International Journal of Geo-Information

MDPI

Article

LLM-Assisted Detection and Community-Based Review of Promotional Texts in OpenStreetMap: An Empirical Study

Pascal Neis

Department of Applied Informatics and Geodesy, School of Technology, Mainz University of Applied Science, 55128 Mainz, Germany; pascal.neis@hs-mainz.de

Abstract

Promotional and SEO-oriented free-text content can undermine the neutrality and quality of OpenStreetMap (OSM). Because such text is multilingual and context-dependent, purely rule-based detection is challenging. This study evaluates large language models (LLMs) for identifying promotional content in OSM description=* tags and investigates their use in a community-based quality assurance workflow. A classification prompt was developed using a pilot dataset and evaluated on a balanced, manually annotated multilingual dataset of 1500 descriptions. Thirty-five open-weight model configurations were compared using precision, recall, F1-score, and runtime, and two conventional text-classification baselines were evaluated for classification performance. The best-performing model, llama3.3:70b, achieved 95.03% precision, 91.87% recall, and an F1-score of 93.42%, outperforming the evaluated conventional baselines. This model was then applied to 16,156 newly added or modified OSM descriptions in a real-world workflow. It identified 2738 descriptions as potentially promotional, resulting in 2599 public changeset comments. Within a 28-day follow-up period, 79.5% of analyzed commented cases were modified or removed, compared with 27.3% during a later comparison period without workflow-generated comments. Because the two conditions were temporally separated and non-randomized, this difference should be interpreted as an observational association.

Keywords: large language models; OpenStreetMap; promotional text detection; human-in-the-loop; community-based quality assurance

Check for updates

Academic Editors: Wolfgang Kainz, Zhipeng Gui and Huayi Wu

Received: 31 August 2026
Revised: 8 November 2026
Accepted: 11 November 2026
Published: 18 November 2026

Copyright: © 2026 by the author. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

OpenStreetMap (OSM) is a major collaborative platform for the collection, maintenance, and dissemination of open geospatial data [1]. As a prominent example of Volunteered Geographic Information (VGI), OSM relies primarily on contributions from a large community of volunteer mappers who collect, edit, and maintain geographic information [2,3]. At the same time, a considerable share of contributions is produced by paid mappers, often referred to as corporate editors, working for large companies. The size and composition of these corporate editing teams have changed markedly over time [4,5], and organizations engaged in such organized editing are expected to document their activities publicly [6].

OSM itself is still largely shaped by its open community of contributors. Once registered, users can directly create, edit, discuss, and delete geographic features. In many regions this openness produces detailed, current data, but it also opens the door to accidental errors, inconsistent tagging, and deliberate manipulation. The OSM community distinguishes ordinary editing mistakes from vandalism, which is generally understood

as the intentional disregard of established community editing norms [7]. Whereas conventional OSM errors often result from limited experience or incomplete local knowledge, promotional and search engine optimization (SEO)-oriented contributions are more likely to reflect the contributor's own commercial interests. Such content may occur in free-text tags such as description=* and primarily serve promotional purposes rather than geographic description. Distinguishing it from legitimate descriptive content can be difficult, particularly when it refers to an existing real-world feature. OSM tagging conventions emphasize that tags should describe verifiable properties of real-world objects. In particular, the description=* tag is intended to contain additional information about a feature rather than advertising, promotional, or SEO-oriented additions [8]. Such uses may undermine the neutrality and reliability of OSM data.

Automatically identifying such content is challenging. Promotional language varies widely across languages and writing styles, ranging from short persuasive phrases and superlatives to longer, keyword-rich descriptions. At the scale of OSM, this challenge is compounded by the volume of daily edits and the diversity of possible tag combinations. Manual review alone is impractical at this scale. Detecting promotional content is therefore a context-sensitive semantic classification task rather than a purely lexical task.

Previous research has established conventional text-classification approaches for spam detection and has increasingly explored large language models (LLMs) for geospatial tasks. However, empirical evidence on the use of LLMs for detecting promotional free-text content in OSM remains limited, particularly in a multilingual and community-governed setting. It is therefore still necessary to assess both the technical suitability of such models and their practical integration into existing community review processes. The following research questions guide the analysis:

- RQ1: How accurately can current open-weight LLMs identify promotional or SEO-oriented content in OSM description=* tags, and how does their performance compare with conventional text-classification baselines?
- RQ2: Which LLM configurations provide the most suitable trade-offs between classification performance, runtime, and practical deployment requirements for a community-facing classification workflow?
- RQ3: How are the likelihood and timing of subsequent modifications associated with the presence or absence of a public changeset comment for potentially promotional entries?

The study therefore contributes both a controlled comparison of LLM-based promotional-text classification in OSM and an empirical assessment of its integration into a community-based review workflow. The proposed approach enables the multilingual detection of promotional content in OSM free-text tags without requiring task-specific model fine-tuning. Furthermore, it integrates automated classification into an established community review mechanism, preserving human judgment and community oversight. By combining controlled experiments with a real-world observational evaluation, it connects model-level performance with the practical requirements and limitations of deploying AI-assisted review in a collaborative VGI environment.

2. Related Work

Research on OSM data quality has been a central topic in the academic study of VGI since the early stages of the field [9]. The openness of the project enables highly current and detailed geographic information, but also gives rise to spatial, thematic, and temporal variation in data quality. Early research therefore focused primarily on comparisons between OSM and authoritative or commercial reference datasets.

The accuracy of OSM data in England and London was assessed using Ordnance Survey data as a reference [10]. The study showed that OSM can achieve considerable geometric

accuracy in well-mapped areas, while completeness varies across regions. A subsequent assessment in France extended this perspective by considering not only positional accuracy but also semantic, temporal, and logical dimensions of data quality [11]. Similarly, an analysis of the development of the OSM road network in Germany found that OSM can be comparable to commercial datasets for certain navigation-related applications, although challenges related to topology and turn restrictions remained [12]. Studies of building footprints have reported comparable patterns. In Munich, OSM buildings exhibited high levels of semantic accuracy and completeness, although their geometries were in some cases simplified relative to authoritative data [13]. Similar findings were reported for building footprints in Lombardy [14].

Intrinsic approaches extend OSM quality assessment beyond comparisons with external reference datasets by incorporating editing histories, contributor activity, and tagging structures [15]. However, these approaches primarily address properties of the data and its production process. Contributions that are deliberately introduced for promotional, disruptive, or other non-mapping purposes raise a different set of quality and credibility concerns.

Vandalism represents a particular challenge for OSM and other VGI projects. A rule-based approach to the automated detection of vandalism in OSM examined different types of suspicious edits and used features derived from changesets, user profiles, and individual edits to identify potentially harmful contributions at an early stage [16].

A study using Pokémon GO as a case showed that external incentives can trigger deliberate cartographic manipulation in OSM [17]. This suggests that problematic edits do not necessarily arise solely from destructive intent, but may also be encouraged by incentives created through external platforms that rely on OSM data. More recent work has proposed attention-based approaches that combine information from changesets, user characteristics, and individual edits [18], extending earlier rule-based and outlier-detection approaches with more flexible data-driven methods. Spam and commercially motivated manipulation constitute related but distinct forms of problematic contributions. Whereas vandalism is commonly associated with destructive or provocative edits, spam and SEO-oriented content are typically driven by promotional or commercial interests. Comparable commercially motivated content has been studied outside VGI, both in research on fake and deceptive reviews [19] and in an analysis of fraudulent business listings on a commercial map service, which identified a form of search engine optimization targeting local listing services [20]. These platforms differ from OSM in their data models and governance, but show that place-related content is a recurring target of commercial manipulation. In the context of VGI, such contributions may affect not only the quality of individual data objects but also the perceived credibility of contributors and the dataset as a whole. Previous research has shown that techniques derived from email spam detection can also be applied to credibility assessment in crowdsourced geographic data [21].

Research on spam and vandalism detection in Wikipedia and Wikidata provides an additional methodological perspective. Like OSM, these platforms rely on open, collaborative editing processes and maintain versioned contribution histories. They therefore face related challenges in distinguishing legitimate contributions from misuse, spam, and deliberate manipulation. Automated vandalism detection in Wikipedia has been formulated as a classification task in which problematic edits are identified using textual and metadata-based features [22]. Other approaches have combined natural-language features, edit metadata, and contributor reputation signals [23], while separate work has focused specifically on spam users and their editing behavior [24].

Detecting promotional or SEO-oriented content in OSM free-text tags is closely related to text classification and spam detection. Text classification involves assigning textual content to predefined categories, for example, “spam” versus “non-spam,” or “promotional” versus

"neutral." Traditional approaches commonly rely on representations such as Bag-of-Words or TF-IDF in combination with classifiers including Support Vector Machines, Logistic Regression, or Random Forest. Comprehensive reviews summarize the development of both traditional and more recent approaches to text classification [25,26].

Efficient baseline methods such as fastText have demonstrated strong performance on large-scale text classification tasks [27]. Transformer-based models such as BERT represent a major methodological development by using contextualized representations that enable models to capture semantic relationships beyond isolated lexical features [28].

