling of borderline cases—particularly when documents lack explicit keywords or direct references to illicit services.

For instance, many *Violence* documents misclassified as *Others* involved forum discussions about politically charged events, such as shootings or historical violence. The GPT-4.1-mini model often interpreted these as general political discourse rather than criminal content, likely due to a strict adherence to the *Violence* category's definition, which emphasizes explicit references to trafficking, blackmail, or torture. Additionally, the qualitative review identified cases of fictional violence — such as depictions of violent acts in comics or video games — labeled as *Violence* by human annotators but classified as *Others* by GPT models. These reasonable discrepancies highlight how differing interpretations of category boundaries contribute to systematic divergences between model predictions and human annotations.

Conversely, the model also reassigned numerous *Others* documents into specific classes, most frequently *Crypto* (5.3%), *Hacking* (3.6%), and *Financial* (2.4%). Qualitative insights suggest that some prevalent terms — such as references to cryptocurrency wallets, malware, or payment services — which are especially frequent in Dark Web documents, may have triggered confident predictions for a specific illicit activity, rendering GPT models as overly sensitive to domain-specific jargon which is not so common in surface web and in other types of documents fed to the models during their training.

Outside the *Others* class, another recurring confusion occurred between *Electronic* and *Financial*: 16.9% of *Electronic* documents were misclassified as *Financial*. This appears to be partially explained by listings that combine electronic devices with services clearly intended for financial fraud in Dark Web marketplaces. The model likely inferred the *Financial* label based on the co-occurrence of banking terms, cryptocurrencies, or carding-related jargon. Although this explanation is plausible for many of the reviewed documents, other factors may also contribute to this confusion.

Less frequent but illustrative errors include *Violence*-to-*Arms* (4.3%), *Crypto*-to-*Hacking* (3.3%), *Crypto*-to-*Financial* (1.8%), and *Porn*-to-*Violence* (1.7%). As suggested by the earlier qualitative review, these misclassifications often reflect overlapping thematic content. For instance, *Violence*-to-*Arms* cases involved violent scenarios where the mention of weapons may have dominated the interpretation. *Crypto*-to-*Hacking* errors typically combined references to hacking tools and cryptocurrency, while *Crypto*-to-*Financial* confusion likely stemmed from the co-occurrence of crypto-related and financial language. In *Porn*-to-*Violence* cases, explicit content with violent undertones may have shifted the model's focus toward violence. These examples illustrate how single-label classification can be limiting when documents blend elements from multiple categories.

In sum, the qualitative review provided additional context to interpret confusion matrix patterns and revealed plausible sources of systematic error. However, it is important to emphasize that these explanations are hypotheses based on sampled cases, not definitive accounts. Other factors — such as annotation inconsistencies, implicit cultural knowledge, or token-level attention behaviors — may also contribute to these misclassifications and warrant further investigation.

5. Discussion

Supervised models, such as fine-tuned BERT, RoBERTa, and DarkBERT, have set the benchmark on the CoDA dataset, achieving F1 scores above 0.920 through domain adaptation and hybrid architectures like TextCNN with topic modeling (Chen et al., 2024; Jin et al., 2023; Shin et al., 2024). These models, however, rely on large, curated labeled datasets and have been evaluated only on the English-language subset of CoDA. In contrast, our study evaluates four GPT-based models (GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, and o4-mini) for zero-shot classification on the full CoDA dataset, including multilingual texts, achieving weighted F1 scores of approximately 0.885 for the top-performing models. This performance surpasses previous zero-shot evaluations on CoDA, such as GPT-3.5-turbo (0.805) (Prado-Sánchez et al., 2024) and LLaMA-3 (0.748) (Chen et al., 2024). The similar F1 scores of GPT-4.1-mini, GPT-4.1, and o4-mini, despite their differences in size and configuration, suggest that performance may be constrained by factors such as prompt design or category definitions.

