Issue Detection and Category Classification in Domain-Specific Technical Logbooks

Afshin Karimi[1,2], Ingmar Hartl[1], Henrik Tünnermann[1], Anne Lauscher[2]
[1]DESY, Hamburg, Germany
[2]Trustworthy AI Lab, University of Hamburg, Germany

Abstract

Operating large-scale research infrastructures such as free-electron lasers produces vast amounts of operator-authored documentation that records daily observations, anomalies, and maintenance actions. These logbooks and incident reports contain valuable operational knowledge but often remain underexplored due to their unstructured, domain-specific language. While large language models (LLMs) show strong generalization in general domains, their effectiveness on such technical operator text has, to the best of our knowledge, not been systematically assessed. We introduce two new datasets from real-world laser operations: (i) a **logbook dataset** annotated for binary *issue detection* (does an entry describe or report an actionable fault?), and (ii) an **operator ticket dataset** annotated for multi-class *issue categorization* (assign each ticket to one of 13 technical categories). The corpora comprise **2,979** logbook entries and **758** tickets from 2022–2024; both are cleaned, anonymized, and suitable for benchmarking classification performance. For the logbook dataset, only textual content is retained, and embedded images or screenshots originally attached to entries are excluded, resulting in a text-only benchmark for issue detection. We evaluate four open LLMs (LLaMA-3, Mistral-Small, Qwen-3-30B, GPT-OSS-120B) under zero-shot, few-shot, and chain-of-thought (CoT) prompting, using multiple semantically equivalent *prompt variants* per setting to assess robustness. Across both tasks, **few-shot prompting** is consistently strongest, with top systems reaching **F1** ≈ 0.84 for logbook issue detection and **Macro-F1** ≈ 0.42 for operator ticket categorization. These results suggest that incorporating a handful of in-domain examples can substantially improve performance on operator-authored technical text, even without fine-tuning. Code and evaluation scripts are publicly available at https://github.com/TAI-HAMBURG/llms-for-xfel-operations. Links to the released datasets are provided in the repository.

Keywords: Issue Detection, Issue Categorization, Operator-authored Logbooks, Large Language Models, Prompting Strategies

1. Introduction

Recent advances in natural language processing (NLP) and large language models (LLMs) have delivered strong gains in text understanding and classification across many domains (Devlin et al., 2019; Brown et al., 2020; Touvron et al., 2023). For instance, these advances now support automation in product reviews (Zhang et al., 2015), news reporting (Vajjala and Shimangaud, 2025), and social media monitoring (Münker et al., 2024). Nevertheless, uptake in scientific and industrial contexts remains limited, especially for operational documentation and technical reporting (Sulc et al., 2024).

In large-scale research infrastructures such as free-electron lasers, particle accelerators, and high-power laser systems, operators continuously record observations, anomalies, and maintenance actions. At the European XFEL and FLASH free-electron laser facilities at DESY (Decking et al., 2020; Rossbach et al., 2019), the documentation of laser operations appears in two complementary forms: "logbook entries" and structured "operator tickets." These records form an important part of the facility's institutional memory, supporting diagnostics, knowledge transfer, and system uptime. However, because the text mixes natural language, technical jargon, and occasional structure, automated analysis is difficult. Manual inspection is labor-intensive, subjective, and inconsistent, which can delay problem resolution and system recovery.

While extensive research exists on text classification in general domains (Bucher and Martini, 2024; Edwards and Camacho-Collados, 2024; Leitner and Rehm, 2025) , general-purpose models often fail to generalize to highly specialized operational contexts. Moreover, most studies on the related domain of log analysis target *machine-generated* data, such as structured IT logs (Guan et al., 2024; Zhong et al., 2024), rather than human-authored reports. Consequently, there remains a clear gap in resources and benchmarks for automated issue detection and categorization in operator-authored technical documentation. To date, no publicly available dataset captures the linguistic and contextual characteristics of operator-generated records in scientific laser systems.

In this paper, we address this gap by introducing two new datasets derived from real-world operator documentation at a large-scale research infrastructure: (i) a **logbook dataset** annotated for binary *issue detection* (issue vs. non-issue), and (ii)

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5996–6008
11–16 May 2026.

an **operator ticket dataset** annotated for multi-class *issue categorization*. Both datasets have been carefully cleaned, anonymized, and prepared for systematic evaluation. We then benchmark several **competitive contemporary LLMs** using three prompting strategies—*zero-shot*, *few-shot*, and *chain-of-thought (CoT)*—to assess their ability to perform accurate and interpretable classification without fine-tuning. Our analysis reveals how LLMs adapt to domain-specific technical text and how reasoning-based prompts influence performance and consistency in classification.

Contributions. This work makes four primary contributions. First, we introduce two domain-specific datasets derived from real laser system operations at the European XFEL and FLASH facilities at DESY: (i) a logbook dataset annotated for issue detection, and (ii) an operator ticket dataset annotated for multi-class issue categorization across 13 normalized technical categories. Both datasets consist of operator-authored narratives produced in a high-stakes operational environment and capture realistic terminology, subsystem interactions, and maintenance workflows. Second, we present a transparent data construction pipeline including expert annotation procedures and an anonymization process combining rule-based redaction with LLM-assisted masking to preserve privacy while retaining technical fidelity. The datasets and processing protocols are released to support reproducible research on domain-specific technical text. Third, we perform a systematic evaluation of open large language models (LLMs) using zero-shot, few-shot, and chain-of-thought prompting strategies, and compare their performance with supervised discriminative baselines through stratified cross-validation. This setup enables a principled comparison between in-context learning and classical text classification approaches in a specialized technical domain. Finally, we analyze model behavior through error analysis, category-level confusions, and boundary ambiguities, highlighting challenges in interpreting operator-authored technical documentation and outlining directions for domain-adaptive prompting and future multimodal extensions for scientific operations.

2. Related Work

Prompting vs. Fine-tuning for Text Classification. Recent work compares prompting-based approaches with fine-tuned models for text classification. Studies such as Leitner and Rehm (2025) and Edwards and Camacho-Collados (2024) report that prompting performs competitively on binary tasks but is less reliable for multi-class classification with closely related categories. Similarly, Bucher and Martini (2024), Botunac et al. (2024), and Wang et al. (2024) show that fine-tuned encoder models often achieve higher accuracy and robustness than zero- or few-shot prompting. However, these evaluations are mostly conducted on general-domain corpora such as reviews or social media. The effectiveness of prompting methods on highly specialized operator-authored technical documentation remains largely unexplored.