This context sensitivity is relevant to OSM free-text content, where promotional intent cannot always be inferred from individual keywords. A website=* tag or a description of a commercial feature, for example, may be entirely legitimate if the content remains factual and relevant to the mapped object. Context-dependent interpretation has also become relevant in recent work on language models for geospatial applications.

Recent reviews show that GPT-based models, LLMs, and generative AI are being applied across a broad range of geospatial tasks, including natural-language interfaces, spatial analysis, map interpretation, and GIS workflows [29]. Evaluations of LLMs on spatial concepts indicate that models can perform well on fundamental concepts, while their performance tends to become less consistent as the complexity of spatial reasoning increases [30]. Further research suggests that performance on geospatial tasks may also depend substantially on model selection and prompting strategy [31]. Approaches such as BB-GeoGPT, GeoGPT, GTChain, and ChatGeoAI illustrate how language models can be combined with geographic knowledge, GIS tools, and tool-using workflows [32–35]. These systems primarily focus on spatial reasoning, GIS automation, tool use, and the translation of natural-language instructions into geospatial operations.

Prompting strategies are an important component of LLM-based classification. Zero-shot prompting allows a model to perform a classification task based on instructions alone, without task-specific labeled examples. Few-shot prompting incorporates selected examples into the prompt and can provide the model with task-specific distinctions and decision boundaries [36]. In the OSM context, examples such as neutral business descriptions, promotional formulations, or borderline cases can convey distinctions that generic instructions alone may not capture. Classification results may depend on prompt formulation, model version, and inference settings [37]. These sources of variation need to be considered when evaluating LLMs for OSM-specific classification tasks.

3. Materials and Methods

The research design combines a controlled evaluation of LLM-based classification approaches with their practical application in OSM. It follows a multi-stage, sequential structure covering the period from January 2025 to August 2026. The individual phases and associated data-processing workflow are summarized in Figure 1. The design was structured to address the three research questions in sequence. RQ1 was examined using the Balanced Evaluation Set to compare the classification performance of the tested open-weight LLMs and conventional baselines. RQ2 used the same benchmark to compare LLM configurations in terms of classification performance, runtime, and practical deployment requirements. RQ3 was addressed through the real-world observational evaluation by comparing the likelihood and timing of subsequent modifications in the commented and non-commented conditions.

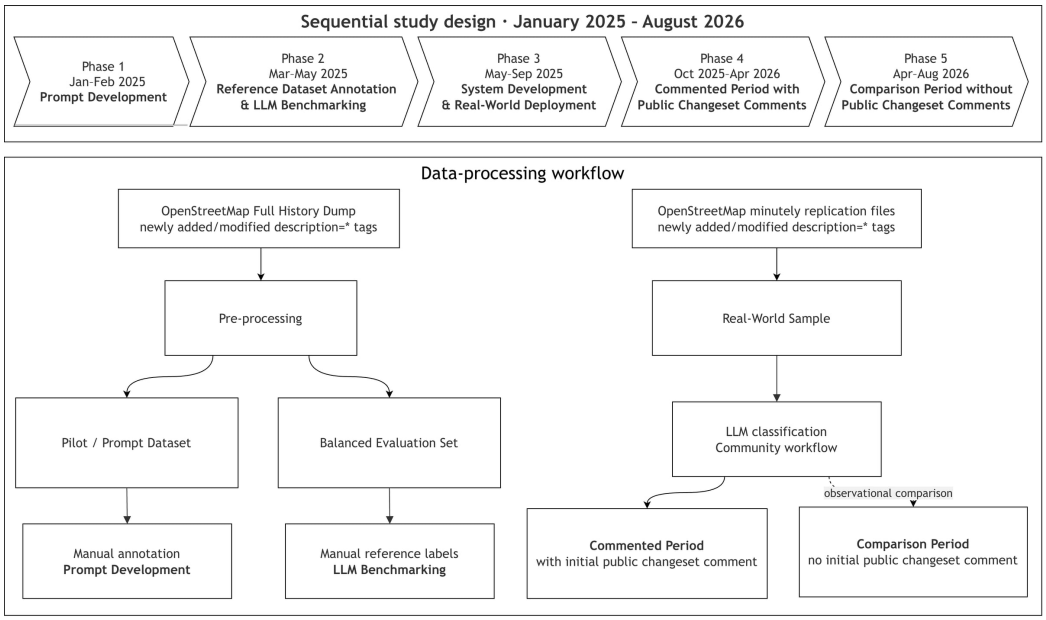


Figure 1. Sequential study design and data-processing workflow.

The study began with the extraction of historical key-value pairs containing description=* tags and the creation of a Pilot Dataset to support the development and iterative refinement of the classification prompt. A separate manually annotated reference dataset was then constructed and used for the systematic comparison of several locally deployed language models. Based on the results of this evaluation, the selected classification approach was transferred to a real-world setting and integrated into a community-based human-in-the-loop workflow. In an evaluation phase, the study examined whether potentially promotional description=* tags were subsequently modified following the public changeset comments posted through the workflow. It further assessed how the likelihood and time to modification differed from those observed during a comparison period in which no public changeset comments were posted. Keeping the phases separate in time and in method was intended to prevent prompt development, model evaluation, and the workflow assessment from influencing one another.

3.1. Data Selection and Corpus Construction

Two OSM data sources were used in the study. The OSM Full History Planet Dump [38] provided the historical data for the Pilot and Balanced Evaluation Datasets, whereas the real-world sample was collected from the OSM minutely replication files [39]. Across all datasets, edits were extracted in which a description=* tag was newly added to or modified on an OSM object. For each observation, the description text, edit timestamp, affected OSM object, contributor account, and associated changeset were recorded. Contributions were then filtered based on the contributor's previous mapping activity, and duplicate description texts were removed. For both the Pilot and Balanced Evaluation Datasets, only descriptions added or modified by accounts with fewer than 42 active mapping days up to and including the date of the respective edit were retained. The threshold was adopted as an OSM-specific heuristic based on the OSM Foundation's Active Contributor Membership criterion [40]. An active mapping day was defined as a calendar day on which the account had created at least one changeset containing an OSM edit. The criterion was used to focus the datasets on contributions from relatively new or infrequently active accounts.

To support consistent manual annotation across languages, the remaining descriptions were translated into English using a locally deployed llama3.3:70b model via Ollama [41].

The translations were used only as an annotation aid; all model evaluations were performed on the original, unmodified descriptions.

The resulting historical data were used to create two temporally separate datasets. The Pilot and Prompt Development Dataset contained 400 descriptions added between 1 and 31 December 2024, comprising 200 promotional and 200 neutral examples. After manual annotation, this dataset was used exclusively for prompt development and the selection of few-shot examples and was not included in the subsequent model comparison.

The Balanced Evaluation Set contained 1500 descriptions added between 1 January and 28 February 2025, comprising 750 promotional and 750 neutral examples. These manually assigned labels served as reference labels for benchmarking the LLM-based classification approaches. The balanced class distribution was chosen to enable controlled comparison using precision, recall, F1-score, and runtime. As this distribution does not reflect the natural prevalence of promotional descriptions in OSM, the results are interpreted as measures of comparative classification performance rather than prevalence estimates.

Following prompt development and model benchmarking, the selected classification approach was deployed on newly added or modified description=* tags obtained from the OSM minutely replication files. The resulting real-world sample covered the period from 1 October 2025 to 15 July 2026. In contrast to the historical datasets, only descriptions added or modified by accounts with no more than three active mapping days at the time of the respective edit were included. This stricter criterion limited the number of cases requiring manual review, community interaction, and follow-up.

Each case in the real-world sample passed through the LLM-based classification and subsequent community workflow. The sample was divided into two consecutive observation periods. During the real-world evaluation from 1 October 2025 to 14 April 2026, potentially promotional descriptions identified by the classification workflow received an initial public changeset comment as part of the study. During the Comparison Period from 15 April to 15 July 2026, no such initial public changeset comment was posted by the study. Cases in both periods were followed for 28 days to record subsequent contributor responses, modifications, or reverts.

The Pilot and Balanced Evaluation Datasets were manually annotated by a single researcher using the binary categories promotional and neutral. Annotation was based on the original description=* text and, where necessary, its English translation. A description was classified as promotional when its primary communicative function was to advertise, attract customers, improve search-engine visibility, or otherwise promote a business, service, or organization rather than provide factual information about the mapped feature. Relevant indicators included superlatives, prices or discounts, calls to action, excessive positive self-presentation, promotional emoji use, and keyword-like passages with little factual information. Classification decisions were based on the communicative function of the description as a whole rather than on individual words or expressions.

To assess the reliability of the reference labels, two additional annotators independently evaluated two separately drawn random samples of 300 descriptions each from the Balanced Evaluation Set. The annotators were blinded to the original reference labels and independently assigned each description to one of the same two categories using the same annotation criteria. Agreement between each additional annotator and the original reference annotation was quantified using percentage agreement and Cohen's kappa.

