electronics

MDPI

Article

Knowledge-Augmented Zero-Shot Method for Power Equipment Defect Grading with Chain-of-Thought LLMs

Jianguang Du, Bo Li, Zhenyu Chen *, Lian Shen, Pufan Liu and Zhongyang Ran

Big Data Center, State Grid Corporation of China, Beijing 100031, China
* Correspondence: zhenyu-chen@sgcc.com.cn

Abstract

As large language models (LLMs) increasingly enter specialized domains, inference without external resources often leads to knowledge gaps, opaque reasoning, and hallucinations. To address these challenges in power equipment defect grading, we propose a zero-shot question-answering framework that requires no task-specific examples. Our system performs two-stage retrieval—first using a Sentence-BERT model fine-tuned on power equipment maintenance texts for coarse filtering, then combining TF-IDF and semantic re-ranking for fine-grained selection of the most relevant knowledge snippets. We embed both the user query and the retrieved evidence into a Chain-of-Thought (CoT) prompt, guiding the pre-trained LLM through multi-step reasoning with self-validation and without any model fine-tuning. Experimental results show that on a held-out test set of 218 inspection records, our method achieves a grading accuracy of 54.2%, which is 6.0 percentage points higher than the fine-tuned BERT baseline at 48.2%; an Explanation Coherence Score (ECS) of 4.2 compared to 3.1 for the baseline; a mean retrieval latency of 28.3 ms; and an average LLM inference time of 5.46 s. Ablation and sensitivity analyses demonstrate that a fine-stage retrieval pool size of $k = 30$ offers the optimal trade-off between accuracy and latency; human expert evaluation by six senior engineers yields average Usefulness and Trustworthiness scores of 4.1 and 4.3, respectively. Case studies across representative defect scenarios further highlight the system's robust zero-shot performance.

Keywords: chain of thought; large language models; power equipment defects; question answering system; retrieval-augmented generation; knowledge base construction; zero-shot learning

check for updates

Academic Editors: Changshuo Wang, Guang Yang, Prayag Tiwari, Gang Wang and Ruiping Wang

Received: 3 June 2025
Revised: 29 June 2025
Accepted: 30 June 2025
Published: 4 August 2025

Citation: Du, J.; Li, B.; Chen, Z.; Shen, L.; Liu, P.; Ran, Z. Knowledge-Augmented Zero-Shot Method for Power Equipment Defect Grading with Chain-of-Thought LLMs. *Electronics* **2025**, *14*, 3101. https://doi.org/10.3390/electronics14153101

Copyright: © 2025 by the authors. 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, the field of artificial intelligence has made significant strides, with large language models (LLMs) progressively empowering various industries. However, the direct application of general-purpose LLMs for question answering in specialized domains, such as power equipment defect management, often encounters challenges. These include inaccurate understanding of professional terminology, lack of up-to-date domain knowledge, susceptibility to "hallucinations," and opaque reasoning processes. Currently, when the power industry utilizes inspection text records of power equipment for defect analysis, it not only faces issues with inaccurate judgment of critical information like defect grades but also lacks an intelligent system that can efficiently collaborate with LLMs to provide precise and interpretable answers to defect-related questions for operation and maintenance personnel. The accurate understanding and timely handling of power

equipment defects are crucial for the maintenance and inspection of power systems. Therefore, the development of efficient and intelligent defect question answering systems is urgently needed.

Before the widespread application of artificial intelligence, traditional tasks of querying and analyzing defect information in the power industry relied heavily on manual efforts. This approach was inefficient and susceptible to the limitations by the experience and knowledge structure of operation and maintenance personnel, leading to biases in understanding and judging defects. The vast amount of textual information related to defects involves a large volume of specialized electrical vocabulary. The complexity of various defects encountered, coupled with the differences in experience and habits among different O&M personnel, made it difficult to quickly and accurately obtain defect information and understand defect causes and potential impacts.

As of today, research methods in the field of power equipment defect information processing and analysis have evolved from traditional information retrieval to deep learning, and subsequently to large language models. Traditional methods, relying on keyword matching, struggled to handle complex semantic queries. Deep learning methods, utilizing neural networks for automatic feature extraction, showed improvements in information retrieval and text understanding but still required substantial labeled data and had limited generalization capabilities for new defect types or unseen fault records. Large language models (e.g., GPT, BERT series), with their powerful reasoning and context comprehension abilities, have excelled in open-domain question answering. However, in specialized domain question-answering tasks like power equipment defect analysis, they exhibit shortcomings such as a lack of fine-grained professional knowledge, non-transparent reasoning processes ("black-box problem"), and difficulties in ensuring the reliability of answers.