Prompting Strategies and Reasoning. Building on standard prompting, recent work explores methods to improve reasoning and stability in classification. Milios et al. (2023) and Münker et al. (2024) show that semantically relevant examples and clear label phrasing can enhance zero-shot performance, while Fechner and Dörpinghaus (2024) and Thaminkaew et al. (2024) demonstrate that label-aware templates aid few-shot generalization. Reasoning-oriented prompting such as *Chain-of-Thought (CoT)* (Wei et al., 2022) improves interpretability but yields mixed gains in categorical classification. We extend these insights by benchmarking zero-shot, few-shot, and reasoning-based prompting for operator-authored technical text.

From Machine Logs to Operator Text. Most prior work on log analysis focuses on structured, machine-generated logs. Recent approaches apply large language models to benchmarks such as HDFS and BGL for anomaly detection and log parsing (Guan et al., 2024; Zhong et al., 2024), while MaintNet (Akhbardeh et al., 2020) studies failure prediction from structured maintenance logs using learned event templates. In contrast, our work studies free-text operator-authored logbook entries describing operational events in natural language, which require contextual understanding rather than template-based log analysis.

Industrial Maintenance and Work-Order Text Classification. Prior work has explored the analysis of human-authored maintenance narratives in industrial environments. For example, Hodkiewicz et al. (Hodkiewicz et al., 2021) propose an NLP pipeline that converts unstructured maintenance work-order descriptions into structured representations for reliability analysis and decision support. Other studies apply machine learning to technician reports to extract operational insights and derive *key performance indicators (KPIs)* and failure categories (Lutz et al., 2023). However, these approaches mainly focus on information extraction in manufacturing or energy systems rather than reproducible supervised benchmarks for issue detection and classification across closely related technical categories. In contrast, our work introduces annotated datasets from scientific laser operations and

evaluates binary issue detection and multi-class categorization in a research setting.

LLMs in Scientific and Operational Domains. LLMs have recently been applied to scientific and operational data, primarily for retrieval and summarization. For example, Sulc et al. (2024) and Sulc et al. (2025) explore retrieval-augmented approaches for accelerator logbooks, while Yin et al. (2024) and Zhang et al. (2025) study domain adaptation and taxonomy enrichment. However, these works do not address issue classification on operator-authored documentation. In contrast, we introduce annotated datasets of operator-authored records and systematically evaluate LLMs for issue detection and categorization in real-world laser system operations.

3. Dataset

3.1. Dataset Construction

This section describes the two datasets used in our experiments: (1) an operator logbook dataset for binary *infrastructure issue detection*, and (2) an operator ticket dataset for multi-class *issue categorization*. Both originate from real documentation of laser system operations at a large research infrastructure. Construction involved data collection, cleaning, anonymization, and (where required) expert annotation to obtain gold labels for evaluation.

3.1.1. Logbook Dataset

Dataset Description. The logbook dataset is derived from a collection of distinct electronic logbooks maintained for several laser systems at a large-scale research infrastructure. These include logbooks for laser operations, diagnostic systems, and related facility components. Our dataset aggregates these parallel streams of operational data, providing a comprehensive and diverse view of facility activities. These logbooks serve as a real-time chronicle of system interactions, capturing the nuances of managing complex scientific instruments and documenting both technical and procedural aspects of facility operations.

Each data point (an *entry*) is a laser-operator-authored note with a concise textual description, sometimes accompanied by screenshots or diagnostic plots. Content ranges from planned work (for example performance tuning or scheduled experiments) to reactive actions following anomalies. This mix of technical and narrative detail makes the dataset suitable for studying unstructured, domain-specific documentation.

Example Entries. To illustrate the content and style of operator-authored logbook records, below are two anonymized examples from the dataset:

> **Example 1 (Issue).** "The beam is moved for vertical direction after [PROJECT_ID]—not known the reason. After [PROJECT_ID] the beam is moved down somehow. It was fixed manually to the previous position."

> **Example 2 (Non-issue).** "Got the Oscil. PD signal on the MBI scope. Classic case of disconnected cable and wrong coupling (should be 50 Ω, not 1 MΩ). This entry was sent to the following experts: [EMAIL] [EMAIL]."

Cleaning. Following data extraction, several preprocessing steps were applied to ensure data completeness. The raw logbook collection contained 8,648 entries. Entries with empty textual content or those consisting solely of screenshots or diagnostic images without accompanying text were removed to maintain meaningful context for language-based analysis. After filtering, 2,979 entries (34.4%) containing usable textual descriptions were retained, while 5,669 entries (65.6%) were excluded because they contained only images or lacked sufficient textual information. Although these image-only records were not included in the current text-based benchmark, they remain valuable for future multimodal analysis. All data were subsequently anonymized using the unified redaction pipeline described in Section 3.2.

Annotation and Validation. We applied a two-step expert annotation procedure. First, subject-matter experts reviewed each entry and assigned a binary label: `issue` if the entry described a malfunction, error state, performance degradation, or an event requiring corrective action; `non-issue` for normal operation, planned maintenance, or successful adjustments. Annotators were English-speaking researchers with experience in laser diagnostics and operations.

Second, a separate expert independently validated a random subset of entries to assess consistency. Disagreements were resolved by discussion, which refined the labeling guidelines. This process improved overall label quality.

The resulting dataset serves as a benchmark for the task of automated infrastructure issue detection. Because it combines entries from multiple laser and subsystem logbooks, it captures realistic linguistic and operational diversity and mirrors the range of conditions seen in real facility operations.

Dataset	#Entries	Years	Median len.	Mean len.	Labels / Categories
Logbook	2,979	2022–2024	121	417	Binary (issue / non-issue)
Operator Tickets	758	2022–2024	872	1,914	13 categories

Table 1: Summary of the two datasets used in this study. Both span the period 2022–2024. "Len." denotes text length in characters, calculated from free-text fields after trimming extreme outliers at the 99th percentile.

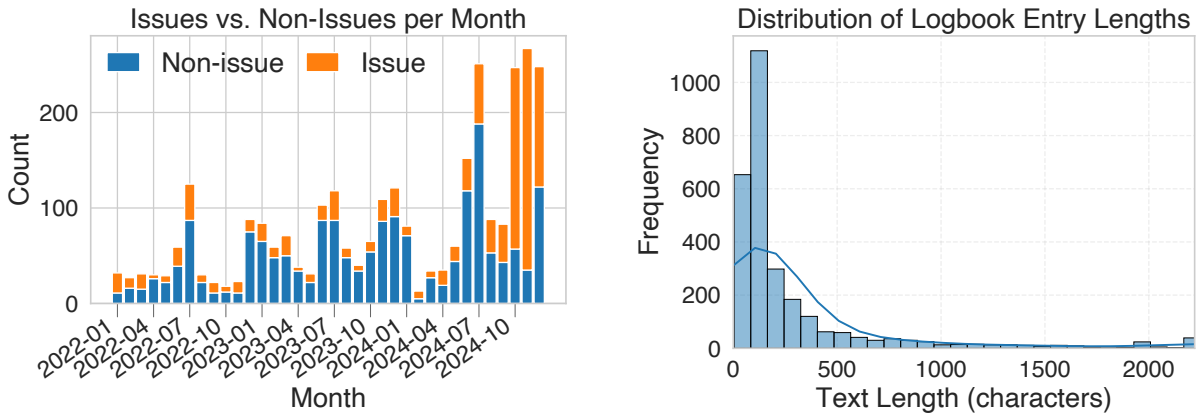


Figure 1: **Logbook dataset overview. Left:** Monthly counts of entries from 2022–2024, split by *issue* vs. *non-issue*. **Right:** Distribution of entry lengths (in characters), exhibiting a heavy right tail—most entries are short, while a small fraction contain lengthy technical descriptions.

