

Explainable Retrieval-Augmented Generation for Multilingual Dark-Pattern Auditing and Digital Consumer Protection

Jobin Roy¹, Akshita Sakshi¹, Shifin Mohammed¹, Het Darshan Mehta^{1,2}, Marco Polignano^{3,1}, Giovanni Semeraro³ and Ernesto William De Luca^{1,2}

¹Otto-von-Guericke University, Universitatplatz 2, 39106 Magdeburg, Germany

²Leibniz Institute for Educational Media | Georg Eckert Institute, Freisestr. 1, 38118 Braunschweig, Germany

³Università degli studi di Bari Aldo Moro, Piazza Umberto I, 70121 Bari, Italy

Abstract

Dark patterns are deceptive textual cues that steer consumers away from their own interests, yet most automated detection systems still treat the problem as plain classification and return only a label. This is insufficient for regulatory auditing and remediation, where stakeholders need to know not only whether a text is manipulative, but also which psychological mechanism it exploits, which evidence supports the decision, what harm it targets, and how the text could be rewritten to remove the manipulation. We present a retrieval-augmented generation (RAG) pipeline that combines dense example retrieval with a locally hosted instruction-following large language model to detect, explain, and rewrite dark-pattern text in a single structured inference step. Our system produces a structured JSON output containing the predicted label, confidence score, psychological mechanism, verbatim evidence span, harm dimension, rationale, and a consumer-safe rewrite. On the EC-DarkPattern corpus, augmented with targeted synthetic examples for rare classes, the deployed SBERT-KNN configuration achieves a macro-F1 of 0.875 and Cohen's κ of 0.905, substantially outperforming both zero-shot prompting and a fine-tuned BERT baseline while simultaneously providing structured explanations that conventional classifiers cannot generate. The same pipeline transfers to the Mathur et al. corpus without retraining, reaching a macro-F1 of 0.922 on the shared categories. We further introduce manually annotated German and Italian test sets and, to our knowledge, present the first cross-lingual evaluation of a retrieval-augmented dark-pattern detection pipeline. Using a multilingual retrieval encoder, the system achieves macro-F1 scores of 0.845 on German and 0.827 on Italian, preserving most of the English performance. Human rubric-based evaluation and LLM-as-a-judge assessment indicate that the generated explanations are generally coherent and useful, although the observed inter-annotator agreement suggests that these results should be interpreted as an initial feasibility study rather than a definitive evaluation of explanation quality. Overall, the results demonstrate that retrieval-augmented generation can extend dark-pattern detection beyond label prediction towards grounded, human-readable, and multilingual compliance support.

Keywords

dark patterns, retrieval-augmented generation, e-commerce, deceptive design, text classification, explainability, few-shot prompting, large language models, cross-lingual transfer, multilingual NLP

1. Introduction

E-commerce platforms increasingly employ textual and interface cues, commonly discussed as dark patterns, that can steer consumers away from their own interests [1, 2]. Examples include artificially urgent messages (URGENCY), misleading pricing disclosures (SNEAKING), and coercive subscription or interaction requirements (FORCED ACTION). Such practices have received increasing regulatory attention. In particular, Article 25 of the EU Digital Services Act prohibits online platforms from designing or operating interfaces in ways that deceive, manipulate, or otherwise materially distort users' ability to

XAI.it 2026 - 6th Workshop on Explainable AI, Generative and Agentic Systems: Trust, Transparency, and Human Oversight, co-located with AxiA 2026, October 6-9, 2026, Perugia, Italy

✉ jobin.roy@st.ovgu.de (J. Roy); akshita.sakshi@st.ovgu.de (A. Sakshi); shifin.mohammed@st.ovgu.de (S. Mohammed); het.mehta@ovgu.de (H. D. Mehta); marco.polignano@uniba.it (M. Polignano); giovanni.semeraro@uniba.it (G. Semeraro); deluca@ovgu.de (E. W. De Luca)

🆔 0009-0007-1741-7346 (J. Roy); 0009-0001-3771-6227 (A. Sakshi); 0009-0000-8372-3877 (H. D. Mehta); 0000-0002-3939-0136 (M. Polignano); 0000-0001-6883-1853 (G. Semeraro); 0000-0003-3621-4118 (E. W. De Luca)



© 2026 Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

make free and informed decisions [3]. This regulatory context motivates scalable tools for identifying potentially manipulative content, although the regulation does not prescribe automated detection as a compliance requirement.

Existing work commonly formulates textual dark-pattern detection as a supervised classification problem [4]. Such systems can identify a category but generally do not provide the additional information needed to inspect a prediction, such as the specific text span supporting it, the associated psychological mechanism, the potential harm, or a possible alternative formulation. These outputs are particularly relevant when automated detection is used as decision support for human auditors rather than as an autonomous compliance determination. In our experiments, zero-shot prompting reaches a macro F1 of 0.596 on the eight-class English evaluation, with particularly weak performance on several less frequent rhetorical categories.

We investigate whether Retrieval-Augmented Generation (RAG) can support this broader task. Our pipeline combines dense semantic retrieval over a labelled example index with a locally hosted instruction-following large language model. For each query, the system retrieves relevant labelled examples as few-shot demonstrations and generates a structured JSON object containing seven fields: label, confidence, psychological mechanism, verbatim evidence span, harm dimension, rationale, and a suggested rewrite. The evidence span provides an inspectable connection between the generated output and the input text, while the additional fields are intended to support subsequent human interpretation and review. We further examine cross-lingual feasibility using manually annotated German and Italian e-commerce test sets.