Table 10 summarizes the F1 scores of prior CoDA studies alongside the present zero-shot GPT models. While supervised models still outperform zero-shot approaches by 6–8 percentage points, it is important to note that prior supervised evaluations were limited to English-only subsets. Our work is, to our knowledge, the first to process the full multilingual CoDA dataset, demonstrating that GPT-based models achieve competitive performance across diverse languages—a practical advantage given the multilingual nature of the Dark Web (Ebrahimi et al., 2022).

The language-stratified evaluation confirms that GPT-based models maintain broadly consistent performance across English and non-English documents. Importantly, differences were small in magnitude and rarely reached statistical significance. Only GPT-4.1

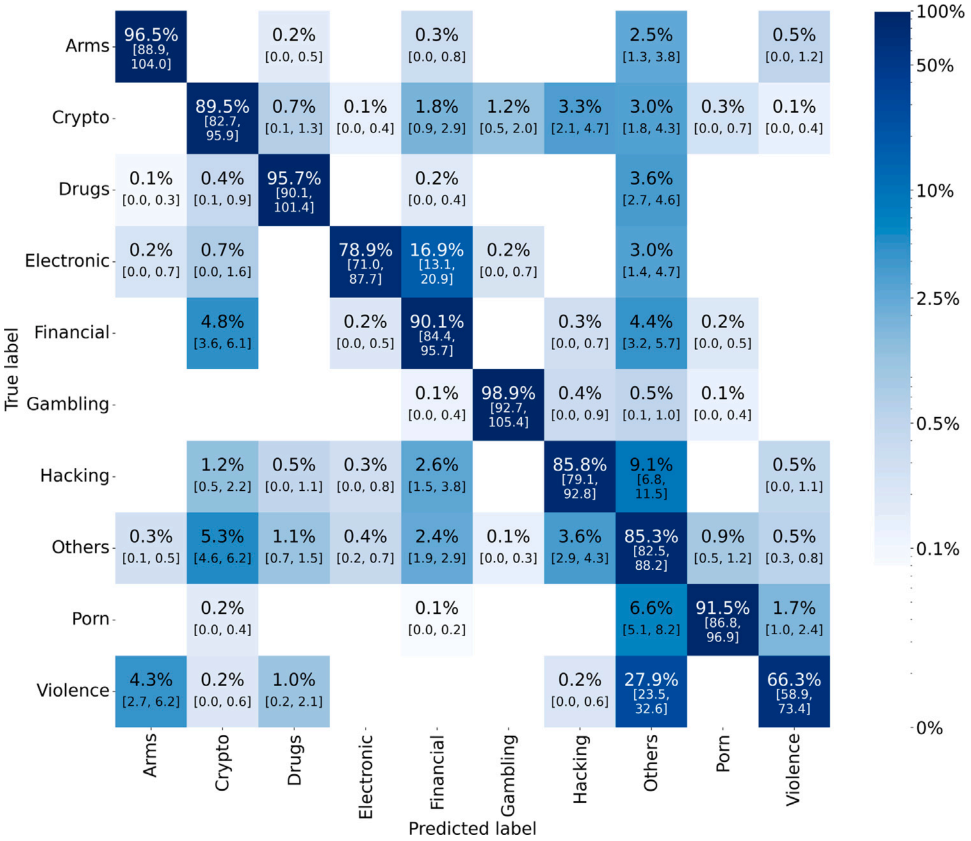

Fig. 2. Normalized Confusion Matrix for GPT-4.1-mini on the CoDA Dataset. Each row corresponds to the true category of documents, and each column corresponds to the predicted category. Cell values show the mean percentage of documents of a true category classified into each predicted category (row-normalized), with 95% bootstrap confidence intervals indicated below the mean. Colors represent the proportion of documents on a logarithmic scale, where darker shades indicate higher classification percentages. Values of 0% are shown as blank cells for readability.

Table 10
Classification performance (F1 score) on the CoDA dataset, comparing the results of the present study with prior studies, including model details, dataset, and experimental setup.