3.1.2. Operator Ticket Dataset

The operator ticketing system is a structured platform for reporting detected issues. Tickets can be created either by e-mail or automatically by monitoring systems (e.g., automated alerts such as power glitches). Each ticket includes both automatically generated metadata (e.g., timestamps, affected system, status, and downtime) and manually entered fields (e.g., issue category and free-text description). The dataset therefore combines structured metadata with rich unstructured text, particularly within the *history* section that records discussions and troubleshooting steps. The issue category selected by the operator serves as the *gold label* for evaluation, so no additional manual annotation was required. The predefined categories follow the facility classification scheme (see Table 4) and support consistent training and evaluation for multi-class classification.

Collection. We collected all tickets created between 2022 and 2024 from the internal incident tracking system. For each ticket we extracted metadata and the full history logs. Metadata were stored in tabular form; conversation histories were stored in JSON to preserve chronology and nesting.

Example Tickets. To provide an impression of the ticket content and structure, below are two anonymized examples illustrating typical operator-authored reports:

> **Example 1 (Interlock System).** *Subject:* [Injektorlaserelog] Sluice door is stuck and interlock is broken. *History excerpt:* "I will turn on the [PROJECT_ID] and [PERSON] will work on [PROJECT_ID]. This entry was sent to the following experts: [PERSON], [PROJECT_ID]."

> **Example 2 (Software / Control System).** *Subject:* [Flash2_PPLaserelog] Remote connection to Amphos PC doesn't work. *History excerpt:* "The issue concerns a failed remote connection to the Amphos control PC. This entry was sent to the following experts: [PERSON]. Direct link to the corresponding logbook entry: [URL]."

Cleaning. After extraction, several preprocessing steps were applied to ensure data completeness and consistency. First, we retained only tickets containing a valid operator-assigned ground-truth issue category, ensuring reliable labels for evaluation. Tickets with missing or invalid category annotations were excluded. Second, we filtered out incomplete tickets lacking a discussion record, retaining only entries that contained at least one *history* section documenting operator interactions. To manage variable-length conversations, histories were truncated to a maximum of 12,000 characters while preserving the earliest and most infor-

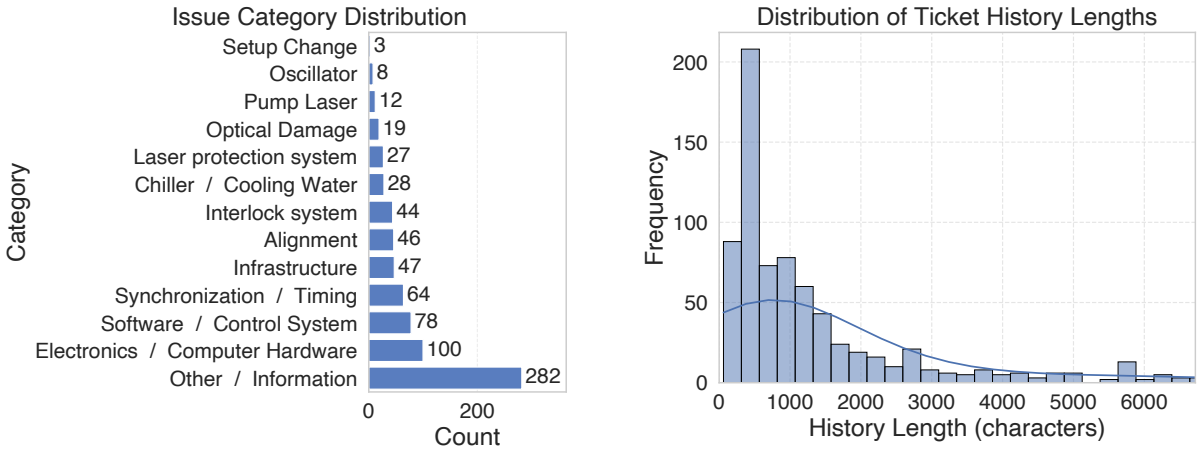


Figure 2: **Operator Ticket dataset. Left:** Distribution of normalized issue categories; the most frequent are *Other / Information*, *Electronics / Computer Hardware*, and *Software / Control System*, while several subsystem categories are comparatively rare. **Right:** Distribution of ticket history lengths (in characters) with a pronounced long tail—most histories are short, but a minority document extended troubleshooting.

mative messages. All tickets were subsequently anonymized using the unified redaction pipeline described in Section 3.2.

3.2. Data Anonymization

All logbook entries and operator tickets were anonymized prior to analysis and release to remove personal and facility-specific identifiers. We used a two-stage redaction pipeline that combines deterministic pattern matching with an LLM-based pass to increase coverage on unstructured text. First, we applied rule-based pre-masking using regular expressions and domain-specific patterns to identify and replace common identifier types, including email addresses, URLs, phone numbers, room/building references, ticket IDs, serial numbers, and known project/facility tokens. We additionally used conservative name heuristics (e.g., capitalized multi-token name patterns and title cues) to mask likely personal names. Second, we ran an LLM-based redaction step that rewrites the text in-place while preserving technical meaning, instructed to conservatively replace any remaining identifying mentions (names, affiliations, user handles, addresses, and project-specific identifiers). To avoid context overflows, long histories were processed in chunks, and a post-processing pass re-applied key regex rules to enforce consistent masking. All detected spans were replaced with standardized placeholder tokens such as `[PERSON]`, `[EMAIL]`, `[PHONE]`, `[URL]`, `[ADDRESS]`, `[TICKET_ID]`, `[SERIAL]`, and `[PROJECT_ID]`. In addition, timestamps were deterministically shifted by up to ± 15 days (preserving temporal order) to reduce re-identification risk while maintaining relative timelines. Finally, we performed manual spot-checking on a random sample from both datasets to validate that no directly identifying personal or facility-specific information remained after anonymization.