For descriptive analysis, the country, original language, and thematic object category were assigned where possible. Country information was derived from the spatial location or geometry of the OSM object using Nominatim [42], while the language of the original description was identified using the locally deployed llama3.3:70b model via Ollama [41]. OSM objects were categorized using a rule-based classification scheme derived from the

tags accompanying description=*. The rules were derived from the OSM Map Features documentation [43] and considered 21 relevant primary keys. Selected amenity=* values were grouped into broader thematic categories, while objects that could not be assigned using the considered keys were classified as unclassifiable.

3.2. LLM-Based and Conventional Classification

The primary classification approach used open-weight LLMs executed locally through Ollama [41] to automatically identify promotional description=* tags. All comparative experiments were conducted on Ubuntu 24.04 LTS using a workstation equipped with two NVIDIA RTX 4090 GPUs, each providing 24 GB of VRAM. Running the models locally provided a common inference environment and kept the processed OSM data on the local system. It also enabled the evaluation of open-weight models of different sizes without relying on proprietary APIs. This choice is relevant in the context of OSM, which is an open-data project largely maintained by volunteers. Relying on commercial API-based models for continuous quality assurance would introduce recurring operational costs and dependencies on external service providers, potentially limiting the long-term sustainability of such an approach. Open-weight models offer greater independence and flexibility for community-oriented applications, although their deployment still requires computational resources and associated operating costs.

During the exploratory phase of the study, 35 models ranging from approximately 1 billion to 405 billion parameters were considered. Candidate models were selected based on their availability in the local inference environment, multilingual capabilities, suitability for semantic text classification, and practical feasibility on the available hardware.

The classification procedure used a fixed system prompt and a standardized user message. The system prompt defined the classification task, the two target classes, relevant indicators of promotional or SEO-oriented content, and a conservative decision rule for ambiguous cases. In addition, 20 multilingual few-shot examples were provided, comprising ten promotional and ten neutral descriptions. The examples were selected to illustrate both clear and borderline cases relevant to the OSM context. The complete prompt, few-shot examples, and message structure are provided in Appendix A (Listing A1 and Listing A2).

Prompt development was conducted exclusively using the Pilot and Prompt Development Dataset. During this phase, the classification criteria, prompt formulation, and few-shot examples were iteratively refined. Proprietary models, including GPT-4, o1, and GPT-4.5, were additionally used during prompt development to explore alternative formulations and identify potential borderline cases. These proprietary models were not included in the subsequent comparative evaluation, and the Balanced Evaluation Set was not used during prompt development. The model deployed in the real-world workflow was selected only after the comparative evaluation on the Balanced Evaluation Set.

For automated evaluation, model output was constrained to the Boolean values True and False, with True representing promotional or SEO-oriented content and False representing neutral content. Responses were normalized before comparison with the manually assigned reference labels. Inference parameters were kept constant across models wherever supported. The temperature was set to 0 to reduce stochastic variation in model outputs. The context window was configured to 2048 tokens, which was sufficient to accommodate the complete prompt, few-shot examples, and individual description texts. Other inference parameters were left at default values.

The classification pipeline began by loading the original description=* text from the prepared dataset and incorporating it into a standardized prompt containing fixed instructions and few-shot examples. This prompt was then submitted to each model through the local Ollama instance. The generated responses were subsequently normalized and stored

before being compared with the corresponding manually assigned reference labels. In addition to class predictions, inference-time measurements were recorded for each model. Runtime was measured as the total wall-clock time required to classify the complete Evaluation Dataset. Timing began after the model had been loaded into the inference environment; model loading time was therefore excluded. No separate warm-up runs were performed. The resulting dataset-level runtime was used as the basis for comparing computational efficiency across models.

For comparison with the LLM-based approaches, two conventional supervised text-classification baselines were additionally evaluated: word-level TF-IDF combined with Logistic Regression and character-level TF-IDF combined with a linear Support Vector Machine (SVM). Both approaches used the original, untranslated description=* texts as input. The classifiers were trained on the 400 manually annotated descriptions from the Pilot and Prompt Development Dataset. Hyperparameters were selected using stratified five-fold cross-validation on this dataset, with F1-score as the optimization criterion. For Logistic Regression, regularization strengths C of 0.1, 1, and 10; word n-gram ranges of (1,1) and (1,2); and minimum document frequencies of 1 and 2 were evaluated. For the linear SVM, character n-gram ranges of (3,5), (3,6), and (4,6); minimum document frequencies of 1 and 2; and the same values of C were considered. The Balanced Evaluation Set was not used during training or hyperparameter selection, and no normalized description texts overlapped between the two datasets. The final classifiers were evaluated on the same 1500 descriptions as the LLM configurations.

3.3. Evaluation Metrics

The classification performance of the tested models was evaluated on the Balanced Evaluation Set using precision, recall, and F1-score [44]. The promotional class was defined as the positive class. Predictions were compared with the manually assigned reference labels to obtain true positives, false positives, true negatives, and false negatives (Table 1).

Table 1. Confusion matrix for the classification of promotional and neutral descriptions.

	Predicted: Promotional	**Predicted: Neutral**
Actual: Promotional	TP	FN
Actual: Neutral	FP	TN

Precision, recall, and F1-score were calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F}_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Precision is particularly relevant in the community-facing workflow because false positives may result in unnecessary review or interaction with contributors, whereas recall reflects the proportion of promotional descriptions identified. The F1-score was used as the primary criterion for model comparison and deployment because the workflow required a balance between identifying promotional content and limiting false-positive classifications.

3.4. Design of the Community Review Workflow

As part of this study, an LLM-assisted workflow was developed for the detection and community-based review of potentially promotional description=* tags (Figure 2).

The workflow was designed so that the LLM does not make any direct modifications to OSM data. Instead, it is used to identify potentially problematic descriptions and initiate a transparent community review process, while the assessment of the reported case and any subsequent modification remain the responsibility of human contributors within the OSM community.

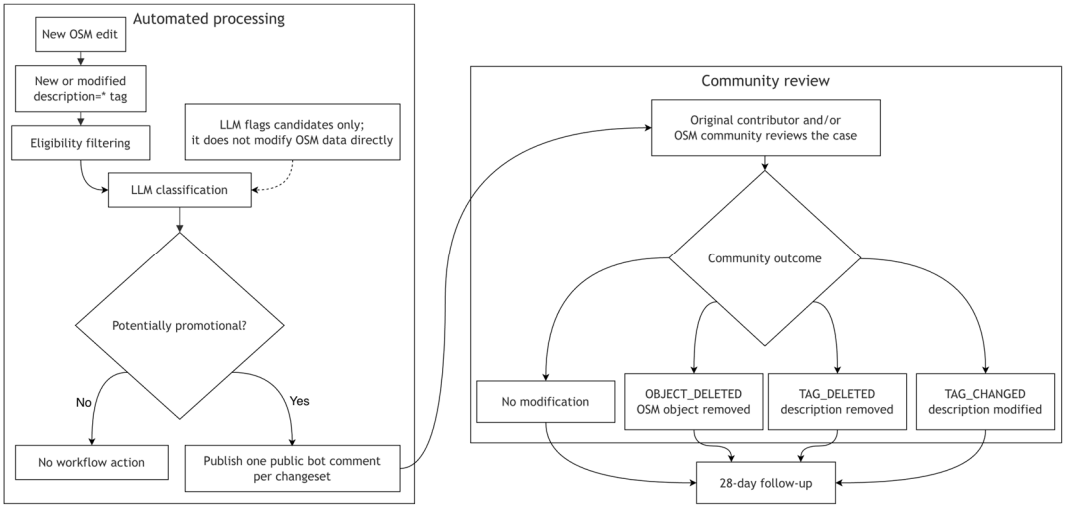


Figure 2. Human-in-the-loop workflow for the detection and community-based review of potentially promotional description=* tags in OSM.

First, newly added or modified description=* tags are detected automatically and subjected to the eligibility criteria described in Section 3.1. Eligible descriptions are then classified by the selected LLM-based approach. Entries classified as neutral receive no further processing within the workflow. For entries classified as potentially promotional, the associated changeset is identified, and a public changeset comment is posted. The automated commenting workflow was designed to reflect key principles of OSM community practices, such as transparency and accountability. Although the workflow did not perform automated modifications to OSM data, its design was informed by principles outlined in the OSM Automated Edits Code of Conduct [45]. A dedicated OSM account [46] was used to make the automated activity publicly identifiable and attributable. The account's profile page documented the process, providing community members with information about the purpose and operation of the workflow. The workflow used public changeset discussions, an established OSM mechanism for providing feedback on specific edits [47]. Comments were formulated as requests for review rather than definitive judgments. The messages described the detected content as potentially problematic rather than as a confirmed violation, allowing contributors to assess the reported content, respond to the comments, and modify the affected entries where appropriate.