Our central hypothesis is that retrieving semantically relevant in-domain demonstrations improves classification by providing local examples of otherwise difficult category distinctions. Retrieval may also provide useful context for structured explanation generation; however, evidence-span extraction alone does not establish explanation faithfulness. We therefore distinguish classification performance and verbatim evidence grounding from the perceived quality of generated explanations, which we examine separately through a small human evaluation.

Our contributions are threefold. First, we present a RAG pipeline for dark-pattern detection and structured decision support without gradient-based fine-tuning, extending label prediction with a verbatim evidence span, psychological mechanism, harm dimension, rationale, and suggested rewrite [4]. Second, we compare eleven prompting and retrieval configurations spanning zero-shot and random few-shot prompting, multiple retrieval strategies, three English encoders, and k -sweep variants. On the augmented EC-DarkPattern test set, the deployed SBERT-KNN configuration reaches a macro F1 of 0.875, compared with 0.596 for zero-shot prompting; the effect of synthetic test instances is examined separately in our evaluation. Third, we introduce manually annotated German and Italian test sets to study the cross-lingual feasibility of the approach. These multilingual experiments are intended as feasibility studies rather than definitive benchmarks because of their limited sample sizes, particularly the single-annotator Italian evaluation.

2. Related Work

Dark patterns and automated detection. Dark patterns were initially catalogued as recurring interface-design strategies that steer users toward decisions they might not otherwise make, and were subsequently systematized in HCI research into broader taxonomies of manipulative design practices [5, 2]. Mathur et al. [1] provided large-scale empirical evidence of their prevalence by analysing approximately 11,000 shopping websites, establishing dark patterns as a widespread phenomenon in online commerce. Subsequent work has also examined how deceptive interfaces affect users’ awareness and behaviour. Bongard-Blanchy et al. [6], for example, found that recognising a manipulative interface does not necessarily prevent users from being influenced by it.

Automated approaches have primarily formulated dark-pattern detection as a supervised classification problem. Yada et al. [4] introduced the EC-DarkPattern corpus of e-commerce text snippets and evaluated BERT-based classifiers over eight dark-pattern categories. Related transformer-based work

has similarly demonstrated the effectiveness of pretrained language models for text classification [7]. These approaches establish that textual dark-pattern cues can be detected computationally, but their principal output is a class prediction. Consequently, additional information that may support human inspection, such as the specific input span associated with a prediction or a natural-language rationale, is generally outside the scope of the classification task.

Retrieval-augmented generation and in-context learning. Retrieval-augmented generation combines parametric language models with information retrieved from an external collection, allowing generation to be conditioned on information selected for the current input [8]. Although originally developed for knowledge-intensive NLP tasks, the same principle is relevant to in-context classification, where retrieved labelled examples can serve as demonstrations for a language model.

The choice of demonstrations can substantially affect in-context learning performance. Nearest-neighbour selection has been shown to provide more informative demonstrations than random sampling when semantic similarity is aligned with the target task [9, 10]. Alternative retrieval strategies aim to modify this relevance–diversity trade-off. Maximal Marginal Relevance (MMR), for example, explicitly balances similarity to the query against redundancy among selected examples [11]. Prototype selection instead represents categories through central examples or class-level representations, while hypothetical-document retrieval methods such as HyDE construct an intermediate representation before retrieval [12]. These strategies provide different ways of selecting demonstrations when class distributions are imbalanced or nearest neighbours are highly redundant.

Explanation and evidence grounding. Natural-language explanations and rationales provide a richer output than class labels, but their evaluation introduces an important distinction between plausibility and faithfulness. An explanation can appear coherent to a human reader without necessarily representing the actual basis of a model’s prediction. In retrieval-grounded generation, requiring an explicit evidence span offers a more directly inspectable connection between the generated output and the source text, but verbatim overlap alone should be interpreted as evidence grounding or schema compliance rather than as proof of explanation faithfulness.

This distinction is particularly relevant to dark-pattern auditing, where an automated prediction may ultimately be reviewed by a human. Prior work examining deceptive interfaces has largely concentrated on taxonomy construction, prevalence, behavioural effects, or detection [2, 1, 4]. Automated generation of structured outputs that combine a prediction with supporting textual evidence and an explanatory rationale remains less explored in this domain.

Cross-lingual retrieval and low-resource evaluation. Most available dark-pattern resources are English-language datasets, whereas deceptive commercial practices occur across languages and jurisdictions. Multilingual sentence representations provide one route to transferring retrieval systems without training a separate encoder for every target language. Sentence-BERT introduced efficient semantic sentence representations for similarity-based retrieval [13], while multilingual extensions align sentence representations across many languages through knowledge distillation [14]. Prior work on cross-lingual in-context learning further suggests that demonstrations can transfer across languages, although performance depends on the correspondence between the target-language input and the available demonstrations [15].

Low-resource categories introduce a related challenge. When labelled examples are sparse, both supervised classifiers and demonstration-based methods may have insufficient examples from which to learn or retrieve. Synthetic augmentation has consequently been explored as a means of increasing representation of minority classes [16]. Such augmentation requires careful evaluation, however, because synthetic examples in an evaluation set may favour systems exposed to similar model-generated language. Separating performance on real and synthetic instances is therefore important when assessing whether gains extend to naturally occurring data.