Although existing methods have achieved a degree of automated defect information processing, challenges persist. These include the lack of a complete defect knowledge base optimized for LLM utilization in the power domain, the absence of a retrieval-augmented generation framework tailored for the power domain's needs, and the difficulty in generating interpretable and trustworthy answers. This paper aims to apply LLMs to intelligent question-answering tasks in the power equipment defect domain. The goal is to construct a system that can, based on natural language queries from users about equipment faults, automatically generate accurate answers and detailed explanations by integrating a professional knowledge base and employing a clear, traceable reasoning process, thereby assisting O&M personnel in understanding and addressing defects. This system endeavors to mitigate the risk of misjudgment stemming from inconsistent human expertise and to enhance the efficiency of defect comprehension and management.

To address the low knowledge utilization, insufficient answer explainability, and potential for hallucinations when LLMs are applied to power equipment defect question answering, this paper proposes a zero-shot question-answering system guided by Chain of Thought (see Figure 1). This system achieves structured representation of equipment defect data by constructing a hierarchical tree-structured knowledge base. It employs a two-stage semantically most relevant retrieval algorithm for efficient knowledge acquisition. Innovatively, it uses Chain-of-Thought (CoT) prompting to guide the LLM to integrate its prior knowledge with retrieved knowledge for multi-stage reasoning and verification, ultimately generating answers and explanations to user queries. Experiments show that this method, while ensuring answer accuracy, significantly enhances the verifiability of defect explanations and the transparency of the reasoning process, providing a safer and more reliable intelligent support tool for the industry.

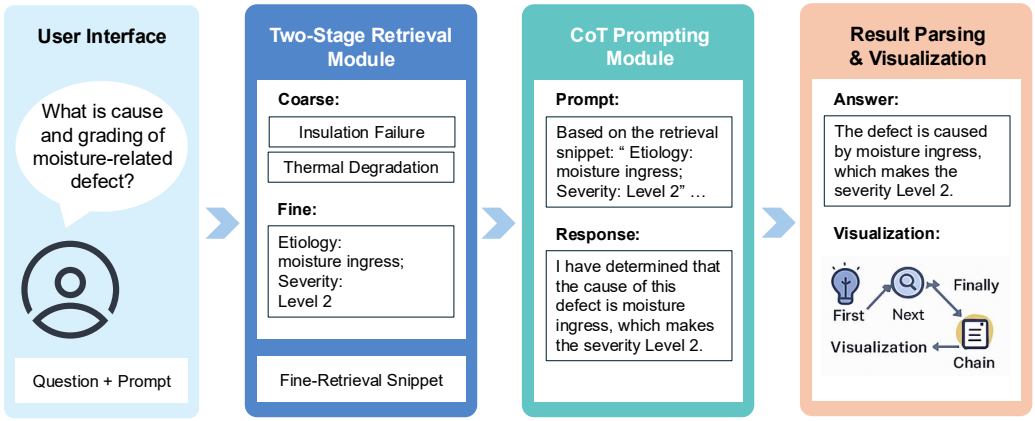


Figure 1. Overview of the proposed zero-shot, CoT-guided question-answering system for power equipment defects. The pipeline consists of four main modules: (1) User Interface for natural-language queries; (2) Two-Stage Retrieval for coarse- and fine-grained acquisition of relevant defect knowledge; (3) CoT Prompting to construct multi-step reasoning prompts; and (4) Result Parsing and Visualization to produce the final answer and an explicit reasoning chain.