If a single changeset contained multiple descriptions classified as potentially promotional, only one public comment was posted for that changeset. Consequently, the number of classified descriptions could exceed the number of published changeset comments. The comment informed the contributor about the potentially problematic content and provided an opportunity for the original contributor or other community members to inspect, discuss, and, where considered appropriate, modify the affected OSM entry.

Following publication of a changeset comment, each affected entry was monitored for a standardized follow-up period of 28 days. Four observable outcomes were distinguished: modification of the description=* value (TAG_CHANGED), removal of the description=*

tag (TAG_DELETED), deletion of the complete OSM object (OBJECT_DELETED), or no observed modification during the follow-up period (NO_MODIFICATION).

The real-world evaluation comprised a commented condition and a comparison condition without public changeset comments. In both conditions, potentially promotional entries identified by the same classification procedure were observed for 28 days. Modification rate and time to modification were then compared between the two conditions.

4. Results

4.1. Composition of the Pilot and Evaluation Datasets

The Pilot and Prompt Development Dataset and the Balanced Evaluation Set comprised 400 and 1500 manually annotated descriptions, respectively. In both datasets, the promotional and neutral classes were equally represented. The descriptions in the Pilot Dataset originated from 99 countries and were written in 42 different languages. The Evaluation Dataset covered a broader geographic and linguistic range, comprising descriptions from 134 countries and 46 source languages. In the Pilot Dataset, English, French, and Spanish were the three most frequently represented languages and together accounted for 65.25% of all descriptions. In the Evaluation Dataset, English, German, and Spanish were the three most frequent languages, jointly accounting for 71.80%. The corresponding language distributions are summarized in Tables 2 and 3.

Table 2. Language distribution of the Pilot and Prompt Development Dataset.

Language	Promotional	Neutral	Total	Share (%)
English	108	93	201	50.25
French	20	15	35	8.75
Spanish	11	14	25	6.25
Other languages	61	78	139	34.75
Total	200	200	400	100.00

Table 3. Language distribution of the Balanced Evaluation Set.

Language	Promotional	Neutral	Total	Share (%)
English	516	350	866	57.73
German	50	76	126	8.40
Spanish	24	61	85	5.67
Other languages	160	263	423	28.20
Total	750	750	1500	100.00

The two datasets also differed in terms of description length. The median description length was 11 words in the Pilot Dataset and 15 words in the Evaluation Dataset. In both datasets, promotional descriptions tended to be longer than neutral descriptions. The difference in median length between the two classes was eight words in the Pilot Dataset and nine words in the Evaluation Dataset.

Differences in dataset composition were also visible across OSM object categories. In the Pilot Dataset, promotional descriptions occurred frequently among objects classified as restaurants and amenities, whereas neutral descriptions were more frequently associated with amenities, buildings, and land-use features. In the Evaluation Dataset, promotional descriptions were common among healthcare facilities, law offices, and insurance offices, while neutral descriptions predominantly referred to buildings and infrastructure-related features.

The independent reliability assessment showed high agreement with the original reference annotations. For the first additional annotator, 270 of 300 classifications agreed

with the reference labels, corresponding to 90.0% agreement and a Cohen's κ of 0.801 (95% CI: 0.734–0.868). For the second additional annotator, agreement was observed for 285 of 300 classifications (95.0%), corresponding to a Cohen's κ of 0.900 (95% CI: 0.851–0.949). These results indicate that the distinction between promotional and neutral descriptions was reproducible to a high degree, while disagreement remained for a subset of cases.

4.2. Model Comparison

To evaluate the LLM-based classification approach, 35 model configurations were tested on the Balanced Evaluation Set. Each configuration was evaluated in three complete runs under identical inference settings. Classification performance varied widely across models, with mean F1-scores ranging from below 1% to 93.42%. In terms of mean F1-score, llama3.3:70b achieved the highest overall classification performance. Across the three runs, the model reached a mean precision of 95.03%, a recall of 91.87%, and an F1-score of 93.42%. Furthermore, a language-group analysis of the model showed higher classification performance for English than for non-English descriptions. For the 866 English descriptions, the model achieved a precision of 97.64%, a recall of 96.12%, and an F1-score of 96.88%. For the 634 non-English descriptions, precision was 88.94%, recall was 82.48%, and the F1-score was 85.59%.

Several smaller models achieved performance comparable to that of larger models. Model size alone did not consistently correspond to higher F1-scores for the classification task examined here. Table 4 summarizes the five model configurations with the highest mean F1-scores together with their precision, recall, and runtime. Complete results for all 35 configurations are provided in Appendix B (Table A1).

Table 4. Classification performance and runtime of the five highest-ranked model configurations.

Model	Precision (%)	Recall (%)	F1-Score (%)	Runtime (s)
llama3.3:70b	95.03	91.87	93.42	716
gemma4:31b	97.42	85.47	91.05	635
llama3.1:8b	85.56	96.40	90.66	374
mixtral:8x22b	93.84	87.33	90.47	3409
llama3:8b	87.31	93.60	90.35	299

The models also differed in their balance between precision and recall. Some configurations showed the positive class less often. This resulted in high precision but lower recall. For example, ministral-3:14b achieved a mean precision of 97.97% but a recall of only 64.40%. A similar precision–recall imbalance was observed for several larger models. The model llama3.1:405b achieved a precision of 98.97% and a recall of 76.67%, while mistral-large:123b reached 97.78% precision and 76.40% recall. Among these examples, deepseek-r1:70b showed the highest precision (99.78%), but identified only 59.47% of descriptions annotated as promotional. Runtime and classification performance varied across the 35 model configurations, as illustrated by the relationship between mean F1-score and dataset-level runtime in Figure 3.

For the complete evaluation set of 1500 descriptions, llama3:8b required an average of approximately 299 s per run, equivalent to approximately 0.20 s per description. The highest-ranked model, llama3.3:70b, required an average of 716 s per run, corresponding to approximately 0.48 s per description. By comparison, mixtral:8x22b required approximately 3409 s and gpt-oss:120b approximately 3565 s to process the same dataset.

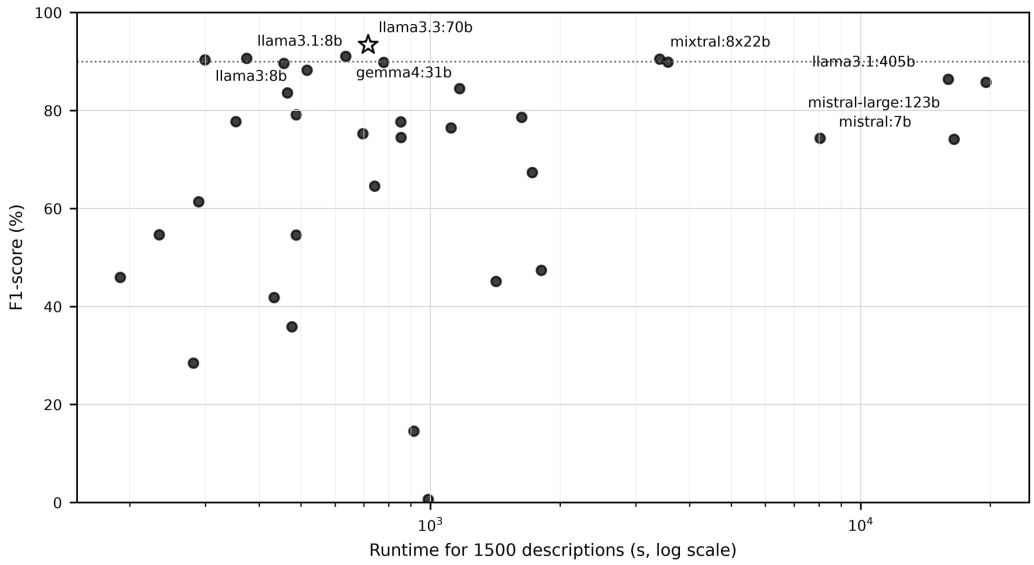


Figure 3. Mean F1-score and mean dataset-level runtime of the 35 evaluated LLM configurations on the Balanced Evaluation Set. Selected models are labeled for readability.

To provide a comparison with conventional text-classification approaches, the two additional baseline classifiers were evaluated on the same Balanced Evaluation Set. Table 5 summarizes their classification performance and inference runtime.

Table 5. Classification performance and inference runtime of the conventional text-classification baselines on the Balanced Evaluation Set.

Model	Precision (%)	Recall (%)	F1-Score (%)	Runtime (s)
TF-IDF (word) + Logistic Regression	72.41	87.47	79.23	0.01
TF-IDF (character) + Linear SVM	75.28	89.33	81.71	0.09

Both conventional classifiers processed the complete evaluation dataset several orders of magnitude faster than the LLM-based approaches, although their classification performance was lower. Word-level TF-IDF combined with Logistic Regression reached an F1-score of 79.23%, while character-level TF-IDF combined with a linear SVM performed somewhat better at 81.71%. The difference was particularly pronounced for precision, which reached 75.28% for the linear SVM compared with 95.03% for llama3.3:70b, whereas recall was more similar at 89.33% and 91.87%, respectively.