Label Reliability. Labels originate from multiple operators working across different shifts and expertise levels. As a result, variations in individual interpretation and documentation style naturally occur. Some tickets are brief or ambiguous, and in such cases operators may select the broad category *Other / Information* even when a more specific label would be appropriate. We did not adjudicate all tickets and did not compute inter-annotator agreement, which implies that a small degree of label noise is expected. Such inconsistencies can occasionally lead to confusions between semantically related categories, especially when similar subsystems or fault types are involved. The potential impact of this label variability is discussed in more detail in our experimental evaluation (Section 4).

On the "Other / Information" Category. Informational entries are common in operational ticketing systems because issue trackers are used not only for fault reporting but also for coordination, status updates, and contextual documentation among stakeholders. Such posts preserve a searchable record within the same infrastructure used for issue management and therefore accumulate alongside actionable incidents. The "Other / Information" category captures entries that do not clearly fit one of the predefined technical subsystems or mainly serve coordination purposes. Some reports are ambiguous, rare, or span multiple subsystems, making strict assignment to a single category difficult. Including this label ensures completeness of the taxonomy and avoids forced misclassification into

unrelated technical classes. Our analysis suggests that combining informational posts and residual issue types under a single label may obscure functional differences. Future work will explore hierarchical taxonomy refinements that separate purely informational entries from residual technical issues, potentially improving interpretability and classification performance.

3.3. Annotation Procedure and Quality Control

Logbook Dataset (Binary Task). Logbook entries were annotated by domain experts with experience in laser system operations. For the annotation task, annotators were instructed to label an *issue* as an entry describing a malfunction, performance degradation, error state, or event requiring corrective action. Entries documenting routine operation, planned maintenance, configuration updates, or informational coordination were labeled as *non-issue*. To assess reliability, a random subset of 100 entries was independently labeled by three annotators. Pairwise Cohen's κ shows a mean agreement of $\kappa = 0.65$ (std = 0.07; range: 0.58–0.74), corresponding to substantial agreement (McHugh, 2012). Compared to earlier pilot annotations, agreement improved after introducing clearer annotation instructions, indicating that explicit guidelines help reduce interpretation differences among annotators. Remaining variability stems from several domain-specific factors. First, logbook entries are often written as short sequential updates referring to the same incident. Because each entry is treated as an independent instance, annotators may differ on whether a brief follow-up message constitutes a separate issue or part of an ongoing one, introducing granularity ambiguity. Second, screenshots and diagnostic plots were excluded from the current text-only dataset; some entries rely on visual context to clearly signal a fault condition, reducing consistency in text-only annotation. Third, interpretation of borderline cases depends strongly on operational experience, and annotator calibration effects may still contribute to disagreement.

Operator Ticket Dataset (Multi-class Task). Ticket labels originate from trained operators assigning one of 13 predefined subsystem categories during incident reporting (Table 4). For reliability assessment, 100 randomly sampled tickets were independently re-labeled by four domain experts using the same category definitions. Pairwise Cohen's κ across annotators shows a mean agreement of $\kappa = 0.56$ (std = 0.07; range: 0.47–0.64), corresponding to moderate agreement (McHugh, 2012). As in the binary task, agreement increased after introducing clearer annotation guidelines for the re-labeling procedure. Disagreements mainly occur between semantically related subsystem categories (e.g., Electronics / Computer Hardware vs. Software / Control System), consistent with the confusion patterns in Section 4. Additionally, informational tickets without a clearly identifiable subsystem can lead to inconsistent interpretations among annotators. Overall, the agreement analysis highlights the inherent complexity of issue classification in operator-authored technical documentation.

Task	Subset	κ (mean $\pm$ std)
Logbook (Binary)	100	0.65 ± 0.07
Tickets (13 classes)	100	0.56 ± 0.07

Table 2: Inter-annotator agreement (pairwise Cohen's κ) on independently re-labeled subsets.

3.4. Dataset Statistics and Visualizations

We analyze the two datasets introduced earlier: (i) the **Logbook dataset** of operator-authored entries and (ii) the **Operator Ticket dataset** of structured issue reports. Covering the period **2022–2024**, both exhibit skewed distributions in text length and reporting frequency—most entries are concise, while a smaller subset contains extensive technical detail. Reporting activity also fluctuates over time, reflecting the irregular nature of free-text operational documentation.

Logbook Dataset. The corpus contains **2,979** entries authored by operators in 2022–2024. Each entry records observations, actions, or incidents. Among these, **1,144** (**38.4%**) are labeled as *issues*, while **61.6%** are non-issue. Mean text length is **417** characters (median 121, min 4, max 20,044), indicating mostly concise notes with a small number of long technical narratives. Yearly volume rose from 514 in 2022 to 1,566 in 2024 (Fig. 1).

Operator Ticket Dataset. This dataset includes **758** tickets across **13** issue categories. Each ticket combines metadata with a detailed *history* of troubleshooting communications. Mean history length is **1,914** characters (median 872, min 61, max 63,338), again with a long-tailed profile. Tickets per year increased from 164 in 2022 to 368 in 2024. The most frequent categories are *Other / Information*, *Electronics / Computer Hardware*, and *Software / Control System*, while subsystems such as *Pump Laser* and *Oscillator* are comparatively rare (Fig. 2).