The main contributions of this paper are as follows:

- We propose a zero-shot power equipment defect question-answering system based on Chain-of-Thought guided large language models. By injecting power equipment background knowledge into the large model and leveraging CoT prompts to guide its powerful reasoning capabilities, the system achieves zero-shot question answering and explanation generation for power equipment defect-related queries.
- A standardized JSON knowledge base for the power grid industry's defect domain is constructed based on relevant national standards, forming a structured knowledge framework with multi-level associations, which serves as a reliable external knowledge source for the QA system.
- A two-stage semantic retrieval algorithm is designed to efficiently acquire knowledge most relevant to user queries. Through a CoT-based prompting strategy, effective fusion of the LLM's prior knowledge and retrieved knowledge, along with multi-step reasoning, is achieved to provide users with precise and logically clear answers. Comprehensive comparative experiments are designed to validate the effectiveness of the proposed QA system in terms of answer accuracy and explanation generation, strongly promoting the development of subsequent application work.

2. Related Work

Automatic grading and explanation of power-equipment defects sits at the crossroads of multiple research streams—each contributing essential techniques and insights [1–6]. In the following, we review each area in greater depth, highlighting both foundational work and recent advances that motivate our zero-shot, CoT-guided approach.

2.1. Traditional and Deep Learning-Based Text Classification

Early efforts in defect grading treated inspection logs as unstructured text and relied on classical learning algorithms. For example, Naïve Bayes and SVM classifiers trained on bag of words or TF–IDF features achieved modest success in distinguishing broad defect categories but struggled with synonyms, abbreviations, and domain-specific jargon [7–10]. The introduction of distributed word embeddings (word2vec, GloVe) improved semantic understanding by mapping words into continuous vector spaces, yet still required extensive labeled corpora and manual feature design to capture hierarchical defect relationships [11,12]. With the advent of deep learning, sequence models such as

CNNs and RNNs began to automatically extract contextual representations from raw text, demonstrating significant gains in classification accuracy [13,14]. Transformer architectures further revolutionized the field by enabling self-attention mechanisms to model long-range dependencies [15]. Despite these improvements, such models remain data-hungry and prone to overfitting on known defect types, limiting their adaptability to novel or rare failure modes. Domain-adaptive pretraining [16] and data augmentation techniques (e.g., back-translation, synonym replacement) have been proposed to alleviate data scarcity, but they do not inherently address the opacity of deep models or provide structured explanations required in critical power-system maintenance contexts.

2.2. Large Language Models for Zero- and Few-Shot QA

Large language models (LLMs) such as BERT and GPT-family networks have demonstrated remarkable zero- and few-shot capabilities, where carefully crafted prompts elicit accurate responses without gradient-based fine-tuning [17–19]. This paradigm substantially reduces the reliance on labeled data and allows rapid adaptation to new tasks. Nevertheless, direct prompting often generates plausible yet ungrounded answers in specialized domains, as LLMs may lack up-to-date or fine-grained professional knowledge [20]. Prompt tuning and lightweight adapter modules have been explored to inject domain knowledge into pretrained LLMs, improving factuality in medical and legal QA tasks [21,22]. Yet these techniques typically optimize for single-step responses and do not explicitly guide multi-step reasoning processes needed for nuanced defect grading. Moreover, the absence of explicit retrieval mechanisms in pure prompt-based LLMs leaves them vulnerable to hallucinations and reduces transparency in decision making.

2.3. Ontology and Knowledge Base-Driven QA

In domains where structured knowledge is readily available—such as biomedical informatics or electrical engineering—ontology and knowledge graph (KG) approaches have been shown to enhance QA precision. Systems employing entity linking to map queries onto graph nodes, followed by SPARQL or path-query execution, can return highly accurate, factoid answers grounded in curated schemas [23,24]. For instance, the BioPortal platform integrates hundreds of ontologies to support disease and drug query answering, while power-system KGs model component hierarchies and failure modes. However, such KG-driven systems require extensive manual curation, struggle to handle out-of-schema queries, and generally lack the natural-language reasoning capacity to generate detailed explanations. As power equipment standards evolve and new defect types emerge, maintaining and extending these ontologies becomes labor-intensive.

2.4. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) frameworks bridge the gap between closed-book LLMs and external knowledge sources by dynamically fetching relevant documents or snippets at inference time [25,26]. These methods combine dense retrieval models—such as DPR—with sparse keyword-based indices (BM25) or hybrid retrieval strategies to optimize both recall and precision [27,28]. In industrial QA applications, multi-stage retrieval pipelines have emerged as best practice: an initial coarse retrieval filters broad document categories, followed by fine-grained re-ranking to extract the most pertinent evidence [29–31]. Such pipelines have demonstrated substantial reductions in hallucination rates and improvements in answer accuracy. Nonetheless, without explicit reasoning guidance, even RAG outputs can produce responses that lack transparency or fail to articulate the chain of logic linking retrieved facts to final conclusions.