4.3. Characteristics of the Real-World Application

During the practical observation period from 1 October 2025 to 14 April 2026, a total of 16,156 newly added or modified description=* tags created by contributors with no more than three active mapping days were screened automatically. Of these, 2738 descriptions (16.9%) were classified by the deployed model as potentially promotional. These descriptions were associated with 2599 distinct changesets, each of which received a public changeset comment through the workflow. The commented changesets originated from 108 countries, and the associated descriptions represented 42 source languages. The United States accounted for the largest share of cases, representing 57% of all commented changesets. India, Poland, Germany, and Spain each accounted for approximately 4%. Overall, 67% of the associated descriptions were written in English.

Among OSM objects that could be assigned to a thematic category, restaurants and healthcare-related facilities were frequent, with dental practices accounting for a notable proportion of the latter group. At a broader level, the affected objects were concentrated primarily in the retail, service, and healthcare sectors. Approximately 15% of the objects could not be assigned to a thematic category using the rule-based classification described in Section 3.1. These nodes, ways, or relations contained no additional tags beyond description=* that permitted a more specific thematic assignment.

4.4. Real-World Evaluation of the Community Review Workflow

Two temporally separated observation periods were compared to assess subsequent modification patterns among entries classified as potentially promotional. The commented condition comprised cases identified during the real-world application period described in Section 4.3. During this period, changesets containing at least one description=* tag classified as potentially promotional received a public changeset comment through the workflow. During a subsequent comparison period, from 15 April to 15 July 2026, potentially promotional descriptions continued to be detected using the same classification procedure, but no workflow-generated public comments were posted.

Of the 2599 commented changesets reported in Section 4.3, 1306 were excluded because their contents had also been manually modified or reverted by the author of this study. Including these cases would have prevented the subsequent modification from being attributed observationally to either community activity or the workflow changeset comment alone. At the same time, because these exclusions were not random, the resulting commented sample may differ systematically from the full set of commented changesets. The final analytical sample comprised 1293 changesets in the commented condition and 1908 changesets in the non-commented comparison condition. Throughout the following analysis, each changeset is treated as one case, irrespective of the number of potentially promotional description=* tags contained within it.

A descriptive comparison of the two conditions showed similar distributions across several characteristics, alongside some differences in individual countries and languages. The United States accounted for 52% of cases in the commented condition and 54% in the non-commented condition, while the median description length was 25 and 27 words, respectively. Nodes represented the large majority of cases in both conditions (89% versus 86%). English was the most frequent description language in both groups, accounting for 71% of descriptions in the commented condition and 74% in the non-commented condition.

Within the predefined 28-day follow-up period, a modification event was observed for 1028 of 1293 cases (79.5%) in the commented condition, compared with 520 of 1908 cases (27.3%) in the non-commented condition. The unadjusted absolute risk difference was 52.3 percentage points (95% CI: 49.3–55.2), and the unadjusted risk ratio was 2.92 (95% CI: 2.70–3.16). Figure 4 shows the cumulative proportion of cases experiencing a first modification event during the 28-day follow-up period.

Differences between the two conditions emerged early in the follow-up period. In the commented condition, 720 of 1293 cases (55.7%) experienced a modification event within the first day. After seven days, the cumulative number had increased to 989 cases (76.5%). In the non-commented condition, the corresponding cumulative proportions were 12.8% after one day and 20.9% after seven days.

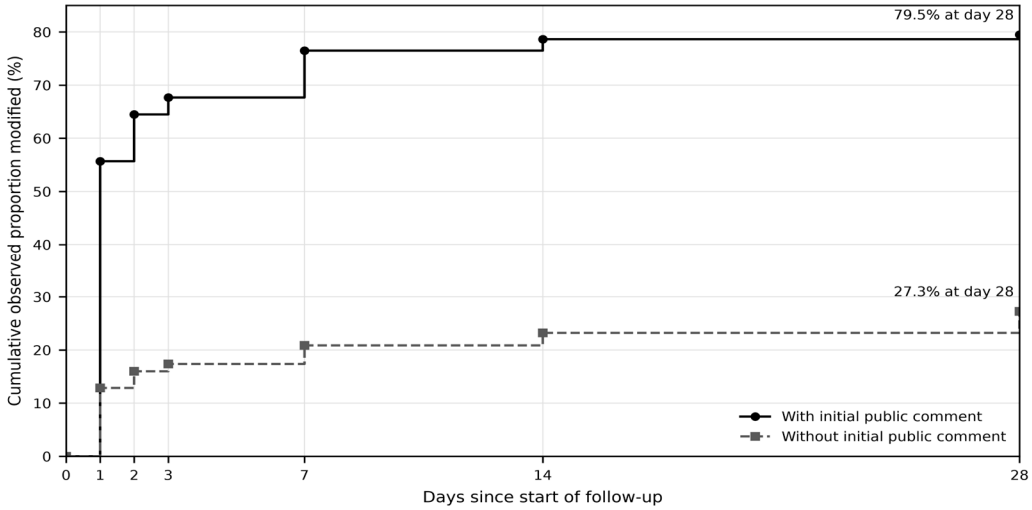


Figure 4. Cumulative proportion of cases with a first observed modification event during the 28-day follow-up period in the commented and non-commented conditions.

Among the 1028 modification events observed in the commented condition, 471 (45.8%) involved a change to the description=* value, 423 (41.1%) involved removal of the tag, and 134 (13.0%) involved deletion of the complete OSM object. Tag-level modifications (TAG_CHANGED and TAG_DELETED) accounted for 87.0% of all observed modification events. In the non-commented condition, 133 of 520 modifications (25.6%) were classified as TAG_CHANGED, 221 (42.5%) as TAG_DELETED, and 166 (31.9%) as OBJECT_DELETED. Complete object deletion was therefore more frequent as a proportion of observed modification events in the non-commented condition (Table 6).

Table 6. Distribution of modification types within the 28-day follow-up period.

Event Type	Commented, n	Share of Modifications (%)	Non-Commented, n	Share of Modifications (%)
TAG_CHANGED	471	45.8	133	25.6
TAG_DELETED	423	41.1	221	42.5
OBJECT_DELETED	134	13.0	166	31.9
Total	1028	100.0	520	100.0

Public changeset communication occurring after case detection was also examined. In the commented condition, 164 of the 1028 modified cases (16.0%) received at least one public changeset comment in addition to the initial workflow-generated changeset comment. In the non-commented condition, 136 of the 520 modified cases (26.2%) received at least one public changeset comment during follow-up. Because no workflow-generated initial comment was present in this condition, these comments represent independently occurring community communication rather than additional comments following an initial changeset comment.

In both conditions, subsequent public comments were most frequently observed for cases involving complete object deletion. Among OBJECT_DELETED cases, at least one such comment occurred in 38.8% of cases in the commented condition and 48.8% in the non-commented condition. Comments were less frequent for tag-level modifications, particularly for TAG_CHANGED events (Table 7).

Table 7. Public changeset comments observed during follow-up, by modification type.

Event Type	Condition	Modified Cases (n)	Cases with Comments (n)	Share (%)
TAG_CHANGED	Commented	471	40	8.5
	Non-commented	133	5	3.8
TAG_DELETED	Commented	423	72	17.0
	Non-commented	221	50	22.6
OBJECT_DELETED	Commented	134	52	38.8
	Non-commented	166	81	48.8
Total	Commented	1028	164	16.0
Total	Non-commented	520	136	26.2

Across all observed cases, irrespective of whether a modification occurred, at least one subsequent public comment was recorded for 164 of 1293 cases (12.7%) in the commented condition and 136 of 1908 cases (7.1%) in the non-commented condition. Among modified cases, the proportions were 16.0% and 26.2%, respectively. These conditional proportions should be interpreted descriptively because the modification rates differed substantially between the two conditions.

5. Discussion

5.1. Interpretation of Model Evaluation and Performance

The reference labels were initially assigned by a single researcher and therefore remain subject to a degree of annotation uncertainty. Distinguishing neutral from promotional content can be difficult when descriptions follow culturally specific writing conventions, require translation, or lie close to the boundary between the two classes. To assess the robustness of the reference annotation, two additional annotators independently classified separately drawn random samples. The results provide evidence that the applied classification criteria were reproducible across annotators. Nevertheless, the remaining disagreements indicate that some descriptions may permit more than one plausible interpretation, consistent with previous research on annotator disagreement [48]. The manually assigned reference labels should therefore still be regarded as reference judgments rather than error-free ground truth.

In a community-facing OSM workflow, false positives and false negatives have different consequences. False positives are more problematic because they may trigger unnecessary public interactions with contributors. False negatives, by contrast, leave potentially promotional descriptions undetected but do not themselves generate additional community interactions. Although F1-score was used as the primary deployment criterion to balance both error types, a more precision-oriented model could therefore be preferable in applications where minimizing unnecessary contributor interactions is the dominant objective. For example, gemma4:31b achieved a higher precision than llama3.3:70b (97.42% versus 95.03%), but at the cost of lower recall (85.47% versus 91.87%). The appropriate model choice therefore depends on the relative operational costs of false positives and false negatives rather than on F1-score alone.