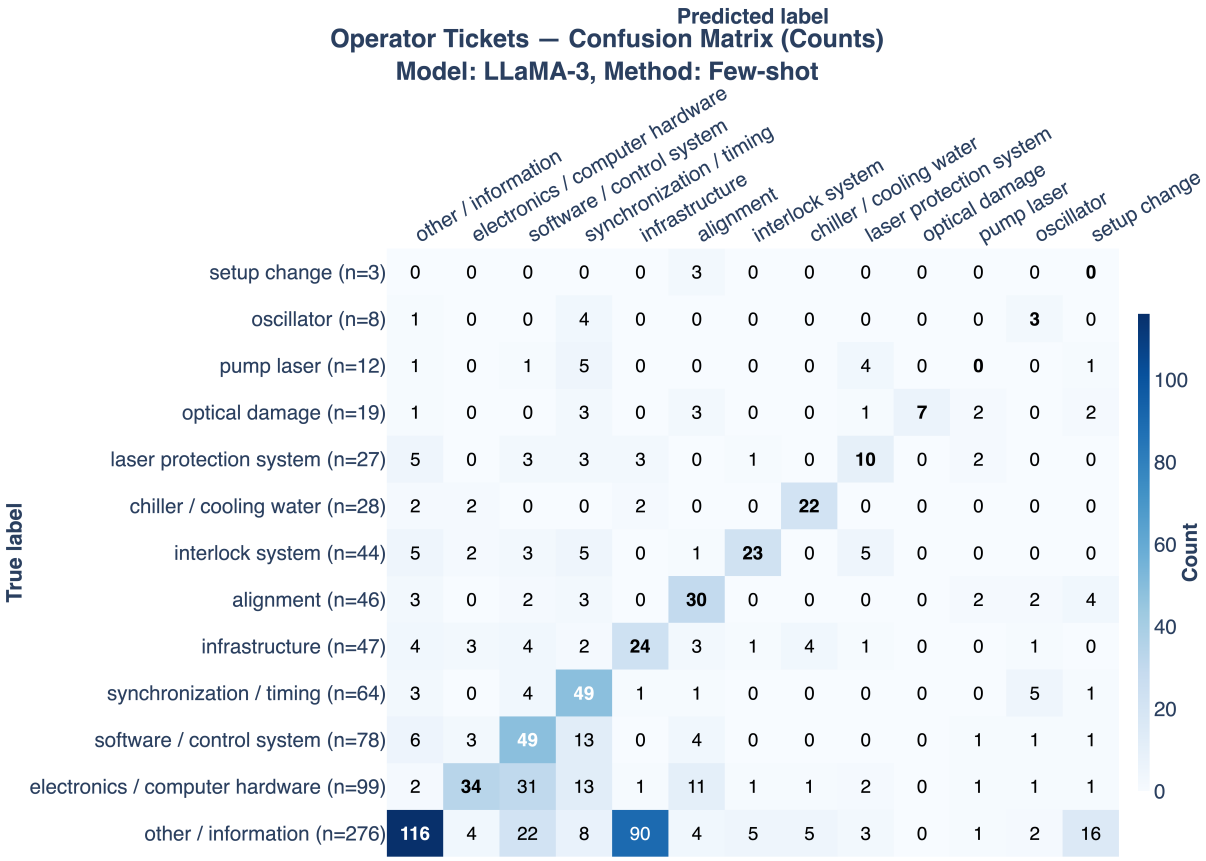


Figure 3: Operator Ticket categorization: confusion matrix (raw counts) for **LLaMA-3 (few-shot)**. Predictions are aggregated over three prompt variants by per-ticket majority vote; rows are true labels (with class support n), columns are predicted labels.

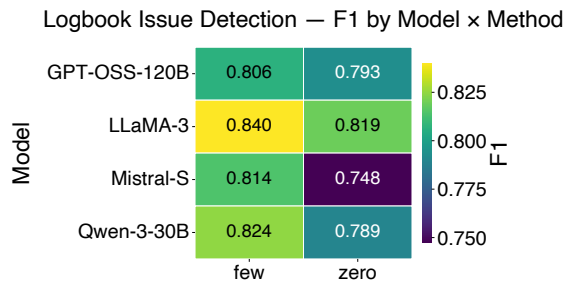


Figure 4: **Logbook issue detection.** Mean **F1 scores** per **model** × **method**. Few-shot prompting outperforms zero-shot, with **LLaMA-3 (few-shot)** performing best.

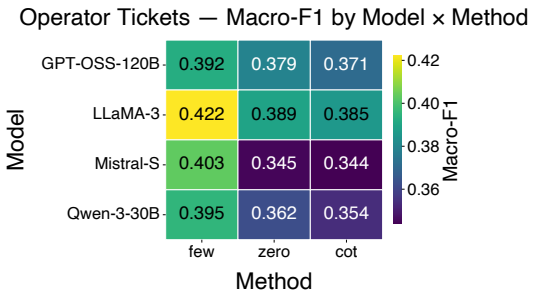


Figure 5: **Operator ticket classification.** Mean **Macro-F1 scores** per **model** × **method**. Few-shot prompting yields the strongest results, with **LLaMA-3 (few-shot)** best overall.

4. Experiments

4.1. Experimental Setup

Models and Prompting. We evaluate four open LLMs served via `Ollama`: *GPT-OSS-120B*, *LLaMA-3.3*, *Mistral-Small 3.2*, and *Qwen-3-30B*. Each model was executed in its quantized form for efficient inference: *GPT-OSS-120B* (117B parameters, MXFP4, 65 GB VRAM), *LLaMA-3.3* (70.6B, Q4_K_M, 43 GB), *Mistral-Small 3.2* (24B, Q4_K_M, 15 GB), and *Qwen-3-30B* (30.5B, Q4_K_M, 19 GB). Each model is tested under three prompting strategies: **zero-shot**, **few-shot**, and **chain-of-thought (CoT)**. For the few-shot setting, we include three exemplars per class, drawn from operator records of different years that are not part of the evaluated dataset. Unless stated otherwise, decoding is deterministic (temperature set to zero, nucleus sampling threshold $p{=}1$), with a generation budget of up to

2,048 tokens and an 8,192-token context window.

Prompt Variants and Robustness. To assess sensitivity to prompt wording, we evaluate multiple semantically equivalent variants for each *model* × *method* pair and report the mean and standard deviation across variants. For the **logbook** task, we use six variants, as pilot runs revealed higher variance due to polarity cues and negation. For the **operator tickets** (multi-class) task we use **3** variants: prompts are longer (label list + exemplars), and pilot runs indicated diminishing returns beyond three (std typically < 0.01 Macro-F1), so we fix three for efficiency.

Tasks and Measures. We evaluate two classification tasks drawn from real operations: (1) **Logbook issue detection** (binary), and (2) **Operator ticket categorization** (multi-class, normalized technical categories). For both tasks we compute **Accuracy**, **Precision**, **Recall**, **F1**, and **Balanced Accuracy** at the instance level. Given class imbalance, we emphasize **F1** for the logbook task and **Macro-F1** for tickets.

Supervised Baselines. We include supervised discriminative baselines to contextualize LLM performance. For both tasks, we train linear classifiers over TF-IDF representations (unigrams and bigrams). Specifically, we evaluate TF-IDF + Linear SVM and Logistic Regression models using 5-fold stratified cross-validation and report mean $\pm$ standard deviation across folds.

4.2. Results

Logbook Issue Detection. Across models, **few-shot prompting** consistently outperforms **zero-shot** on F1 (Fig. 4), with the average F1 score being approximately **0.03 higher** in the few-shot setting. The best-performing configuration is **LLaMA-3 (few-shot)** (F1 $\approx$ **0.840**), showing small variance (standard deviation $\approx$ 0.005) across prompt variants, which indicates stable behavior across different prompt formulations. A supervised **TF-IDF + Linear SVM** baseline achieves **F1 = 0.806** with a standard deviation of **0.007** under 5-fold stratified cross-validation (Table 3). While this lexical baseline already performs strongly, few-shot LLaMA-3 achieves an additional improvement of about **0.03 F1** (approximately **3 percentage points**), suggesting that in-context learning captures semantic regularities beyond surface n-gram features. Compared to the random baseline (F1 $\approx$ 0.50), all models demonstrate substantial improvements. LLaMA-3 also maintains a clear lead over the second-best model, **Qwen-3-30B (few-shot)** (F1 $\approx$ **0.824**).