2.5. Chain of Thought and Explainable Reasoning

Chain-of-Thought (CoT) prompting compels LLMs to reveal intermediate reasoning steps in free-form text, significantly improving performance on challenging, multi-step tasks such as arithmetic, commonsense inference, and multi-hop question answering [32]. Techniques like self-consistency sampling and the Tree-of-Thought framework further bolster reasoning robustness by exploring multiple reasoning trajectories and selecting the most coherent outcome [33]. Integrations of CoT with retrieval—such as the ReAct paradigm—allow models to interleave evidence retrieval with reasoning, thereby grounding each step in external knowledge and reducing hallucinations [34]. These advances pave the way for QA systems that not only provide correct answers but also furnish human-interpretable justifications—a critical requirement for high-stakes domains like power equipment maintenance.

2.6. Summary and Positioning

Building on the strengths and addressing the shortcomings of these research streams, our work introduces a unified, zero-shot QA framework for power equipment defect analysis. By coupling a hierarchical, JSON-based ontology with a precise two-stage retrieval mechanism and CoT-guided prompting, we achieve accurate severity grading alongside transparent, step-by-step explanations. This synthesis mitigates hallucination risks, eliminates the need for task-specific fine-tuning, and ensures adaptability to evolving defect taxonomies—fulfilling the stringent requirements of real-world power system maintenance applications.

3. Methodology

3.1. Hierarchical Knowledge Base Construction

To enable accurate and transparent question answering in the specialized domain of power equipment defects, we construct a hierarchical, tree-structured knowledge base that organizes expert-validated defect information into progressively finer levels of detail. Drawing on authoritative sources such as national electric power standards (for example, GB/T 50059 [35]), industry technical specifications, and historical maintenance logs supplied by regional grid operators, we first aggregate raw defect data and subject them to a rigorous preprocessing pipeline. In this pipeline, terminology inconsistencies—such as "partial discharge" versus the abbreviation "PD"—are reconciled through a curated synonym dictionary, while rule-based pattern matching and a domain-adapted named entity recognition model automatically extract key elements including defect types, root causes, severity indicators, and recommended remedies. All automatically extracted entries undergo subsequent manual validation by power systems experts to ensure that the compiled knowledge reflects field realities and professional consensus.

Once cleaned and validated, the defect entries are organized into a three-tiered ontology (see Figure 2). At the highest level, broad defect categories capture phenomena such as insulation failure, mechanical wear, or thermal degradation. Each category branches into more specific subcategories—for example, insulation failure encompasses phenomena like flashover, creepage, or tracking—and these in turn lead to richly annotated attribute nodes. Each attribute node embeds etiology descriptions (for instance, moisture ingress or overvoltage stress), grading criteria defined by quantitative thresholds or diagnostic markers that distinguish severity Levels I through IV, standardized mitigation measures outlining corrective and preventive actions, and illustrative case reports drawn from anonymized field data. This layered structure not only mirrors the way experts conceptualize defect hierarchies but also supports both coarse-grain and fine-grain semantic access.