The results also show that high classification performance was not restricted to the largest models. Both llama3.1:8b and llama3:8b achieved mean F1-scores above 90% while requiring less inference time than several larger configurations. This observation is consistent with recent research showing that the suitability of language models for text classification depends not only on model scale but also on factors such as prompt design, task characteristics, and computational efficiency [49]. Research on few-shot open-set classification likewise

indicates that appropriately designed classification approaches can enable substantially smaller models to approach the performance of much larger models in specific settings [50].

The language-specific analysis also revealed a notable performance difference between English and non-English descriptions. This finding indicates that the strong overall performance should not be assumed to apply uniformly across languages. Because the non-English subset combines descriptions from many languages with substantially smaller individual sample sizes, the present analysis does not establish whether the observed difference is concentrated in particular languages. Larger language-specific samples would be required for a more detailed assessment.

The comparison with conventional text-classification approaches further showed that the observed performance was not limited to differences among LLM configurations. Both TF-IDF baselines achieved comparatively high recall but substantially lower precision and F1-scores than the best-performing LLM. The difference was pronounced for precision. This distinction is relevant for the community-facing workflow, where false positives may result in unnecessary interactions with contributors. At the same time, these results do not establish that LLMs are universally superior to conventional text-classification methods, as only two representative baseline approaches were evaluated and their performance may depend on the available training data and feature configuration.

Overall, the model evaluation provides clear answers to RQ1 and RQ2. With respect to RQ1, current open-weight LLMs achieved high classification performance for promotional or SEO-oriented content in OSM and, in the present benchmark, performed better than the evaluated conventional baselines, although performance was lower for non-English descriptions. Regarding RQ2, the results indicate that no single model was optimal across all evaluation criteria. llama3.3:70b provided the strongest overall balance, while higher-precision or smaller and faster models may be preferable depending on operational priorities. The observed variation in performance was not systematically related to model size, and the present study was not designed to isolate its underlying causes; possible factors include differences in model architecture, training, instruction tuning, multilingual capabilities, and interactions with the fixed prompt. These findings suggest that future research could investigate task-specific fine-tuning of smaller models using multilingual OSM datasets and the incorporation of additional contextual information, such as surrounding OSM tags. Whether such adaptations improve classification performance remains to be established experimentally.

5.2. Significance of the Community Review Workflow

With respect to RQ3, cases receiving a public changeset comment were modified more frequently and earlier during the 28-day follow-up than cases in the non-commented comparison period. Because assignment to the two conditions was determined by calendar period rather than randomization, this difference cannot be interpreted as a causal effect of the changeset comment. A descriptive comparison showed that the two conditions were similar in several observed characteristics, including the dominant country, description language, OSM object type, and median description length, although differences remained for individual countries and languages. Nevertheless, differences in contributor composition, seasonal mapping activity, or the overall level of community review may have contributed to the observed modification rates. More general temporal changes in the prevalence or characteristics of promotional contributions may also have affected the observed modification rates. In addition, because the commented period preceded the non-commented comparison period, previous workflow interactions may have influenced later contributor or community behavior. These factors cannot be separated from the effect of the public changeset comment in the present observational design. The observed difference should

therefore be interpreted as an association between the presence of a workflow-generated comment and subsequent modification patterns. One possible explanation is that the public changeset comment increased the visibility of the affected entry and prompted contributors to inspect it.

The subsequent modifications did not usually involve extended public discussion. Among modified cases, an additional public changeset comment was recorded for 16.0% of cases in the commented condition and 26.2% in the non-commented condition. In some cases, contributors may therefore have acted on the initial comment without replying publicly; in others, the modification may have occurred independently. Inspection of individual case histories showed that changes were made both by original contributors and by other OSM participants. Where the original mapper revised the affected entry, the changeset comment may also have served as feedback on OSM-specific expectations concerning promotional content. However, because the actor responsible for each modification and the reason for the change were not recorded systematically, the relative importance of these different pathways cannot be quantified. A longitudinal study would be needed to determine whether this feedback has a lasting effect on later editing behavior.

5.3. Ethical Implications and Unintended Incentives

By leaving all decisions to OSM contributors, the workflow preserves human judgment. This is consistent with research on community-based machine learning systems in Wikipedia, which emphasizes contributor engagement, support for diverse workflows, and system trustworthiness [51]. This human review remains important even when classification performance is high. Automation bias may lead users to place too much confidence in automated recommendations, and explanations alone do not necessarily prevent this effect [52]. A positive LLM classification should be treated as a reason to inspect an entry, not as a verdict on the contribution or its author. In practice, reviewers still need to consider the original description together with the surrounding tags and the mapped object.

The deployment also produced a small number of objections to the use of the bot itself. Three negative comments were recorded during the practical application. The author, who also maintained the bot, acknowledged the concerns raised, manually reviewed each reported case, and communicated the findings in a respectful follow-up reply in the changeset discussion. None of these objections explicitly demonstrated a false-positive classification, and manual inspection of the disputed cases found promotional characteristics according to the criteria used in this study. No further responses from the contributors who had raised the objections were observed following these exchanges.

These cases suggest that acceptance of an automated workflow depends on more than classification accuracy. Contributors may object to the use of automated assessment or to the way in which such feedback is delivered, even when the underlying case appears to meet the classification criteria. With the small number of such comments, however, the study provides too little evidence to assess broader community acceptance.

The three-active-day eligibility threshold also needs to be interpreted carefully. Cases were restricted to contributors with no more than three active mapping days in order to keep the number of classifications, public interactions, and follow-up cases manageable. This threshold was a sampling and prioritization rule; it was not intended as an indicator that new contributors are more likely to submit promotional content. Because contributors with more extensive mapping histories were excluded from the real-world analysis, the observed patterns may not be representative of the full OSM contributor population and should not be generalized to more experienced contributors without further evaluation.

6. Conclusions

This study investigated whether current open-weight LLMs can reliably detect promotional or SEO-oriented content in OSM description=* tags and how such classification could be integrated into a community-based quality assurance workflow. In the controlled evaluation, performance varied considerably across models. The best-performing model, llama3.3:70b, reached a precision of 95.03%, a recall of 91.87%, and an F1-score of 93.42%. Classification errors nevertheless remain relevant when predictions may lead to direct interactions with OSM contributors. Although llama3.3:70b achieved the highest mean F1-score, several smaller models also exceeded 90% F1 while requiring less inference time. For deployment, model choice should therefore consider both classification performance and the expected workload. In the community workflow, entries receiving a public changeset comment were modified more often within the 28-day follow-up period than entries in the comparison condition (79.5% vs. 27.3%), and modifications generally occurred earlier.

The results indicate that LLM-based classification can support OSM quality assurance by identifying and prioritizing potentially problematic free-text content. It should be used for prioritization rather than automated decision-making. Human review remains important because false classifications can lead to unnecessary interactions with contributors. Future work should extend the language-specific evaluation using larger samples for individual languages and investigate the factors underlying performance differences between model configurations. The community workflow should also be evaluated with more comparable observation groups and, where feasible, a randomized study design.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were not required for this study. The study evaluated a public, community-based quality-assurance workflow in OpenStreetMap. The unit of analysis was the OSM data object and its description=* tag, not the individual contributor. No individuals were recruited as research participants, and no individual-level responses or behavior were analyzed. Subsequent modifications to the analyzed OSM data could have been made by any member of the OSM community and were assessed solely at the data-object level.

Informed Consent Statement: Not applicable. The study analyzed publicly available OSM contributions and public changeset activity. No individuals were recruited as research participants, and no individual contributors are identified in the manuscript.

Data Availability Statement: The underlying OpenStreetMap data used in this study are publicly available through the OSM project and its associated data distribution services. Derived datasets supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the author used ChatGPT (OpenAI, GPT-5.6 Sol) for linguistic and stylistic revision of author-written text, including improvements to clarity, grammar, and academic phrasing. The tool was not used to generate or analyze the study data. The author reviewed and edited all AI-assisted output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A

Listing A1. System prompt used in the experiments.

```
You are a multilingual expert specialized in detecting SEO spam in OpenStreetMap
(OSM) edits.
You understand natural language in many languages and can identify commercial,
promotional, or unrelated content that does not belong in OSM descriptions.
```