Study	Model	Setup	Dataset	F1
Chen et al. (2024)	LLaMA-3	Zero-shot	CoDA (English only)	0.748
Prado-Sánchez et al. (2024)	GPT-3.5-turbo	Zero-shot	CoDA (Porn excluded)	0.805
This work	GPT-4.1-mini, GPT-4.1, o4-mini	Zero-shot	CoDA (full dataset)	≈ 0.886
Jin et al. (2023)	DarkBERT	Supervised	CoDA (English only)	0.945
Chen et al. (2024)	BERT, RoBERTa, DarkBERT, fine-tuned LLaMa-3	Supervised	CoDA (English only)	≈ 0.945
Shin et al. (2024)	TextCNN + topic modeling	Supervised	CoDA (English only)	0.962

showed a statistically significant gap, achieving higher F1 scores on English texts ($\Delta = 0.031$ [0.006, 0.056], $p = 0.011$), suggesting a systematic bias toward English content. By contrast, GPT-4.1-mini and o4-mini displayed mild, non-significant advantages for English, while GPT-4.1-nano performed slightly better on non-English texts, though again without statistical significance ($p > 0.05$). These findings indicate that, overall, zero-shot GPT models demonstrate robust multilingual generalization, with limited evidence of consistent or systematic performance degradation in non-English languages. This result aligns with prior research showing that LLMs outperform fine-tuned supervised models in cross-lingual contexts thanks to their strong generalization abilities (Chen et al., 2024; Zhao et al., 2023).

Beyond classification performance, output stability is a critical metric for evaluating large language models (LLMs). We measured the Total Agreement Rate across 10 runs (*TARa@10*) and found a high value of 0.970 for GPT-4.1-mini, indicating strong reproducibility under deterministic prompts with zero temperature. Recent research highlights that output variability, even in deterministic settings, can undermine the reliability of LLM-based systems (Atil et al., 2025), and single-output evaluations may fail to meet scientific reliability standards without explicit stability checks (Reiss, 2023). Robust prompt design is essential for consistent

LLM performance (Barrie et al., 2025). However, the cost-effective GPT-4.1-nano model exhibited stability issues, producing responses outside the required category labels (e.g., *Electronics* instead of *Electronic*), resulting in a lower *TARa@10* of 0.928 and an F1 score of 0.793. Similarly, o4-mini, despite matching top performers in F1 score, showed reduced stability (*TARa@10* of 0.902) due to limited configurability, such as the absence of temperature parameter, impacting its reliability in operational settings. These stability challenges highlight the importance of model selection and prompt optimization, particularly when considering the factors contributing to the performance gap between zero-shot and supervised models.

Qualitative analysis of misclassified documents in the CoDA dataset reveals that many discrepancies between zero-shot model predictions and human-assigned labels stem from ambiguities in classification protocols, particularly in categories like *Violence* and *Electronic*. For example, documents discussing political events or fictional violence lead to differing interpretations of *Violence*, while those combining illegal electronics and financial fraud expose vague definitions in the *Electronic* category. Similar challenges are observed in other domains: Kotzé and Senekal (2024) reported that GPT-3 diverged from human annotators when classifying WhatsApp messages about past violence due to missing contextual guidelines, and Li et al. (2024) noted that GPT-3.5-turbo disagreed with annotators on hateful content due to unclear label boundaries, with low Krippendorff's alpha values (0.4–0.53) indicating unreliable annotation schemes (Krippendorff, 2004). These findings suggest that the apparent underperformance of zero-shot models often reflects limitations in classification protocols rather than model deficiencies, emphasizing the need for refined taxonomies to improve model-human alignment.

In summary, our study highlights the trade-offs between supervised and zero-shot models for Dark Web content classification on the CoDA dataset. Supervised models like DarkBERT and RoBERTa achieve superior F1 scores above 0.920 but are limited to English data and require extensive labeled datasets. In contrast, zero-shot models such as GPT-4.1-mini, GPT-4.1, and o4-mini offer competitive performance with F1 scores of approximately 0.885, excelling in multilingual settings by processing the full CoDA dataset, including 1145 non-English documents. Their effectiveness in low-resource and diverse linguistic contexts underscores their practical value for real-world applications. However, their performance may be partially constrained by ambiguities in the classification protocol, particularly in categories like *Violence* and *Electronic*, where overlapping or vague definitions often lead to misclassifications. Additionally, challenges in output stability, particularly with models like GPT-4.1-nano and o4-mini, reveal areas for improvement. Refining prompt design and developing more precise taxonomies could bridge the performance gap, enabling zero-shot models to better address the complex and multilingual nature of Dark Web content classification.