Positioning of the present work. The literature above leaves a gap between classification-oriented dark-pattern detection and systems intended to support human inspection of potentially manipulative text. We address this gap by combining retrieval-based in-context classification with a structured output containing a predicted category, verbatim evidence span, psychological mechanism, harm dimension, rationale, and suggested rewrite. We compare multiple retrieval strategies and encoders and examine the approach on English as well as small German and Italian test sets. The multilingual experiments are treated as feasibility studies rather than definitive benchmarks, and evidence-span matching is evaluated as input grounding rather than as a direct measure of explanation faithfulness.

3. Dataset, Cross-Lingual Test Sets, and Augmentation

3.1. EC-DarkPattern Dataset

We use the EC-DarkPattern dataset [4], which contains 2361 product-description snippets labelled across eight categories. Each snippet is a short text excerpt from an e-commerce product listing and is assigned one of the eight dark-pattern categories used throughout this study. After de-duplication and removal of records with missing labels, we retain 2356 real examples.

The original dataset split provides limited representation of the rarest categories in the evaluation partition. To ensure that all eight categories are represented in the training, validation, and test sets, we construct a stratified 70/15/15 split using a fixed random seed. Synthetic examples for the three rarest categories are described separately in Section 3.4. The resulting augmented experimental corpus contains 2501 examples, of which 145 are synthetic.

The resulting class distribution remains imbalanced. NOT DARK PATTERN is the dominant category, whereas FORCED ACTION, SNEAKING, and OBSTRUCTION have substantially fewer examples. This imbalance motivates both the targeted augmentation of these categories and our comparison of random and retrieval-based demonstration selection. Because the evaluation split contains a small number of synthetic instances, we distinguish performance on the complete augmented test set from performance on the real-only subset in Section 6.

3.2. German Dark-Pattern Test Set

To evaluate cross-lingual transfer, we manually collected and annotated 102 candidate German product-description snippets from Otto, Zalando, Shein, and MediaMarkt, sampled across listing pages, subscription flows, and checkout screens. Three annotators (the paper authors) independently labelled each snippet using the EC-DarkPattern taxonomy, with the annotation guidelines extended by per-category German signal phrases (for example, “Nur noch X Stück” for SCARCITY and “automatische Verlängerung” for SNEAKING). Curation reduced the 102 candidates to 76 unique items after de-duplication, then to a final set of 71 test examples after excluding 5 snippets lacking a majority label. The resulting set has 6–12 examples per class and is balanced by design, unlike the English corpus (Table 1).

3.3. Italian Dark-Pattern Test Set

To extend the cross-lingual evaluation to a second target language, we additionally collected and annotated Italian product-description snippets from over twenty Italian e-commerce surfaces (fashion, books, classifieds, electronics, groceries, travel, telecoms, news paywalls, consumer-comparison portals; e.g. Yoox, IBS, Subito, ePrice, Coop, Booking, TIM, Sky, Corriere, Facile.it). One annotator labelled each snippet under the EC-DarkPattern taxonomy and performed two verification passes: a borderline re-check against the annotation guidelines, and a regulatory-and-scope pass that in particular relabelled items mandated by Article 17-bis of the Italian Consumer Code (transposing EU Omnibus Directive 2019/2161, which requires retailers to display the lowest price in the preceding 30 days) as NOT DARK PATTERN, since a legally required disclosure is not a manipulative interface element even when its strikethrough presentation visually resembles one. Per-row audit fields (original_category,

Table 1

Per-category counts of the German ($n=71$, three-annotator) and Italian ($n=77$, single-annotator) cross-lingual test sets. Both are balanced by design, unlike the English corpus. In the Italian set NOT DARK PATTERN is enlarged relative to a strict reading by the 12 items moved into it during verification, 10 of them out of MISDIRECTION (6 Art. 17-bis statutory disclosures, 3 free-text private-seller listings, 1 shipping-policy disclosure; Section 3.3).

Category	German	Italian
Not Dark Pattern	12	29
Forced Action	6	6
Misdirection	8	6
Obstruction	7	7
Scarcity	9	7
Sneaking	8	7
Social Proof	10	6
Urgency	11	9
Total	71	77

recategorization_reason) are preserved in the JSONL release so pre-verification labels remain recoverable under stricter taxonomic readings. The released set contains 77 verified examples across all eight categories with at least 6 per class (Table 1); the dominant class NOT DARK PATTERN accounts for 38% versus 47% in the training distribution. The Italian set is similar in size to the German set (77 vs. 71) but single-annotator, and the results below are reported as a feasibility probe rather than a parallel benchmark to German.

3.4. Synthetic Augmentation

The EC-DarkPattern corpus is strongly imbalanced, with particularly limited representation of FORCED ACTION, SNEAKING, and OBSTRUCTION. We therefore generate 145 synthetic examples for these three categories using Qwen3 8B. For each generation prompt, the model receives one or two real examples together with a structured description of the corresponding manipulative technique, psychological mechanism, and harm dimension, and is instructed to generate diverse paraphrases. A bigram Jaccard similarity filter with a threshold of 0.4 is subsequently applied to remove near-duplicate generations. The retained synthetic examples have distinct-2 ratios between 0.95 and 0.99 and pairwise bigram overlap below 0.015. The generation prompts are provided in the accompanying repository. Synthetic augmentation is applied only to the English EC-DarkPattern experiments. The German and Italian test sets described below are manually collected and annotated and contain no synthetically generated examples.