Listing A1. *Cont.*

```
TASK:
Classify a given text as SEO spam or not.
Output exactly one word: "True" or "False".
No explanations.

RETURN "True" if ANY of the following are present:
• Promotional/marketing intent (e.g., "best", "cheap", "No.1", "top-rated",
"official", "limited time", "discount %", "free consultation").
• Repeated brand/product mentions or keyword stuffing.
• Spammy patterns: URLs/domains, link shorteners, social/contact handles
(WhatsApp/Telegram/Line/Kakao/WeChat/Instagram), email addresses, phone numbers
with call-to-action, coupon codes, price lists, emojis used as marketing,
repeated hashtags, UTM/tracking params, or obfuscated text (e.g., R@r¡ble,
c@sino).
• High-risk verticals: crypto/NFT, gambling/casino/betting, escort/adult
services, payday loans, locksmith bailouts, pharma/supplements, fake tech
support, ticketing scalps.
• Off-topic, unrelated to the mapped feature (generic promotional paragraph,
city-wide ad text).
• Multi-location or region-stuffed phrases intended for search visibility rather
than identification of one place (e.g., multiple cities, departments, regions in
one phrase).

RETURN "False" if:
• The text is a normal place name or neutral/concise description of the mapped
feature.
• The text is too short (1--2 words) **and** contains no spammy pattern.
• The text is mostly hashtags/emojis with no promotional or commercial meaning
and no other spammy pattern.
• You are unsure.

Ambiguity rule:
Default to "False" unless a "True" condition clearly matches.

STRICT OUTPUT:
Only "True" or "False".
```

Listing A2. User Prompt and few-shot examples.

```
You will see examples of text followed by the correct label.
After the examples, classify the FINAL text. Output only "True" or "False".

Examples:
Text: "Best locksmith in Madrid---24/7 emergency service, call +34 600 123 123,
low prices!"
Answer: True

Text: "Parque del Retiro"
Answer: False

Text: "Café artesanal con bollería. Abre 8:00--17:00."
Answer: False

Text: "Buy crypto here --- official site: https://short.ly/nft?ref=deal"
Answer: True
```

Listing A2. *Cont.*

```
Text: "The phrase ‘We Are All Rarible CEO’ empowering artists via token
governance on Rarible and TokenSmart."
Answer: True

Text: "Historic church from the 17th century; baroque façade."
Answer: False

Text: "Cheap tickets! -50% TODAY ONLY!! Visit www dot tickets-fast dot com"
Answer: True

Text: "🔧 Taller mecánico. Cambios de aceite y frenos."
Answer: False

Text: "Call now +49 170 5555555 | WhatsApp: @quickunlock | #Berlin
#Schluesseldienst #cheap"
Answer: True

Text: "Аптека. Круглосуточно."
Answer: False

Text: "Online casino España | Bono 200% | Juega ya ▶️"
Answer: True

Text: "Museum ticket desk"
Answer: False

Text: "Escort VIP---book via Telegram @citystarsvip"
Answer: True

Text: "薬局(処方箋受付)"
Answer: False

Text: "Loan approval in minutes---apply now! 3.9% APR*"
Answer: True

Text: "Restaurante familiar de comida casera."
Answer: False

Text: "Best dentist in Paris | Invisalign | Free consultation | contact:
smile@clinic.fr"
Answer: True

Text: "R@r¡ble NFT marketplace --- earn, stake, vote"
Answer: True

Text: "Calle Mayor, 12"
Answer: False

Text: "It is a sandy, quite horrible, but playable football field. Occasionally
used by clubs to train (for cheap)"
Answer: False

Now classify the FINAL text:
```

Appendix B

Table A1 reports the complete model comparison on the Balanced Evaluation Set. The dataset contained 1500 manually annotated 'description=*' tags, consisting of 750 promotional descriptions and 750 neutral descriptions. Unless otherwise noted, values are averaged across all complete runs recorded in the evaluation logs. Accuracy, precision, recall, and F1-score are reported in percentages. Runtime is reported as the mean duration in seconds for classifying all 1500 entries. The columns TP, TN, FP and FN report the mean confusion-matrix values across valid runs. Models are sorted by mean F1-score in descending order.

Table A1. (a) Classification performance. (b) Runtime and confusion-matrix values.

Rank	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	llama3.3:70b	93.53	95.03	91.87	93.42
2	gemma4:31b	91.60	97.42	85.47	91.05
3	llama3.1:8b	90.07	85.56	96.40	90.66
4	mixtral:8x22b	90.80	93.84	87.33	90.47
5	llama3:8b	90.00	87.31	93.60	90.35
6	gpt-oss:120b	90.53	96.34	84.27	89.90
7	llama3.1:70b	90.67	98.88	82.27	89.81
8	glm-4.7-flash:bf16	89.45	88.01	91.33	89.64
9	gemma4:26b	89.31	98.00	80.27	88.25
10	llama3.1:405b	87.93	98.97	76.67	86.40
11	mistral-large:123b	87.33	97.78	76.40	85.78
12	deepseek-v2:236b	85.95	94.40	76.44	84.48
13	gemma4:e4b	85.47	95.86	74.13	83.61
14	qwen2.5:32b	82.47	97.65	66.53	79.14
15	qwen3.5:122b	82.27	98.99	65.20	78.62
16	ministral-3:14b	81.53	97.97	64.40	77.72
17	qwen2.5:72b	81.60	98.57	64.13	77.71
18	deepseek-v2.5:236b	80.80	98.73	62.40	76.47
19	llama3:70b	80.00	98.49	60.93	75.29
20	deepseek-r1:70b	79.67	99.78	59.47	74.52
21	gpt-oss:20b	79.15	96.59	60.44	74.36
22	mistral:7b	77.24	85.85	65.24	74.14
23	llama2:70b	53.80	52.08	95.38	67.37
24	llama4:16x17b	73.73	99.17	47.87	64.57
25	llama3.2:1b	45.47	47.51	86.67	61.38
26	deepseek-v2:16b	43.67	45.73	67.87	54.64
27	qwen3:32b	68.60	98.61	37.73	54.58
28	deepseek-r1:8b	65.40	98.73	31.20	47.41
29	llama2:7b	49.29	49.19	43.11	45.95
30	mistral-small:22b	64.33	97.78	29.33	45.13
31	gemma4:e2b	62.98	97.56	26.62	41.83
32	deepseek-r1:32b	60.87	99.39	21.87	35.85
33	llama3.2:3b	58.07	96.90	16.67	28.44
34	llama2:13b	33.02	20.11	11.42	14.57
35	phi3.5:3.8b	50.11	76.67	0.31	0.61

Rank	Model	Duration (s)	s/1k Tags	TP	TN	FP	FN
1	llama3.3:70b	716	477	689	714	36	61
2	gemma4:31b	635	423	641	733	17	109
3	llama3.1:8b	374	250	723	628	122	27
4	mixtral:8x22b	3409	2273	655	707	43	95
5	llama3:8b	299	200	702	648	102	48
6	gpt-oss:120b	3565	2376	632	726	24	118
7	llama3.1:70b	778	519	617	743	7	133

Table A1. *Cont.*

Rank	Model	Duration (s)	s/1k Tags	TP	TN	FP	FN
8	glm-4.7-flash:bf16	456	304	685	656.7	93.3	65
9	gemma4:26b	516	344	602	737.7	12.3	148
10	llama3.1:405b	15,970	10,646	575	744	6	175
11	mistral-large:123b	19,540	13,026	573	737	13	177
12	deepseek-v2:236b	1169	779	573.3	716	34	176.7
13	gemma4:e4b	465	310	556	726	24	194
14	qwen2.5:32b	487	324	499	738	12	251
15	qwen3.5:122b	1629	1086	489	745	5	261
16	ministral-3:14b	353	235	483	740	10	267
17	qwen2.5:72b	853	568	481	743	7	269
18	deepseek-v2.5:236b	1116	744	468	744	6	282
19	llama3:70b	696	464	457	743	7	293
20	deepseek-r1:70b	854	569	446	749	1	304
21	gpt-oss:20b	8042	5361	453.3	734	16	296.7
22	mistral:7b	16,473	10,982	489.3	669.3	80.7	260.7
23	llama2:70b	1724	1149	715.3	91.7	658.3	34.7
24	llama4:16x17b	741	494	359	747	3	391
25	llama3.2:1b	289	192	650	32	718	100
26	deepseek-v2:16b	234	156	509	146	604	241
27	qwen3:32b	487	325	283	746	4	467
28	deepseek-r1:8b	1810	1207	234	747	3	516
29	llama2:7b	190	127	323.3	416	334	426.7
30	mistral-small:22b	1421	947	220	745	5	530
31	gemma4:e2b	433	288	199.7	745	5	550.3
32	deepseek-r1:32b	476	317	164	749	1	586
33	llama3.2:3b	281	187	125	746	4	625
34	llama2:13b	915	610	85.7	409.7	340.3	664.3
35	phi3.5:3.8b	987	658	2.3	749.3	0.7	747.7