5.1. Limitations of the study

This study has several limitations that should be considered when interpreting the results.

First, the evaluation is based exclusively on the CoDA dataset, which — although carefully balanced and widely adopted — does not fully capture the diversity and messiness of real-world Dark Web content. As a result, the evaluation may overestimate performance relative to more heterogeneous or noisy datasets found in operational settings.

Second, the experiments are limited to four GPT-based models (GPT-4.1-nano, GPT-4.1-mini, GPT-4.1, and o4-mini), all evaluated in a zero-shot configuration with a single prompt structure. This focus was chosen to assess the performance of recent GPT models on the CoDA dataset, but it limits the scope to a specific subset of large language models within the GPT ecosystem. Consequently, the results may not generalize to other zero-shot approaches, nor to other LLM families like Claude, Gemini, or LLaMA, which may exhibit different performance due to variations in pretraining corpora, alignment strategies, or moderation policies. Expanding the evaluation to include these approaches was beyond the scope of this study but is a critical direction for future research to understand the broader landscape of zero-shot classification for Dark Web content.

Third, although the study includes detailed performance metrics and a qualitative review of misclassified documents, it does not offer fine-grained interpretability or insight into model reasoning. Zero-shot prompting does not expose intermediate representations or token-level attributions, limiting our ability to understand why the model made a particular decision. Moreover, the study did not explore alternative prompting strategies — such as few-shot prompting, chain-of-thought reasoning, or task decomposition — which may further affect performance and stability.

Finally, while numerous classification discrepancies between model predictions and ground-truth labels were manually reviewed, the study does not include a systematic process to determine whether these discrepancies were due to model errors, annotation inconsistencies, or genuine semantic ambiguities in the data. Prior work has shown that test set label noise is pervasive in NLP benchmarks and can distort performance assessments (Northcutt et al., 2021), and that subtle flaws in annotation guidelines can introduce systematic bias (Parmar et al., 2024). Without adjudication or re-annotation, it is difficult to determine the true reliability of the ground truth, which in turn limits the interpretability of false positives and false negatives in model outputs.

6. Conclusions

This study presents the most comprehensive evaluation to date of recent GPT models for zero-shot classification of Dark Web content, including multilingual texts. Our results show that models such as GPT-4.1-mini, GPT-4.1, and o4-mini achieve weighted F1 scores close to 0.885, representing a substantial improvement over previous zero-shot baselines such as GPT-3.5-turbo (Prado-Sánchez et al., 2024) and LLaMA-3 (Chen et al., 2024). Although these models do not outperform supervised approaches like DarkBERT (Jin et al., 2023) or hybrid neural architectures such as TextCNN with topic modeling (Shin et al., 2024) on English-language documents, they offer considerable practical advantages, particularly their ability to generalize to multilingual content without domain-specific training.

Beyond performance, we evaluated the stability of model predictions using the *TARa@10* metric. Our results indicate that models like GPT-4.1-mini achieve high stability (0.970), making them suitable for operational scenarios (Atil et al., 2025; Reiss, 2023). In contrast, lighter or less controllable variants such as GPT-4.1-nano and o4-mini showed notable limitations in consistency across repeated inferences.

Our findings also reveal systematic difficulties in classifying certain categories, such as *Violence* and *Electronic*, stemming from semantic ambiguity and inconsistencies in human annotation. Particularly, we observed a strong tendency of GPT models to assign ambiguous documents to the catch-all class *Others*, even when weak signals of specific illicit activities were present. This conservative strategy helps models avoid incorrect specific labels but reduces their precision in borderline cases.