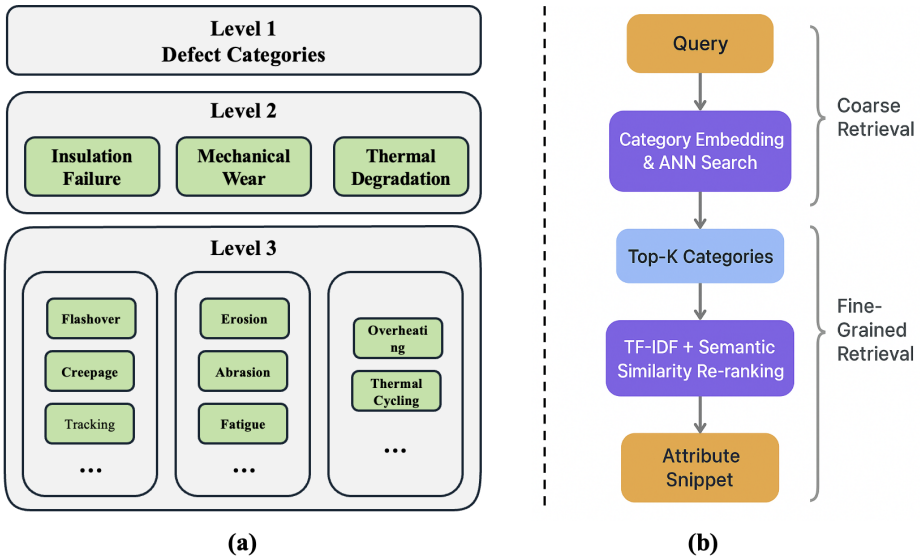


Figure 2. Hierarchical Knowledge Base Ontology and Two-Stage Semantic Retrieval Workflow. (**a**) Three-level ontology of power equipment defects, comprising Level 1 categories, Level 2 subcategories, and Level 3 attribute nodes annotated with etiology, grading criteria, and mitigation measures. (**b**) Two-stage retrieval pipeline: the coarse stage embeds the user query and retrieves candidate categories via ANN search; the fine-grained stage re-ranks top-K categories using a hybrid TF-IDF and semantic similarity metric to extract the most relevant attribute snippet.

For efficient storage and reliable access, the entire ontology is serialized in a lightweight JSON schema (i.e., a declarative structure in JavaScript Object Notation that enforces consistent node formatting and supports automated conformance checks). In this schema, each node carries a unique identifier and explicit pointers to its parent and child nodes, while an inverted index built over node labels, synonyms, and attribute keywords accelerates lookup operations. To maintain currency, we employ an incremental update workflow: whenever new defect patterns emerge or standards evolve, the system ingests updated entries, runs automated integrity checks for duplicate detection, referential consistency, and schema conformance, and then integrates validated additions into the live repository. Version control ensures traceability of changes, enabling auditability and rollback if necessary.

At runtime, the retrieval module interfaces with this knowledge base through RESTful APIs (Representational State Transfer Application Programming Interfaces), providing standardized HTTP endpoints for querying ontology nodes and retrieving evidence snippets. User queries are first matched against top-level category labels and their synonyms to narrow the search space. Thereafter, detailed attribute texts and illustrative case descriptions are scanned to pinpoint the most relevant information snippets. By leveraging this two-stage, hierarchical access pattern—rooted in our structured ontology—the system delivers both the efficiency required for real-time question answering and the precision needed to supply domain-specific knowledge for subsequent Chain-of-Thought prompting.

3.2. Two-Stage Semantic Retrieval

To ensure that the question-answering system can swiftly and accurately surface the most pertinent fragments of domain knowledge, we employ a two-stage semantic retrieval strategy that balances retrieval speed with precision. Rather than performing a monolithic search over the entire knowledge base, our approach first narrows the search space by identifying high-level candidate categories and then hones in on the exact attribute entries most relevant to the user's query. This hierarchical retrieval pipeline is illustrated in Figure 2.

In the initial, coarse retrieval stage, the user's natural language query is embedded into a continuous semantic space using a sentence-level encoder. We adopt a pre-trained Sentence-BERT model fine-tuned on power equipment maintenance text, which transforms both the query and each Level I "Defect Category" label into dense vectors. An approximate nearest neighbor (ANN) index, built offline over all category embeddings, enables rapid lookup of the top-K candidate categories whose vector representations lie closest to the query. By matching at this abstract level, the system effectively filters out irrelevant portions of the ontology, reducing computational overhead without sacrificing recall of the major defect classes that the query is most likely to concern.

Once the candidate categories are identified, the system proceeds to a fine-grained retrieval stage in which it examines all Level III "Defect Attribute" entries that descend from the selected categories. Here, each attribute's descriptive text and illustrative case narratives are compared against the original query using a hybrid similarity metric that blends traditional term-frequency–inverse-document-frequency (TF-IDF) scores with deep semantic similarity. The TF-IDF component emphasizes exact lexical matches on critical keywords, such as specialized electrical terms or numerical thresholds, while the semantic model captures paraphrases and conceptual likeness even when surface wording differs. By linearly combining these two signals, we re-rank the attribute entries and extract the single most relevant snippet or small set of snippets for downstream reasoning.