Method	Logbook F1	Ticket Macro-F1
Random baseline	0.50	0.083
TF-IDF (1–2g) + Linear SVM	0.806 ± 0.007	0.408 ± 0.057
TF-IDF (1–2g) + Logistic Regression	0.777 ± 0.005	0.383 ± 0.063
LLaMA-3 (few-shot)	$\mathbf{0.840 \pm 0.005}$	$\mathbf{0.421 \pm 0.008}$

Table 3: Comparison of strongest prompting configuration and supervised discriminative baselines. Results for TF-IDF models are reported using 5-fold stratified cross-validation (mean $\pm$ std).

Zero-shot prompting tends to trade precision for higher recall, producing more false positives.

Operator Ticket Categorization. Figure 5 summarizes **Macro-F1** for all models and prompting strategies. The task is challenging due to $K = 13$ closely related issue categories and an underlying random Macro-F1 baseline of $\approx$ 0.083. A supervised **TF-IDF + Linear SVM** baseline achieves **Macro-F1 = 0.408 $\pm$ 0.057** under 5-fold cross-validation. Few-shot prompting again performs best across models, with **LLaMA-3 (few-shot)** reaching a peak Macro-F1 $\approx$ **0.421** ($\pm$ 0.008). This indicates that LLMs remain competitive with, and slightly outperform, strong lexical discriminative baselines even in multi-class classification across **13 closely related technical subsystem categories**. On average, few-shot leads both zero-shot and CoT by $\approx$ 0.01–0.02 in absolute Macro-F1. **Zero-shot** and **CoT** prompting yield slightly lower scores, suggesting that explicit reasoning steps (*i.e.*, CoT) may not be beneficial for concise operator texts or that the overhead of the prompt reduces the context available for the ticket text itself. The overall performance range is tight, spanning from the worst model's 0.385 to LLaMA-3's 0.421, confirming consistent model behavior across variants.

Table 3 summarizes the strongest configuration across prompting and supervised baselines. Few-shot LLaMA-3 consistently achieves the highest performance on both tasks.

Confusion Matrix and Detailed Error Analysis. Figure 3 shows the confusion matrix for the best-performing configuration, **LLaMA-3 (few-shot)**. The model demonstrates acceptable performance on high-frequency, distinct classes, achieving strong **Recall** for *Chiller/Cooling Water* ($\approx$ **84.6%**) and *Synchronization/Timing* ($\approx$ **77.8%**). However, a detailed error analysis reveals two primary failure modes that constrain the overall Macro-F1 score:

Failure Mode 1: Misclassification of Informational Tickets. The largest source of misclassification involves the general *Other/Information* category. In many cases, tickets that are in-

formational or do not correspond to a specific subsystem are incorrectly assigned to more specific technical categories by the model. For example, several *Other/Information* tickets are predicted as *Infrastructure* or other subsystem classes. This behavior suggests that the LLM tends to infer a concrete technical cause even when the ticket text does not clearly indicate one. Because the *Other/Information* category represents non-actionable or informational entries, its boundaries are inherently less explicit, which makes it difficult for the model to distinguish these cases from genuine subsystem faults.

Failure Mode 2: Semantic Confusion and Critical Recall Gaps. Errors frequently occur between semantically related systems, highlighting the difficulty in distinguishing between related failures from a short ticket description. Most notably, there is strong bidirectional confusion between *Electronics/Computer Hardware* (51.5% Recall) and *Software/Control System* (62.8% Recall), where **31 true Electronics tickets** were misclassified as Software. More critically, the model exhibits complete failure (**0.0% Recall**) on the lowest-frequency classes, *Pump Laser* and *Setup Change*. The 12 tickets for *Pump Laser*, for instance, were scattered primarily across *Synchronization/Timing* and *Laser Protection System*, confirming that the long-tailed data distribution provides insufficient distinguishing examples for these rare fault types.

Representative Error Cases. Binary (False Positive). *Gold: Not Issue → Predicted: Issue* "leakage sensor leakage sensor prepared alarm." The presence of terms such as *leakage* and *alarm* strongly signals fault-like language, leading the model to predict an issue. However, the entry documents a preparatory configuration rather than an active malfunction, illustrating how lexical cues can trigger false positives when context is underspecified. **Binary (False Negative).** *Gold: Issue → Predicted: Not Issue* "[PERSON] elogbook entry was sent to following experts: [EMAIL]." Although labeled as an issue, the entry primarily contains forwarding metadata without explicit technical description. The lack of observable fault terminology leads the model to treat it as informational, highlighting annotation-boundary ambiguity between operational coordination and incident reporting. **Multi-class (Hardware vs. Software Confusion).** *Gold: Electronics / Computer Hardware → Predicted: Software / Control System* "Polarization in user panel cannot be changed remotely anymore." Remote control failures can stem from hardware components, firmware, network interfaces, or control software. The linguistic surface form emphasizes remote interaction, which biases the model toward a software/control interpretation despite the underlying hardware-related categorization. **Multi-class ("Other" Sinkhole).** *Gold: Interlock System → Predicted: Other / Information* "[XFELelog] alic_server.xml ... Direct link to e-logbook entry: [URL]." Entries dominated by metadata, links, or forwarding notes lack explicit fault descriptors. In such underspecified cases, the model defaults to the broad *Other / Information* category, reflecting both class imbalance and semantic ambiguity in short coordination-oriented tickets.

5. Conclusion

We introduced two datasets derived from real-world operations of a large-scale scientific facility: a binary **logbook issue detection** dataset and a multi-class **operator ticket categorization** dataset. These resources provide one of the first benchmarks for studying LLM-based understanding of operator-authored technical documentation in complex experimental environments.

Our evaluation of four open LLMs under zero-shot, few-shot, and chain-of-thought prompting shows that **few-shot prompting consistently provides the strongest performance**, improving F1 by approximately **0.03** compared to zero-shot setups. Few-shot LLaMA-3 also slightly outperforms strong supervised baselines such as TF-IDF + Linear SVM, indicating that in-context learning can capture domain-specific patterns beyond lexical features. While chain-of-thought reasoning did not significantly improve classification accuracy, it produced interpretable rationales that may support explainable annotation workflows. Overall, the results indicate that open LLMs can effectively interpret highly technical operator-authored records with minimal in-domain adaptation.

Future work will focus on improving dataset quality and extending analysis beyond single-entry classification. Planned directions include clustering related logbook entries referring to the same incident, incorporating multimodal information from screenshots and diagnostic plots, refining the taxonomy to reduce category ambiguity, and strengthening annotation guidelines and annotator training to improve agreement.