References

1. Chen, C.; Jiang, Y.; Feng, Y.; Liu, Y.; Meng, L. Explaining the Urban Mechanisms behind OpenStreetMap Updates: Evidence from 88 German Cities. *Comput. Environ. Urban Syst.* **2026**, *128*, 102456. [CrossRef]
2. Goodchild, M.F. Citizens as Sensors: The World of Volunteered Geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]
3. Haklay, M.; Weber, P. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
4. Sarkar, D.; Anderson, J.T. Corporate Editors in OpenStreetMap: Investigating Co-editing Patterns. *Trans. GIS* **2022**, *26*, 1879–1897. [CrossRef]
5. Anderson, J.; Sarkar, D.; Palen, L. Corporate Editors in the Evolving Landscape of OpenStreetMap. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 232. [CrossRef]
6. OpenStreetMap Wiki-Organised Editing/Activities. Available online: https://wiki.openstreetmap.org/wiki/Organised_Editing/Activities (accessed on 15 August 2026).
7. OpenStreetMap Wiki-Vandalism. Available online: https://wiki.openstreetmap.org/wiki/Vandalism (accessed on 15 August 2026).
8. OpenStreetMap Wiki-Key: Description. Available online: https://wiki.openstreetmap.org/wiki/Key:description (accessed on 15 August 2026).
9. Sun, Y.; Meng, L.; Camero, A.; Auer, S.; Zhu, X.X. A Deep Dive into OpenStreetMap Research since Its Inception (2008–2024): Contributors, Topics, and Future Trends. *Int. J. Geogr. Inf. Sci.* **2026**, *40*, 1–55. [CrossRef]
10. Haklay, M. How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environ. Plan. B Plan. Des.* **2010**, *37*, 682–703. [CrossRef]
11. Girres, J.; Touya, G. Quality Assessment of the French OpenStreetMap Dataset. *Trans. GIS* **2010**, *14*, 435–459. [CrossRef]
12. Neis, P.; Zielstra, D.; Zipf, A. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet* **2011**, *4*, 1–21. [CrossRef]
13. Fan, H.; Zipf, A.; Fu, Q.; Neis, P. Quality Assessment for Building Footprints Data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 700–719. [CrossRef]
14. Brovelli, M.A.; Zamboni, G. A New Method for the Assessment of Spatial Accuracy and Completeness of OpenStreetMap Building Footprints. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 289. [CrossRef]

15. Barron, C.; Neis, P.; Zipf, A. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Trans. GIS* **2014**, *18*, 877–895. [CrossRef]
16. Neis, P.; Goetz, M.; Zipf, A. Towards Automatic Vandalism Detection in OpenStreetMap. *ISPRS Int. J. Geo-Inf.* **2012**, *1*, 315–332. [CrossRef]
17. Juhász, L.; Novack, T.; Hochmair, H.; Qiao, S. Cartographic Vandalism in the Era of Location-Based Games—The Case of OpenStreetMap and Pokémon GO. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 197. [CrossRef]
18. Tempelmeier, N.; Demidova, E. Attention-Based Vandalism Detection in OpenStreetMap. In *Proceedings of the ACM Web Conference 2022, Virtual Event, Lyon, France, 25–29 April 2022*; ACM: New York, NY, USA, 2022; pp. 643–651.
19. Paul, H.; Nikolaev, A. Fake Review Detection on Online E-Commerce Platforms: A Systematic Literature Review. *Data Min. Knowl. Discov.* **2021**, *35*, 1830–1881. [CrossRef]
20. Huang, D.Y.; Grundman, D.; Thomas, K.; Kumar, A.; Bursztein, E.; Levchenko, K.; Snoeren, A.C. Pinning Down Abuse on Google Maps. In *Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017*; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2017; pp. 1471–1479.
21. Koswatte, S.; McDougall, K.; Liu, X. VGI and Crowdsourced Data Credibility Analysis Using Spam Email Detection Techniques. *Int. J. Digit. Earth* **2018**, *11*, 520–532. [CrossRef]
22. Potthast, M.; Stein, B.; Gerling, R. Automatic Vandalism Detection in Wikipedia. In *Advances in Information Retrieval*; Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; Volume 4956, pp. 663–668, ISBN 978-3-540-78645-0.
23. Adler, B.T.; De Alfaro, L.; Mola-Velasco, S.M.; Rosso, P.; West, A.G. Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In *Computational Linguistics and Intelligent Text Processing*; Gelbukh, A., Ed.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6609, pp. 277–288, ISBN 978-3-642-19436-8.
24. Green, T.; Spezzano, F. Spam Users Identification in Wikipedia Via Editing Behavior. *Proc. Int. AAAI Conf. Web Soc. Media* **2017**, *11*, 532–535. [CrossRef]
25. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150. [CrossRef]
26. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A Survey on Text Classification: From Traditional to Deep Learning. *ACM Trans. Intell. Syst. Technol.* **2022**, *13*, 1–41. [CrossRef]
27. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics; Volume 2, Short Papers. *arXiv* **2016**, arXiv:1607.01759.
28. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North, Minneapolis, MN, USA, 2–7 June 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.
29. Wang, S.; Hu, T.; Xiao, H.; Li, Y.; Zhang, C.; Ning, H.; Zhu, R.; Li, Z.; Ye, X. GPT, Large Language Models (LLMs) and Generative Artificial Intelligence (GAI) Models in Geospatial Science: A Systematic Review. *Int. J. Digit. Earth* **2024**, *17*, 2353122. [CrossRef]
30. Hojati, M.; Feick, R. Large Language Models: Testing Their Capabilities to Understand and Explain Spatial Concepts (Short Paper). In *Leibniz International Proceedings in Informatics (LIPIcs)*; Dagstuhl Publishing: Wadern, Germany, 2024; Volume 315, pp. 31:1–31:9. [CrossRef]
31. Xu, L.; Zhao, S.; Lin, Q.; Chen, L.; Luo, Q.; Wu, S.; Ye, X.; Feng, H.; Du, Z. Evaluating Large Language Models on Geospatial Tasks: A Multiple Geospatial Task Benchmarking Study. *Int. J. Digit. Earth* **2025**, *18*, 2480268. [CrossRef]
32. Zhang, Y.; Wei, C.; Wu, S.; He, Z.; Yu, W. GeoGPT: Understanding and Processing Geospatial Tasks through An Autonomous GPT. *arXiv* **2023**, arXiv:2307.07930.
33. Zhang, Y.; Wang, Z.; He, Z.; Li, J.; Mai, G.; Lin, J.; Wei, C.; Yu, W. BB-GeoGPT: A Framework for Learning a Large Language Model for Geographic Information Science. *Inf. Process. Manag.* **2024**, *61*, 103808. [CrossRef]
34. Zhang, Y.; Li, J.; Wang, Z.; He, Z.; Guan, Q.; Lin, J.; Yu, W. Geospatial Large Language Model Trained with a Simulated Environment for Generating Tool-Use Chains Autonomously. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *136*, 104312. [CrossRef]
35. Mansourian, A.; Oucheikh, R. ChatGeoAI: Enabling Geospatial Analysis for Public through Natural Language, with Large Language Models. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 348. [CrossRef]
36. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
37. Zhuo, J.; Zhang, S.; Fang, X.; Duan, H.; Lin, D.; Chen, K. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. *arXiv* **2024**, arXiv:2410.12405.
38. Planet OSM-Complete OSM Data History. Available online: https://planet.openstreetmap.org/planet/full-history/ (accessed on 15 August 2026).
39. Planet OSM-Replication Diffs. Available online: https://planet.openstreetmap.org/replication/minute/ (accessed on 15 August 2026).

40. OpenStreetMap Foundation-Active Contributor Membership. Available online: https://supporting.openstreetmap.org/active-contributor-membership/ (accessed on 15 August 2026).
41. Ollama-Build with Open Models, on Your Computer and in the Cloud. Available online: https://ollama.com/ (accessed on 15 August 2026).
42. OpenStreetMap Wiki-Nominatim. Available online: https://wiki.openstreetmap.org/wiki/Nominatim (accessed on 15 August 2026).
43. OpenStreetMap Wiki-Map Features. Available online: https://wiki.openstreetmap.org/wiki/Map_features (accessed on 15 August 2026).
44. Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
45. OpenStreetMap Wiki-Automated Edits Code of Conduct. Available online: https://wiki.openstreetmap.org/wiki/Automated_Edits_code_of_conduct (accessed on 2 November 2026).
46. NeisBot | OpenStreetMap. Available online: https://www.openstreetmap.org/user/NeisBot (accessed on 15 August 2026).
47. OpenStreetMap Foundation-Introducing Changeset Discussions. Available online: https://blog.openstreetmap.org/2014/11/02/introducing-changeset-discussions/ (accessed on 15 August 2026).
48. Xu, J.; Theune, M.; Braun, D. Leveraging Annotator Disagreement for Text Classification. *arXiv* **2024**, arXiv:2409.17577.
49. Li, L.; Sleem, L.; Gentile, N.; Nichil, G.; State, R. Small Language Models in the Real World: Insights from Industrial Text Classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), Vienna, Austria, 28–30 July 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 971–982.
50. Singh, A.K.; Yannakoudakis, H. Few-Shot Open-Set Classification via Reasoning-Aware Decomposition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 13865–13886.
51. Smith, C.E.; Yu, B.; Srivastava, A.; Halfaker, A.; Terveen, L.; Zhu, H. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020*; ACM: New York, NY, USA, 2020; pp. 1–14.
52. Vered, M.; Livni, T.; Howe, P.D.L.; Miller, T.; Sonenberg, L. The Effects of Explanations on Automation Bias. *Artif. Intell.* **2023**, *322*, 103952. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.