Of the 145 synthetic English examples, 103 occur in the training partition and 42 in the validation and test partitions, including 19 synthetic instances in the 376-example test set. Synthetic instances therefore constitute approximately 5% of the complete English test set. Because the generator used in the RAG pipeline is also Qwen3 8B, performance on these instances may partly reflect similarity to model-generated language rather than generalisation to naturally occurring e-commerce text. We consequently report the complete 376-instance evaluation for comparability with the experimental development setting and additionally report performance after excluding the 19 synthetic test instances. The latter provides a real-only evaluation over 357 naturally occurring examples. Examples of prompts used for the task are available here: [Github Link](#)

4. System Architecture

The proposed system combines example retrieval with structured generation for dark-pattern detection and explanation. Rather than asking the language model to infer a category from the input alone, the pipeline first retrieves semantically similar labelled examples and provides them to a locally hosted

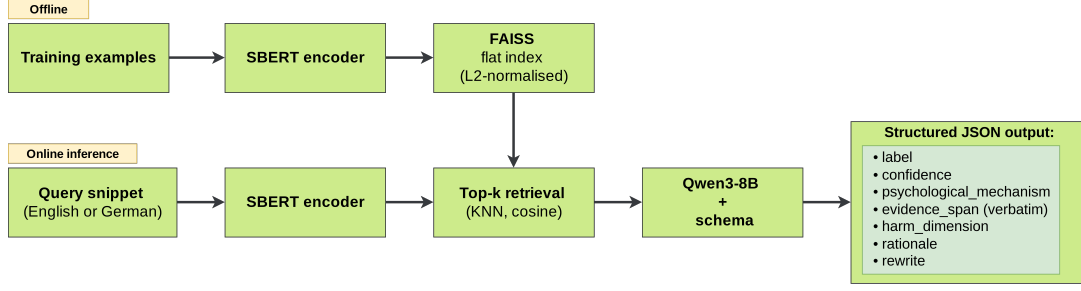


Figure 1: Overview of the retrieval-augmented dark-pattern pipeline. Offline, labelled training examples are encoded and stored in a FAISS index. Online, a query is encoded, the top- k examples are retrieved, and the language model generates a structured JSON output. The same pipeline is used for English, German, and Italian evaluation, with the retrieval configuration adapted as described in Section 5.8.

instruction-following language model as in-context demonstrations. The model then produces both a category prediction and a set of structured explanatory fields intended to support subsequent human inspection.

Figure 1 illustrates the two stages of the pipeline. During offline indexing, labelled training examples are embedded into a dense vector space and stored in a nearest-neighbour index. At inference time, the query is embedded using the corresponding encoder, relevant examples are retrieved, and these demonstrations are included in the prompt provided to the generator. The model returns a single JSON object containing the predicted label, confidence, psychological mechanism, evidence span, harm dimension, rationale, and suggested rewrite. The architecture therefore integrates retrieval, classification, and structured explanation generation within a single inference pipeline.

4.1. Offline Retrieval Index

Let x_i denote a labelled training example and $f(\cdot)$ the sentence encoder. Each example is represented by its L2-normalized embedding

$$z_i = \frac{f(x_i)}{\|f(x_i)\|_2}.$$

The normalized vectors are stored in a FAISS index, allowing cosine similarity to be computed through inner product. At inference time, the query is encoded using the same encoder as the indexed examples.

We compare Sentence-BERT (SBERT), BERT-base, and RoBERTa-large as English retrieval encoders, and multilingual SBERT for the cross-lingual experiments. SBERT is designed for semantic similarity and sentence-level retrieval [13], whereas the BERT and RoBERTa representations provide alternative encoder baselines. Multilingual SBERT provides a shared representation space across languages [14]. The retrieval encoder and selection strategy are varied while the downstream generator and structured output schema remain fixed.

For the augmented English experiments, synthetic training examples are indexed together with real training examples. Their inclusion increases the number of available demonstrations for the three underrepresented categories described in Section 3.4. The potential effect of synthetic data on evaluation is considered separately by reporting results on both the complete and real-only English test sets.

4.2. Neighbour Selection

For a query representation q and indexed example z_i , standard nearest-neighbour retrieval uses

$$s(q, z_i) = q^\top z_i.$$

The k examples with the highest similarity scores are supplied to the generator as in-context demonstrations. We use KNN as the primary retrieval strategy and compare it with three alternative selection methods.

Maximal Marginal Relevance (MMR) balances similarity to the query against redundancy among the selected demonstrations [11]. Prototype-based selection incorporates class-level representations when ranking candidate demonstrations. HyDE first generates a hypothetical representation of the query and uses that representation for retrieval [12]. Comparing these strategies allows us to examine whether direct semantic similarity, diversity, or class-level structure provides the most useful demonstrations for this task.

We additionally vary the number of demonstrations k . This tests whether increasing the retrieved context improves classification or whether a smaller set of highly similar examples is sufficient. The corresponding comparisons are reported in Section 6.

4.3. Structured Generation

The generator receives the input text, the retrieved demonstrations, and an instruction specifying a fixed JSON schema. It produces seven fields: `label`, `confidence`, `psychological_mechanism`, `evidence_span`, `harm_dimension`, `rationale`, and `rewrite`. The `label` is restricted to the eight EC-DarkPattern categories, while the `evidence_span` is instructed to reproduce the most directly relevant substring of the input. The remaining fields provide a structured description of the predicted manipulation and a suggested alternative formulation.

Formally, the generation step can be represented as

$$p(y, e, m, h, r, w \mid x, \mathcal{R}(x)),$$

where x is the input, $\mathcal{R}(x)$ denotes its retrieved demonstrations, y is the predicted label, e the evidence span, m the psychological mechanism, h the harm dimension, r the rationale, and w the suggested rewrite.