6. Acknowledgments

We gratefully acknowledge DESY (Deutsches Elektronen-Synchrotron), a member of the Helmholtz Association, for providing the operational environment in which the datasets used in this study were created. The logbook entries and operator tickets analyzed in this work originate from routine facility operations at DESY. We sincerely thank the laser operators and technical

experts whose documentation efforts and domain expertise made this dataset possible, as well as the experts who contributed to the annotation and validation of issue categories.

We further acknowledge financial support from DASHH (Data Science in Hamburg – Helmholtz Graduate School for the Structure of Matter). Parts of this research were funded by and carried out at the University of Hamburg. The work of Anne Lauscher is funded under the Excellence Strategy of the German Federal Government and the Federal States.

7. Ethical Considerations and Limitations

All operator ticket and logbook data used in this study were fully anonymized in accordance with institutional data protection and privacy guidelines prior to analysis. The datasets contain no personal or identifying information and are shared only in anonymized form following institutional approval. While the models show strong in-domain performance, their generalizability beyond the specific operational environment remains to be validated. Furthermore, overlapping or ambiguous issue categories may introduce label noise, which could affect classification consistency. All model predictions are used solely for research purposes, and no automated decisions are deployed without human verification.

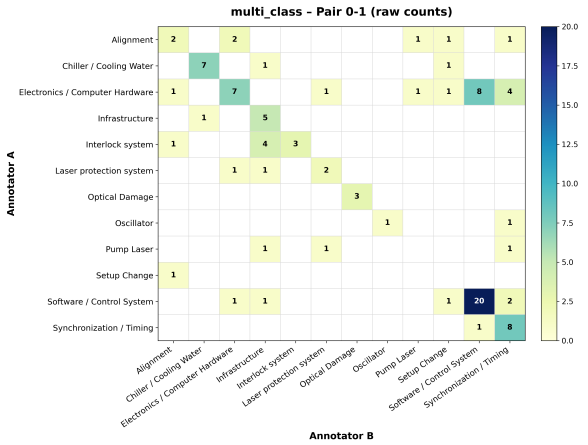
8. Supplementary Materials

Operator Ticket Label Definitions Table 4 lists the operator ticket categories and their definitions used in the multi-class classification task.

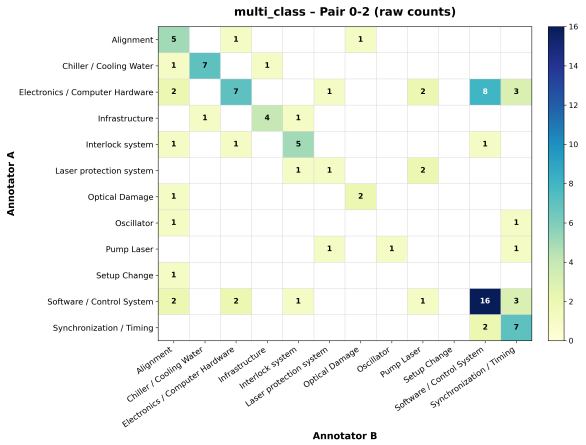
Annotator Confusion Matrices To further analyze the annotation process, we include pairwise confusion matrices for the 100-sample re-annotation subset. These visualizations highlight where annotators disagree and reveal common confusions between semantically related categories. Annotator 0 corresponds to the original labels collected before the introduction of detailed written annotation guidelines. The remaining annotators performed the re-annotation after receiving the guidelines. Comparing the resulting matrices shows that the guidelines improved annotation consistency: overall agreement increased and confusions between similar categories were reduced. These visualizations therefore complement the inter-annotator agreement statistics reported in Section 3.3 by providing a more detailed view of how the guidelines affected labeling behavior.

Category	Short definition
Alignment	Beam or mechanical alignment problems affecting laser performance (e.g., crystal misalignment, beam pointing drift, inactive beam stabilization).
Chiller / Cooling Water	Cooling system faults unrelated to house water, including chiller malfunction, low water flow, overheating, or corrosion.
Electronics / Computer Hardware	Failures of electronics, controllers, computers, or hardware components (e.g., MTCA crates/cards, controllers, power supplies, cameras, or storage devices).
Infrastructure	Facility-level issues such as power supply failure, network outages, house water problems, or laboratory environmental control failures.
Interlock System	Safety interlock faults involving access control or hardware components (e.g., door contacts, shutters, or interlock control units).
Laser Protection System	Malfunctions in laser protection mechanisms, including false triggers or failure to trip during fault conditions.
Optical Damage	Damage or degradation of optical components affecting beam quality or power and requiring inspection or replacement.
Oscillator	Oscillator instability or failure, including loss of mode-locking, component degradation, or startup problems.
Pump Laser	Failures in the pump laser system such as driver electronics faults, pump diode failures, or instability in pump output.
Setup Change	Planned configuration modifications or adjustments to the experimental or laser setup (e.g., wavelength change, pulse-shape adjustment, beam profile optimization).
Software / Control System	Problems in control software, automation scripts, or system configuration (e.g., feedback loop failures, control panel freezes, device communication errors).
Synchronization / Timing	Timing and synchronization issues such as loss of sync lock, clock drift, timing signal failures, or delay scan errors.
Other / Information	Informational or coordination entries that do not describe a technical fault or require corrective action.

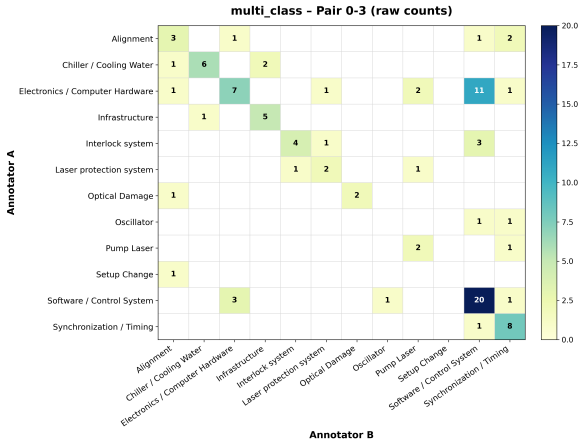
Table 4: Operator ticket categories and short definitions used for the multi-class classification task.