In conclusion, this evaluation establishes a performance baseline for GPT-based models in zero-shot classification of Dark Web content, highlighting their effectiveness in multilingual settings and the importance of prompt design and clear category definitions. Future work should explore refined prompting strategies, explanatory reasoning mechanisms, and improvements in annotation guidelines to enhance model-human alignment in complex or ambiguous classification tasks.

CRediT authorship contribution statement

Adrián Domínguez-Díaz: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Luis de-Marcos:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Funding acquisition. **Víctor-Pablo Prado-Sánchez:** Writing – original draft, Software, Data curation. **Daniel Rodriguez:** Writing – review & editing, Funding acquisition. **José-Javier Martínez-Herráiz:** Resources, Project administration, Funding acquisition.

Funding

This publication is part of the project "Project for the analysis and recovery of criminal evidence associated with hidden networks" (PARCHE), with reference PID2021-125645OB-I00, funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Prompt template

The following is the exact prompt used for all models during zero-shot classification:

> You are given the text content of a darknet website. Your task is to classify this content into one of the predefined categories listed below.
>
> Each class includes a name and a description. Read the text carefully and select the single most appropriate class that best matches the main topic of the content. If the content does not clearly belong to any specific category, assign it to "Others".
>
> Categories:
>
> - Arms: Any type of non-lethal or lethal weapons, including guns, ammunition, explosives, knives, missiles, or chemical weapons.
>
> - Crypto: Services or technologies related to cryptocurrency, such as wallets, mining, laundering, mixing, multiplying, scamming, and escrow.
>
> - Drugs: Legal or illegal drugs, including medications, steroids, painkillers, viagra, cannabis, hashish, meth, benzodiazepines, ecstasy, opioids, and psychedelics.
>
> - Electronic: Sale of or information about stolen or hacked electronic devices such as mobile phones, laptops, tablets, etc.
>
> - Financial: Counterfeit, cloned, or stolen money or identifications (e.g., credit cards, bills, certificates, passports), fiat currency transfers (e.g., PayPal), ATM skimmers, magnetic card readers, etc.
>
> - Gambling: Gambling, betting, casinos, lotteries, or any related services.
>
> - Hacking: Hacking tools, guides, groups, or services, including ransomware, malware, exploits, DDoS attacks, cracking, or botnets.
>
> - Porn: Pornographic or sexually explicit content, including child pornography.
>
> - Violence: Content related to human trafficking, hitmen, kidnapping, poisoning, torture, extortion, sextortion, sex slavery, or blackmail.

- Others: Any content that does not clearly belong to the categories above, such as login pages, error messages, or unrelated information.

The darknet website content is provided below, enclosed in triple quotes.

""" {content} """

Respond only with a JSON object with ‘‘category’’ as its single key and the name of the selected category as its value.

Data availability

Data will be made available on request.

References

Al-Nabki, M. W., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2019). ToRank: Identifying the most influential suspicious domains in the Tor network. *Expert Systems with Applications*, *123*, 212–226.

Al-Nabki, M. W., Fidalgo, E., Alegre, E., & de Paz, I. (2017). Classifying illegal activities on tor network based on web textual contents. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 1, long papers* (pp. 35–43). Valencia, Spain: Association for Computational Linguistics.

Alaidi, A. H. M., Alairaji, R. M., Alrikabi, H. T. S., Aljazaery, I. A., & Abbood, S. H. (2022). Dark web illegal activities crawling and classifying using data mining techniques. *International Journal of Interactive Mobile Technologies (IJIM)*, *16*(10), 122–139.

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., & Baldwin, B. (2025). Non-determinism of ‘‘deterministic’’ LLM settings. arXiv:2408.04667, [cs].

Avarikioti, G., Brunner, R., Kiayias, A., Wattenhofer, R., & Zindros, D. (2018). Structure and content of the visible darknet. arXiv:1811.01348, [cs].

Barrie, C., Palaiologou, E., & Törnberg, P. (2025). Prompt stability scoring for text annotation with large language models. arXiv:2407.02039, [cs].