The evidence span is intended to make the textual basis associated with the generated output directly inspectable. We evaluate whether this field is reproduced as a verbatim substring of the input (Section 5). This measure establishes extractive grounding and schema compliance, but should not by itself be interpreted as evidence that the complete generated rationale is faithful to the model’s internal decision process.

The generator also produces a suggested rewrite intended to remove or reduce the manipulative wording while preserving the underlying product information. Rewrite adequacy is not independently evaluated in this study. We therefore treat this field as a candidate reformulation for subsequent human review rather than as a verified compliant or consumer-safe replacement.

4.4. Generation Model

We use Qwen3 8B [17] as the instruction-following generator. The model produces the complete structured output in a single generation call. Explicit chain-of-thought prompting is not used; instead, the prompt requests only the structured fields required by the task. This choice keeps the returned output concise and directly usable by downstream evaluation and auditing interfaces.

The generation model is held fixed across the principal retrieval comparisons so that differences among these configurations can be attributed to the examples supplied through retrieval rather than to a change in the downstream language model. Generation parameters and hardware are reported in Section 5.

4.5. Cross-Lingual Retrieval

For German and Italian evaluation, we retain the same general retrieval-generation architecture but replace the English retrieval representation with a multilingual sentence encoder. English training examples and target-language queries can therefore be represented in a shared multilingual embedding space.

The cross-lingual experiments do not constitute a pure encoder ablation. In the multilingual condition, the encoder is accompanied by the similarity-threshold and diversity settings described in Section 5.8. We consequently interpret the reported cross-lingual improvement as the effect of the complete multilingual retrieval configuration rather than attributing it solely to the encoder. This distinction is important because the individual contributions of these components are not separately ablated in the present study.

5. Experimental Evaluation

5.1. Generation Setup

The generator is Qwen3 8B [17], served locally through Ollama 0.17.0. We use a generation temperature of 0.0 to reduce sampling variability across configurations. JSON output mode is used where supported, together with a fallback parser that extracts the last valid JSON object when the model surrounds the structured response with additional text.

Inference is performed on a single NVIDIA GeForce RTX 5060 Laptop GPU with 8 GB VRAM. Retrieval encoders and the supervised classification baseline run on CPU to avoid memory contention with the generator. The same generator, generation settings, system prompt, output schema, and English evaluation split are used across the principal prompting and retrieval configurations.

5.2. Prompting and Retrieval Configurations

We evaluate eleven configurations designed to separate the effect of retrieval from the effect of simply providing demonstrations. The *no-retrieval* conditions comprise zero-shot prompting and random few-shot prompting with $k = 3$ and $k = 5$. The latter sample labelled demonstrations without using query-dependent retrieval and therefore control for the presence and number of examples in the prompt.

The retrieval-augmented conditions use query-dependent demonstration selection. With SBERT as the retrieval encoder, we compare KNN, Maximal Marginal Relevance (MMR), prototype-based retrieval, and HyDE at $k = 5$, together with a KNN configuration at $k = 2$. We additionally evaluate KNN retrieval with BERT-base and RoBERTa-large representations, including a RoBERTa $k = 2$ condition. All configurations use the same eight-class output taxonomy and structured generation schema.

This design allows two related comparisons. First, zero-shot and random few-shot prompting provide no-retrieval baselines against which the effect of query-dependent example selection can be assessed. Second, the retrieval configurations allow us to compare alternative strategies for selecting demonstrations while keeping the downstream generator fixed. Classification differences between these conditions can therefore be associated with the prompting and retrieval configuration; we do not, however, use this comparison alone to infer that retrieval causes higher-quality explanations.

5.3. Supervised Classification Baseline

As a supervised baseline, we fine-tune `bert-base-uncased` for five epochs using AdamW, a learning rate of 2×10^{-5} , and batch size 16. The baseline is trained on 1647 real training examples without synthetic augmentation and achieves a macro F1 of 0.785 on the corresponding evaluation reported in Section 6.

This baseline provides a reference for label-prediction performance, but it is not a like-for-like comparison with the complete RAG pipeline. The BERT classifier produces only category predictions, whereas the generative configurations additionally produce confidence, psychological mechanism, evidence span, harm dimension, rationale, and rewrite fields. Moreover, the BERT baseline and the augmented RAG evaluation differ in their exposure to synthetic data. We therefore avoid attributing differences between these systems solely to model architecture or retrieval.

5.4. English Evaluation Protocol

The primary English evaluation uses the stratified test partition described in Section 3. The complete test set contains 376 instances, including 19 synthetically generated examples from the three augmented minority categories. We refer to this as the *complete* English test set.

Because synthetic examples constitute approximately 5% of this test set and were generated using the same model family employed by the downstream generator, we additionally evaluate the saved predictions after excluding these 19 instances. This produces a *real-only* test subset containing 357 naturally occurring EC-DarkPattern examples. No predictions are regenerated for this analysis; the same model outputs are rescored after filtering the evaluation instances. Reporting both views allows us to distinguish performance on the experimental test partition from performance on naturally occurring examples alone.

Unless explicitly marked as real-only, the configuration results in Table 2 refer to the complete 376-instance test set. The real-only analysis is reported separately in Section 6.