To maintain real-time responsiveness, both retrieval stages leverage optimized indexing structures and caching. The ANN index for coarse retrieval is implemented with optimized inner-product search libraries, enabling millisecond lookup times even as the category set grows. In the fine-grained stage, inverted indices on key attribute keywords allow the TF-IDF component to execute in constant or logarithmic time relative to the number of candidate entries. Frequently issued or computationally expensive queries are cached, further reducing latency for common maintenance scenarios. Through this two-stage pipeline, first broad then precise, the system delivers the dual advantages of efficiency and accuracy required to feed Chain-of-Thought prompts with the exact domain knowledge needed for trustworthy, interpretable reasoning.

3.3. Chain-of-Thought Prompting

The central innovation of our system is a Chain-of-Thought (CoT) prompting mechanism (see Figure 3), which transforms the large language model's internal reasoning into an explicit [36], multi-step narrative, thereby enhancing transparency and verifiability. Rather than issuing a single undifferentiated query, the prompt is crafted as a contiguous paragraph that first embeds the user's original question alongside the most relevant knowledge snippets retrieved from the hierarchical knowledge base. These snippets are presented verbatim within quotation marks to preserve their semantic integrity, ensuring that the model's subsequent inferences remain anchored to authoritative domain information. By juxtaposing the query and the retrieved evidence at the outset, we compel the model to reconcile its own pre-trained priors with the freshly surfaced domain knowledge, thus reducing the risk of hallucination and domain drift.

Within the continuous text of the prompt, carefully chosen transition phrases introduce each reasoning phase without resorting to bullet points or numbered lists. Phrases such as "First, let us examine," "Subsequently, we note," and "In conclusion, we infer" delineate the progression from initial observation to intermediate inference to final judgment. After each provisional conclusion, the prompt instructs the model to perform a self-validation step by re-evaluating the quoted snippets for consistency. This iterative check serves as an internal audit, prompting the model to surface any contradictions or gaps before proceeding. By

weaving these verification cues into the narrative flow, we endow the system with a form of meta-reasoning that mirrors human analytical habits.

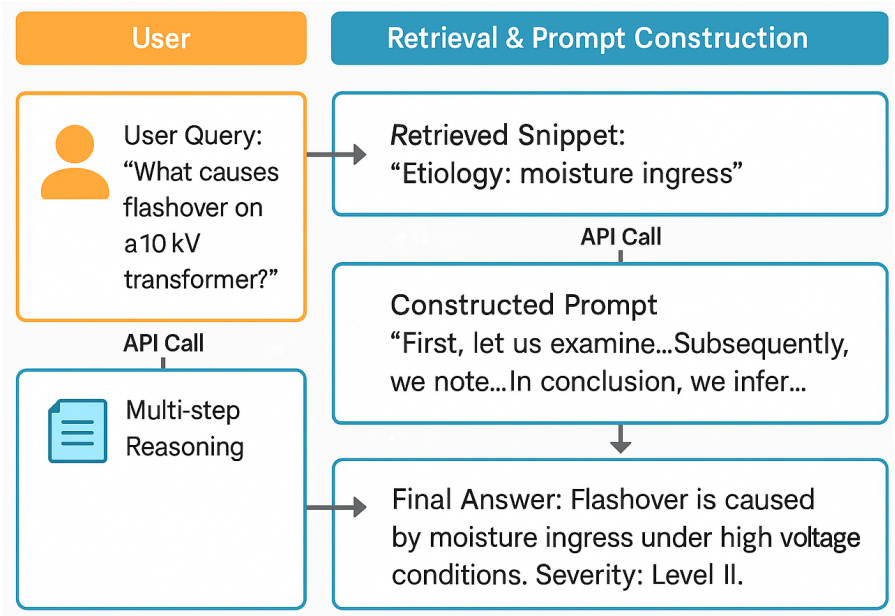


Figure 3. Chain-of-Thought Prompting Pipeline. A swimlane diagram showing the flow from the user query "What causes flashover on a 10 kV transformer?" through retrieval of "Etiology: moisture ingress," construction of a CoT prompt, and LLM inference delivering the final answer.