Bauer, N., Preisig, M., & Volk, M. (2024). Offensiveness, hate, emotion and GPT: Benchmarking GPT3.5 and GPT4 as classifiers on Twitter-specific datasets. In R. Kumar, A. K. Ojha, S. Malmasi, B. R. Chakravarthi, B. Lahiri, S. Singh, & S. Ratan (Eds.), *Proceedings of the fourth workshop on threat, aggression & cyberbullying @ LREC-COLING-2024* (pp. 126–133). Torino, Italia: ELRA and ICCL.

Belal, M., She, J., & Wong, S. (2023). Leveraging ChatGPT as text annotation tool for sentiment analysis. arXiv:2306.17177, [cs].

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *57*(1), 289–300.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Vol. 33*, In *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates, Inc..

Cascavilla, G., Catolino, G., & Sangiovanni, M. (2022). Illicit darkweb classification via natural-language processing: Classifying illicit content of webpages based on textual information. In *Proceedings of the 19th international conference on security and cryptography* (pp. 620–626). arXiv:2312.04944, [cs].

Chae, Y., & Davidson, T. (2025). Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociological Methods & Research*, 00491241251325243.

Chen, H., Diao, Y., Xiang, H., Huo, Y., Xie, X., Zhao, J., Wang, X., Sun, Y., & Shi, J. (2024). Decode the dark side of the language: Applications of LLMs in the dark web. In *2024 IEEE 9th international conference on data science in cyberspace (DSC)* (pp. 224–231). http://dx.doi.org/10.1109/DSC63484.2024.00037.

Cilleruelo, C., de Marcos, L., Junquera-Sánchez, J., & Martínez-Herráiz, J.-J. (2021). Interconnection between darknets. *IEEE Internet Computing*, *25*(3), 61–70.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davison, A. C., & Hinkley, D. V. (1997). Bootstrap methods and their application. In *Cambridge series in statistical and probabilistic mathematics*, Cambridge: Cambridge University Press.

Domínguez-Díaz, A., Goyanes, M., De-Marcos, L., & Prado-Sánchez, V. P. (2024). Comparative analysis of automatic gender detection from names: evaluating the stability and performance of ChatGPT versus Namsor, and Gender-API. *PeerJ Computer Science*, *10*, Article e2378.

Ebrahimi, M., Chai, Y., Samtani, S., & Chen, H. (2022). Cross-lingual cybersecurity analytics in the international dark web with adversarial deep representation learning. *MIS Quarterly*, *46*(2), 1209–1226.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185, Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1987.10478410.

Efron, B., & Tibshirani, R. (1993). An introduction to the bootstrap. In *Monographs on statistics and applied probability*, (57), New York: Chapman & Hall.

Elma, K. J., Vishal, S., & Varun, M. (2024). Unveiling the dark: Analyzing and categorizing dark web activities using bi-directional LSTMs. In *2024 2nd international conference on networking and communications (ICNWC)* (pp. 1–6). Conference Name: 2024 2nd International Conference on Networking and Communications (ICNWC) ISBN: 9798350365269 Place: Chennai, India Publisher: IEEE.

Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2024). A bibliometric review of large language models research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, *15*(5), 91:1–91:25.

Goyanes, M., De-Marcos, L., & Domínguez-Díaz, A. (2024). Automatic gender detection: a methodological procedure and recommendations to computationally infer the gender from names with ChatGPT and gender APIs. *Scientometrics*, *129*(11), 6867–6888.

Gupta, S. (2022). Hate speech detection using OpenAI and GPT-3. *International Journal of Emerging Technology and Advanced Engineering*, *12*(5), 132–138.

He, S., He, Y., & Li, M. (2019). Classification of illegal activities on the dark web. In *Proceedings of the 2nd international conference on information science and systems* (pp. 73–78). New York, NY, USA: Association for Computing Machinery.

Jin, Y., Jang, E., Cui, J., Chung, J.-W., Lee, Y., & Shin, S. (2023). DarkBERT: A language model for the dark side of the internet. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 7515–7533). Toronto, Canada: Association for Computational Linguistics.