5.5. Classification Metrics

Classification performance is evaluated using macro-averaged F1, precision, and recall over the fixed eight-class taxonomy. Macro averaging assigns equal weight to each category and is therefore particularly appropriate for the imbalanced English dataset. We also report Cohen’s κ as a chance-corrected measure of agreement between predicted and reference labels. Per-class results are inspected where relevant to determine whether aggregate improvements are concentrated in frequent categories or also extend to the less represented classes. Because the German and Italian test sets are small and deliberately more balanced than the English corpus, their results are interpreted primarily as cross-lingual feasibility evidence rather than as directly comparable benchmark scores.

Table 2

Pipeline comparison on the EC-DarkPattern v2 test set (376 examples), macro-averaged over the fixed eight-class taxonomy. Deployed RAG configuration in **bold** (SBERT KNN $k=5$); the supervised BERT baseline in *italics*. All RAG configurations were generated in a single inference session under one prompt, so the ranking is apples-to-apples.

Configuration	Macro F1	Macro P	Macro R	Cohen κ	Span Exact
<i>BERT fine-tuned</i>	<i>0.785</i>	<i>0.830</i>	<i>0.777</i>	<i>0.932</i>	—
Zero-shot	0.596	0.707	0.647	0.741	0.989
Random few-shot $k=5$	0.616	0.745	0.668	0.747	0.989
Random few-shot $k=3$	0.676	0.782	0.703	0.765	0.992
RAG SBERT HyDE $k=5$	0.817	0.834	0.851	0.851	0.997
RAG SBERT KNN $k=2$ (k -sweep)	0.832	0.881	0.847	0.893	0.995
RAG SBERT Prototype $k=5$	0.844	0.863	0.873	0.890	0.992
RAG SBERT Diversity $k=5$	0.851	0.873	0.873	0.908	0.997
RAG BERT KNN $k=5$	0.854	0.861	0.890	0.902	0.997
RAG RoBERTa KNN $k=2$	0.859	0.879	0.878	0.904	0.997
RAG SBERT KNN $k=5$	0.875	0.878	0.909	0.905	0.997
RAG RoBERTa KNN $k=5$	0.896	0.900	0.915	0.923	0.995

The strongest gains come from the retrieval-augmented configurations. The deployed SBERT KNN model with $k = 5$ achieves a macro F1 of 0.875, macro precision of 0.878, macro recall of 0.909, and Cohen’s $\kappa = 0.905$. This represents a substantial improvement over both zero-shot prompting and the fine-tuned BERT baseline (macro F1 = 0.785), while requiring no gradient-based fine-tuning. In addition to improved classification performance, the proposed RAG pipeline produces a complete structured explanation comprising the predicted label, evidence span, psychological mechanism, harm dimension, rationale, and consumer-safe rewrite, making the system more suitable for compliance-oriented auditing than a conventional classifier. Although the RoBERTa-large KNN configuration achieves a slightly

higher macro F1 (0.896), SBERT KNN with $k = 5$ was selected as the deployed configuration because it provides consistently strong performance while serving as the common retrieval backbone for both the English and cross-lingual experiments presented in this work.

The Table 2 also shows that the advantage is not driven by a single lucky metric. Precision, recall, and chance-corrected agreement all improve together, indicating that the model is not merely overpredicting frequent classes or exploiting a threshold artifact. In particular, the retrieval-based system preserves high recall while also increasing precision, which suggests that the retrieved examples are giving the generator both a better label prior and a better local decision boundary. This is important for dark-pattern detection because the cost of missing a manipulative cue can be higher than the cost of flagging a borderline case for review.

5.6. Evidence-Span Grounding

The structured output requires the model to return an `evidence_span` corresponding to a substring of the input. We evaluate this requirement using exact substring matching between the generated span and the original input text. We additionally inspect case-insensitive matches where appropriate. This metric measures whether the model follows the extractive evidence requirement of the output schema. A high match rate indicates that the returned evidence can be traced directly to the input, but it does not establish that the evidence is sufficient for the predicted category or that the generated rationale is faithful to the model’s internal decision process. We therefore refer to this measure as *evidence-span grounding* or *schema compliance*, rather than as an explanation-faithfulness metric.

5.7. Explanation-Quality Evaluation

Explanation quality is evaluated separately from classification and evidence-span grounding. For the deployed English RAG configuration, we conduct the rubric-based human evaluation described in Section 6.2. The evaluation considers coherence, faithfulness to the presented input, specificity, and non-circularity of the generated rationale. Importantly, this assessment is conducted only for outputs from the deployed retrieval configuration and does not include a matched no-retrieval explanation condition. It therefore provides descriptive evidence about the perceived quality of the generated explanations, but does not establish that retrieval itself causes an improvement in explanation quality. The zero-shot and random few-shot comparisons are used to quantify retrieval-related differences in classification performance, not explanation quality.

5.8. Cross-Lingual Evaluation

Cross-lingual performance is evaluated on the independently collected German and Italian test sets described in Sections 3.2 and 3.3. We report macro F1, Cohen’s κ , and accuracy, together with per-class inspection of categories that transfer successfully or remain difficult. For the multilingual retrieval condition, we use `paraphrase-multilingual-MiniLM-L12-v2` with KNN retrieval at $k = 5$, a similarity threshold of $\tau = 0.15$, and an MMR diversity parameter of $\alpha = 0.3$. The multilingual condition therefore changes more than the retrieval encoder alone. Because the encoder, similarity-threshold mechanism, and diversity fallback are not independently ablated, differences from the English-encoder condition are attributed to the complete multilingual retrieval configuration rather than to the multilingual encoder in isolation.