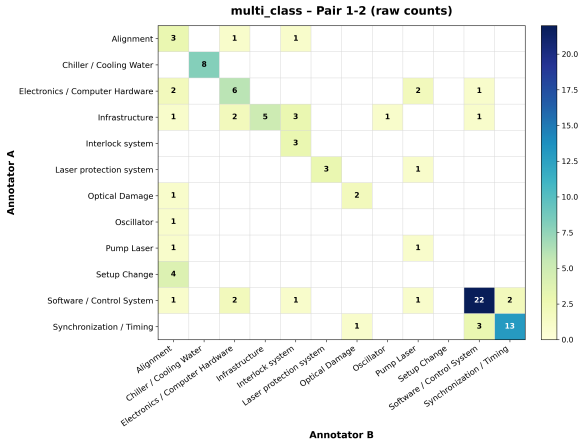
(a) Annotator 0 vs 1



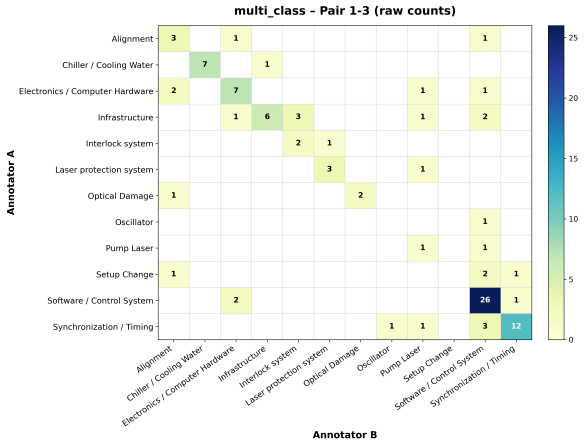
(b) Annotator 0 vs 2



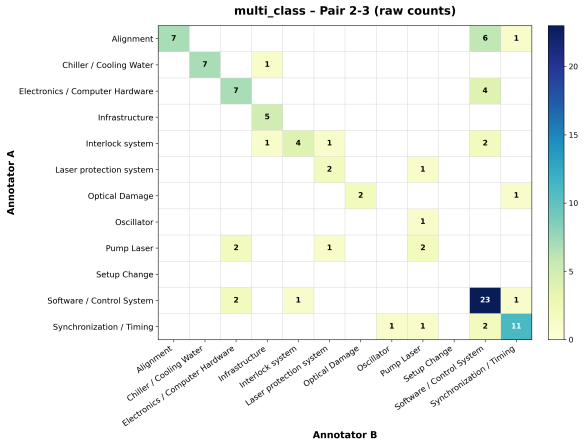
(c) Annotator 0 vs 3



(d) Annotator 1 vs 2



(e) Annotator 1 vs 3



(f) Annotator 2 vs 3

Figure 6: Pairwise confusion matrices between annotators on the 100-sample re-annotation subset of the operator ticket dataset. Rows correspond to labels assigned by the first annotator and columns correspond to labels assigned by the second annotator. These matrices complement the Cohen's κ agreement statistics reported in Section 3.3.

9. Bibliographical References

A Akhbardeh, F Lin, C Zhang, Y Zhou, S Natarajan, S Mohan, J Goldstein, F Suchanek, A Gray, J Wiener, et al. 2020. Maintnet: Predicting unplanned downtime using event correlation mining. *ArXiv preprint*, abs/2005.12443.

Ive Botunac, Marija Brkić Bakarić, and Maja Matetić. 2024. Comparing fine-tuning and prompt engineering for multi-class classification in hospitality review analysis. *Applied Sciences*, 14(14):6254.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned'small'llms (still) significantly outperform zero-shot generative ai models in text classification. *ArXiv preprint*, abs/2406.08660.

Winfried Decking, S Abeghyan, P Abramian, A Abramsky, A Aguirre, C Albrecht, P Alou, M Altarelli, P Altmann, K Amyan, et al. 2020. A mhz-repetition-rate hard x-ray free-electron laser driven by a superconducting linear accelerator. *Nature photonics*, 14(6):391–397.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia. ELRA and ICCL.

Richard Fechner and Jens Dörpinghaus. 2024. No train, no pain? assessing the ability of llms for text classification with no finetuning. In *19th Conference on Computer Science and Intelligence Systems*, pages 9–16.

Wei Guan, Jian Cao, Shiyou Qian, Jianqi Gao, and Chun Ouyang. 2024. Logllm: Log-based anomaly detection using large language models. *ArXiv preprint*, abs/2411.08561.

Melinda Hodkiewicz, Ming-Tzu Ho, et al. 2021. Pipeline for machine reading of unstructured maintenance work order records. *Reliability Engineering & System Safety*.

Elena Leitner and Georg Rehm. 2025. Exploring the limits of llms for german text classification: Prompting and fine-tuning strategies across small and medium-sized datasets. *Journal for Language Technology and Computational Linguistics*, 38(2):1–12.

Marc-Alexander Lutz, Bastian Schäfermeier, Rachael Sexton, Michael Sharp, Alden Dima, Stefan Faulstich, and Jagan Mohini Aluri. 2023. Kpi extraction from maintenance work orders—a comparison of expert labeling, text classification and ai-assisted tagging for computing failure rates of wind turbines. *Energies*, 16(24):7937.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184, Singapore. Association for Computational Linguistics.

Simon Münker, Kai Kugler, and Achim Rettinger. 2024. Zero-shot prompt-based classification: topic labeling in times of foundation models in german tweets. *ArXiv preprint*, abs/2406.18239.

Jörg Rossbach, Jochen R Schneider, and Wilfried Wurth. 2019. 10 years of pioneering x-ray science at the free-electron laser flash at desy. *Physics reports*, 808:1–74.

Antonin Sulc, Alex Bien, Annika Eichler, Daniel Ratner, Florian Rehm, Frank Mayet, Gregor Hartmann, Hayden Hoschouer, Henrik Tuennermann, Jan Kaiser, et al. 2024. Towards unlocking insights from logbooks using ai. *ArXiv preprint*, abs/2406.12881.

Antonin Sulc, Thorsten Hellert, Aaron Reed, Adam Carpenter, Alex Bien, Chris Tennant, Claudio Bisegni, Daniel Lersch, Daniel Ratner, David Lawrence, et al. 2025. elog analysis for accelerators: status and future outlook. *ArXiv preprint*, abs/2506.12949.

Thanakorn Thaminkaew, Piyawat Lertvittayakumjorn, and Peerapon Vateekul. 2024. Prompt-based label-aware framework for few-shot multi-label text classification. *IEEE Access*, 12:28310–28322.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Sowmya Vajjala and Shwetali Shimangaud. 2025. Text classification in the llm era–where do we stand? *ArXiv preprint*, abs/2502.11830.

Zhiqiang Wang, Yiran Pang, Yanbin Lin, and Xingquan Zhu. 2024. Adaptable and reliable text classification using large language models. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 67–74. IEEE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kai Yin, Bo Li, Chengkai Liu, Ali Mostafavi, and Xia Hu. 2024. Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *ArXiv preprint*, abs/2406.15477.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2025. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. In *Proceedings of the ACM on Web Conference 2025*, pages 2032–2042.

Aoxiao Zhong, Dengyao Mo, Guiyang Liu, Jinbu Liu, Qingda Lu, Qi Zhou, Jiesheng Wu, Quanzheng Li, and Qingsong Wen. 2024. Logparser-llm: Advancing efficient log parsing with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 4559–4570. ACM.