Jin, Y., Jang, E., Lee, Y., Shin, S., & Chung, J.-W. (2022). Shedding new light on the language of the dark web. In *Proceedings of the 2022 conference of the north American chapter of the association for computational linguistics: human language technologies* (pp. 5621–5637). Seattle, United States: Association for Computational Linguistics.

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, *6*, Article 100048.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radlinski, L., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, *99*, Article 101861.

Kotzé, E., & Senekal, B. A. (2024). Evaluating the GPT-3.5 and GPT-4 large language models for zero-shot classification of south african violent event data. In *2024 international conference on artificial intelligence, big data, computing and data communication systems (icABCD)* (pp. 1–7). http://dx.doi.org/10.1109/icABCD62167.2024.10645261.

Krippendorff, K. (2004). *Content analysis: an introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Li, L., Fan, L., Atreja, S., & Hemphill, L. (2024). "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, *18*(2), 30:1–30:36.

Liu, Y. L., Blodgett, S. L., Cheung, J., Liao, Q. V., Olteanu, A., & Xiao, Z. (2024). ECBD: Evidence-centered benchmark design for NLP. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 16349–16365). Bangkok, Thailand: Association for Computational Linguistics.

Loukas, L., Stogiannidis, I., Malakasiotis, P., & Vassos, S. (2023). Breaking the bank with ChatGPT: Few-shot text classification for finance. In C.-C. Chen, H. Takamura, P. Mathur, R. Sawhney, H.-H. Huang, & H.-H. Chen (Eds.), *Proceedings of the fifth workshop on financial technology and natural language processing and the second multimodal AI for financial forecasting* (pp. 74–80). Macao.

Mathebula, M., Modupe, A., & Marivate, V. (2024). ChatGPT as a text annotation tool to evaluate sentiment analysis on South African financial institutions. *IEEE Access*, *12*, 144017–144043.

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv:2103.14749, [stat].

Padiu, B., Iacob, R., Rebedea, T., & Dascalu, M. (2024). To what extent have LLMs reshaped the legal domain so far? A scoping literature review. *Information*, *15*(11), 662.

Parmar, M., Mishra, S., Geva, M., & Baral, C. (2024). Don't blame the annotator: Bias already starts in the annotation instructions. arXiv:2205.00415, [cs].

Prado-Sánchez, V.-P., Domínguez-Díaz, A., de Marcos, L., & Martínez-Herráiz, J.-J. (2024). Clasificación zero-shot de contenidos de la Dark Web mediante GPT-3.5: Evaluación de rendimiento y análisis de errores del clasificador. In *IX jornadas nacionales de investigación en ciberseguridad, 2024, ISBN 978-84-09-62140-8, págs. 286-292* (pp. 286–292). Antonia M. Reina Quintero, Section: IX Jornadas Nacionales de Investigación En Ciberseguridad.

Reiss, M. V. (2023). Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. arXiv:2304.11085, [cs].

Sennad, M., Ellaky, Z., & Benabbou, F. (2025). Enhancing DarkWebActivities classification using embedding methods. In *2025 5th international conference on innovative research in applied science, engineering and technology (IRASET)* (pp. 1–8). http://dx.doi.org/10.1109/IRASET64571.2025.11008275.

Shin, G.-Y., Jang, Y., Kim, D.-W., Park, S., Park, A.-R., Kim, Y., & Han, M.-M. (2024). Dark side of the web: Dark web classification based on TextCNN and topic modeling weight. *IEEE Access*, *12*, 36361–36371.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.

Tariq, A., Luo, M., Urooj, A., Das, A., Jeong, J., Trivedi, S., Patel, B., & Banerjee, I. (2024). Domain-specific LLM development and evaluation – a case-study for prostate cancer. Pages: 2024.03.15.24304362.

Zhao, B., Jin, W., Del Ser, J., & Yang, G. (2023). ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification. *Neurocomputing*, *557*, Article 126708.