6. Results and Discussion

The results (Table 3) expose a two-tier transfer pattern. *Zero-shot* (F1 = 0.668) transfers signal-rich categories well: SCARCITY and SOCIAL PROOF reach F1 = 1.000 from unambiguous German surface signals (“nur noch X Stunden/Stück”), and URGENCY reaches 0.957. The rhetorical categories, by contrast, collapse (MISDIRECTION 0.000, SNEAKING 0.400). An intermediate English-encoder condition (SBERT,

Table 3

Cross-lingual evaluation on the German ($n=71$) and Italian ($n=77$) test sets, with English RAG SBERT KNN $k=5$ as in-language reference. Deployed multilingual configuration per language in **bold**; the Italian translated-index row is the k -sweep best.

Configuration	n	Macro F1	Cohen κ	Accuracy
English RAG SBERT KNN $k=5$ (reference)	376	0.875	0.905	0.931
<i>German ($n=71$)</i>				
Zero-shot	71	0.668	0.707	0.747
RAG Multilingual KNN $k=5$	71	0.845	0.854	0.873
<i>Italian ($n=77$)</i>				
Zero-shot	77	0.771	0.692	0.740
RAG Multilingual KNN $k=5$ (shared)	77	0.827	0.739	0.779
RAG Multilingual _{IT} KNN $k=1$ (transl.)	77	0.799	0.724	0.766

$k=1$, $F1 = 0.696$) barely improves on zero-shot, because the low-similarity English neighbours retrieved for German queries add little overall. Switching to the *multilingual encoder* ($k=5$ with the τ /MMR fallbacks) raises macro F1 to 0.845 ($\kappa = 0.854$, accuracy 87.3%): SCARCITY and SOCIAL PROOF reach 1.000, URGENCY 0.957, NOT DARK PATTERN 0.917, MISDIRECTION recovers to 0.875, while SNEAKING (0.545), OBSTRUCTION (0.667), and FORCED ACTION (0.800) remain the hardest cross-lingual categories.

6.1. Italian Probe

To check whether the cross-lingual pattern observed for German also holds for a second Romance-family target language, we replicate the zero-shot and RAG-multilingual configurations on the 77-example Italian test set (Section 3.3). The encoder is again paraphrase-multilingual-MiniLM-L12-v2, with identical hyperparameters to the German condition ($\tau=0.15$, $\alpha=0.3$). We evaluate two index variants: a *shared-space* index (English training texts encoded with the multilingual encoder, so Italian queries retrieve English neighbours in a shared embedding space; no translation involved), and a *translated* index (training texts machine-translated en $\rightarrow$ it via Helsinki-NLP/opus-mt-en-it and then encoded with the same multilingual encoder, so Italian queries retrieve Italian neighbours). The Italian set is small and single-annotator, so the results (Table 3, lower block) are reported as a feasibility probe rather than a benchmark contribution.

Italian results match the German transfer profile. The deployed shared-space configuration (multilingual encoder, KNN $k=5$, $\tau=0.15$, $\alpha=0.3$, identical to the German condition) reaches macro $F1 = 0.827$ on the 77-example Italian test set, preserving 94% of the in-language English reference (0.875, Table 3), comparable to German’s 97% (0.845/0.875). Unlike English and German, the Italian zero-shot baseline is already strong ($F1 = 0.771$): on this small, balanced, signal-rich set the model classifies well without retrieval, so RAG adds a smaller margin (+0.056) than it does on the imbalanced English corpus. The translated index built via Helsinki-NLP/opus-mt-en-it reaches $F1 = 0.799$ at $k=1$, below the shared-space $k=5$ result (0.827), so we again deploy the shared-space index. Per-class results follow the German two-tier profile: signal-rich categories such as FORCED ACTION ($F1 = 1.000$), OBSTRUCTION (1.000), SOCIAL PROOF (1.000), SCARCITY (0.875), and URGENCY (0.800) transfer cleanly, while MISDIRECTION ($F1 = 0.476$ after the Art. 17-bis verification pass) and NOT DARK PATTERN (0.696) are the hardest cross-lingual categories; its residual items rely on Italian-specific retail conventions (MSRP-based savings claims, unverifiable superlative price labels) that lack close English neighbours in the EC-DarkPattern index. The cross-lingual transferability of the pipeline is thus a property of the multilingual-encoder-plus-RAG architecture, not of any particular Romance-family target.

6.2. Human Evaluation of Explanation Quality

Classification performance alone does not establish whether the generated explanations are useful for human interpretation. We therefore conducted a rubric-based human evaluation of the explanations produced by the deployed configuration (RAG SBERT KNN, $k = 5$). Four explanation-quality dimensions were assessed using a five-point Likert scale: *Coherence* (fluency and logical consistency), *Faithfulness* (whether the explanation reflects the input without unsupported claims), *Specificity* (whether the explanation refers to the concrete manipulative phrase rather than providing a generic description), and *Non-circularity* (whether the explanation provides meaningful reasoning rather than merely restating the predicted label). The composite score is the sum of these four dimensions (maximum = 20). We deliberately restrict the evaluation to explanation quality; rewrite adequacy and psychological mechanism correctness are left for future work.

We sampled 30 predictions stratified across the eight categories according to their frequency in the test set (27 correctly classified and 3 misclassified instances; records 6, 9, and 14). Six independent annotators (anonymised as Annotators 1–6) evaluated each explanation, and their assessments were compared with the automated rubric-based evaluator Prometheus 7B v2.0 [18].

Overall, both the human evaluators and Prometheus assign relatively high scores to the generated explanations. Coherence receives the highest average human score (4.55), suggesting that the explanations are generally fluent and logically consistent. Specificity receives the lowest average score (4.16), indicating that some explanations remain relatively generic despite correctly identifying the underlying manipulation. Prometheus assigns similar overall ratings, although it evaluates Specificity slightly more favourably (4.53) and Non-circularity somewhat more conservatively (4.03 versus 4.27). The mean composite scores are also comparable (17.29 for human evaluation and 17.93 for Prometheus).

These results should, however, be interpreted with appropriate caution. The six annotators exhibited noticeable differences in scoring behaviour, with mean composite scores ranging from 15.17 to 19.07 out of 20, resulting in weak chance-corrected inter-annotator agreement. Consequently, the human evaluation should be regarded as an initial feasibility assessment of explanation quality rather than a definitive benchmark. A larger evaluation with refined annotation guidelines, additional annotators, and explicit assessment of rewrite quality and psychological mechanism identification constitutes important future work.

7. Conclusion

This paper investigated whether retrieval-augmented generation can extend textual dark-pattern detection beyond label prediction toward structured, human-readable decision support. By combining dense example retrieval with a locally hosted instruction-following model, the proposed pipeline produces a category prediction together with a verbatim evidence span, psychological mechanism, harm dimension, rationale, and suggested rewrite in a single structured output. The experiments show that query-dependent retrieval substantially improves classification over zero-shot and random few-shot prompting on the EC-DarkPattern evaluation. The deployed SBERT KNN configuration achieves strong performance without gradient-based fine-tuning, while the broader comparison demonstrates that retrieval strategy and representation choice affect classification performance. Cross-lingual experiments further provide preliminary evidence that multilingual retrieval can support transfer to German and Italian without target-language fine-tuning. Human evaluation indicates that the generated explanations are generally coherent and useful, although the limited sample size and weak inter-annotator agreement mean that these findings should be interpreted as feasibility evidence rather than definitive validation of explanation quality.

Overall, the results suggest that retrieval-augmented generation is a promising approach for extending dark-pattern detection from classification toward inspectable, multilingual decision support. Future work should evaluate the generated rewrites and psychological mechanisms directly, strengthen multilingual annotation and benchmarking, and investigate explanation faithfulness beyond verbatim evidence-span grounding.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4.5 and Google Gemini 3.1 for grammar and spelling checking. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, A. Narayanan, Dark patterns at scale: Findings from a crawl of 11k shopping websites, in: *Proceedings of the ACM on Human-Computer Interaction*, volume 3, 2019, pp. 81:1–81:32. doi:10.1145/3359183.
- [2] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, A. L. Toombs, The dark (patterns) side of UX design, in: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 534:1–534:14. doi:10.1145/3173574.3174108.
- [3] European Commission, Regulation (EU) 2022/2065 of the european parliament and of the council (Digital Services Act), *Official Journal of the European Union*, 2022.
- [4] S. Yada, A. Mathur, T. Naous, M. Guo, EC-DarkPattern: A dataset and benchmark for dark pattern detection in E-Commerce product listings, *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2022) 1–14.
- [5] H. Brignull, Dark patterns: Deception vs. honesty in UI design, *Interaction Design, Smashing Magazine*, 2010.
- [6] K. Bongard-Blanchy, A. Rosset, S. Rivas, S. Doublet, V. Koenig, G. Lenzini, “I am definitely manipulated, even when I am aware of it. it’s just too convenient!”: Dark patterns from the end-user perspective, in: *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, 2021, pp. 763–776. doi:10.1145/3461778.3462086.
- [7] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, 2019, pp. 312–317.
- [8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in Neural Information Processing Systems* 33 (2020) 9459–9474.
- [9] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen, What makes good in-context examples for GPT-3, in: *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 2022, pp. 100–114. doi:10.18653/v1/2022.deelio-1.10.
- [10] O. Rubin, J. Herzig, J. Berant, Learning to retrieve prompts for in-context learning, in: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, 2022, pp. 2655–2671. doi:10.18653/v1/2022.naacl-main.191.
- [11] J. Carbonell, J. Goldstein, The use of MMR, diversity-based reranking for reordering documents and producing summaries, in: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 335–336. doi:10.1145/290941.291025.
- [12] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 1762–1777. doi:10.18653/v1/2023.acl-long.99.
- [13] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.
- [14] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 4512–4525. doi:10.18653/v1/2020.emnlp-main.365.

- [15] G. I. Winata, A. Madotto, Z. Lin, R. Liu, J. Yosinski, P. Fung, Language models are few-shot multilingual learners, in: Proceedings of the 1st Workshop on Multilingual Representation Learning, 2021, pp. 1–15. doi:10.18653/v1/2021.mrl-1.1.
- [16] D. Kochanek, A. Grumpe, C. Wöhler, Similarity-based composite prompting for LLM data augmentation in low-resource text classification, arXiv preprint arXiv:2403.12345, 2024.
- [17] Q. Team, Qwen technical report, arXiv preprint arXiv:2309.16609, 2023.
- [18] S. Kim, J. Shin, Y. Cho, J. J. Han, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, M. Seo, Prometheus: Inducing fine-grained evaluation capability in language models, in: Proceedings of the Twelfth International Conference on Learning Representations, 2024. URL: <https://openreview.net/forum?id=8euJaTveKw>.