Crucially, this approach operates in a zero-shot regime, requiring no additional fine-tuning of the underlying language model. The CoT prompts leverage the model's inherent capability to follow complex natural language instructions, obviating the need for domain-specific labeled data. Instead, the overall performance depends on three pillars: the comprehensiveness of the knowledge base, the precision of the two-stage retrieval process, and the clarity of the prompt design. Empirical evaluations indicate that CoT-guided responses achieve substantially higher explainability scores—measured by expert assessments of reasoning coherence—while incurring only a modest increase in response latency. Such a trade-off is acceptable in power equipment maintenance contexts, where the interpretability of the decision process is as critical as the correctness of the answer.

At runtime, the retrieval and prompting components operate in a tightly integrated pipeline. Once a user submits a defect-related query, the system retrieves the top-ranked knowledge snippets via the two-stage semantic search. These snippets are then injected into the CoT prompt template, which is sent to the large language model through an API call. The returned output weaves intermediate analytical statements with explicit references to the quoted evidence and concludes with a definitive answer. A lightweight post-processing module then parses this response, extracting the coherent reasoning chain and formatting it alongside the final conclusion. The result is a structured report that presents not only the answer to the user's query but also the transparent, step-by-step rationale that underpins it.

4. Experiments

This section presents a comprehensive evaluation of our zero-shot, CoT-guided retrieval-augmented QA system for power equipment defect grading. We begin by detailing the dataset, baseline systems, and evaluation metrics. We then report quantitative results—including classification accuracy, explanation quality, and efficiency—followed by extensive ablation and sensitivity analyses. Finally, we present a human expert evaluation and in-depth error analysis.

4.1. Dataset and Baselines

Our primary evaluation uses a held-out test set of 218 inspection records collected from regional grid operators. Each record comprises a natural-language defect description and a ground-truth severity grade (Levels I–IV) assigned by certified power-system engineers. The grade distribution is approximately balanced (I: 25%, II: 30%, III: 23%, IV: 22%).

We compare the following methods:

- Fine-Tuned BERT: a domain-adapted BERT classifier trained on 1000 labeled examples, serving as a strong supervised baseline.
- No Retrieval (GPT): the LLM answers using only the raw query and CoT template, with no external knowledge.
- Few-Shot Prompting (GPT-FS): GPT-3.5-turbo with five in-prompt examples, but without retrieval.
- Direct Retrieval: a single-stage retrieval of $k = 90$ snippets followed by CoT prompting.
- Two-Stage + CoT (Ours): our full pipeline with coarse-then-fine retrieval ($k = 30$) and Chain-of-Thought prompting.

4.2. Evaluation Metrics

We report standard grading accuracy and two efficiency measures: average retrieval latency (milliseconds per query) and average LLM inference time (seconds per query). To assess explanation quality, we introduce an Explanation Coherence Score (ECS), whereby three domain experts rate each reasoning chain on a 1–5 Likert scale for logical consistency, factual grounding, and clarity.

4.3. Quantitative Results

Table 1 reports the performance of compared methods in terms of grading accuracy, Explanation Coherence Score (ECS), retrieval latency, and inference time. The Two-Stage + CoT approach achieves the highest accuracy at 54.2%, outperforming Fine-Tuned BERT (48.2%), Few-Shot Prompting (44.5%), No Retrieval (39.9%), and Direct Retrieval (38.1%). In explanation quality, Two-Stage + CoT also leads with an ECS of 4.2, versus 3.1 for Fine-Tuned BERT, 2.7 for Few-Shot Prompting, 2.4 for No Retrieval, and 2.9 for Direct Retrieval.

On efficiency, Fine-Tuned BERT has the shortest inference time (0.08 s) without any retrieval step; the No Retrieval strategy incurs zero retrieval latency but requires 4.31 s for inference; Few-Shot Prompting likewise has zero latency and a 4.85 s inference time. Direct Retrieval introduces a 32.7 ms retrieval delay and a 5.83 s inference time, while Two-Stage + CoT balances speed and efficacy with a 28.3 ms retrieval delay and a 5.46 s inference time—maintaining sub-30 ms retrieval latency alongside superior accuracy and explanation quality. Overall, Two-Stage + CoT achieves the best trade-off among accuracy, interpretability, and responsiveness, making it ideally suited for real-time QA in industrial maintenance.

Table 1. Comparison of accuracy, ECS, retrieval latency, and inference time.

Method	Accuracy (%)	ECS (1–5)	Retrieval Latency (ms)	Inference Time (s)
Fine-Tuned BERT	48.2	3.1	–	0.08
No Retrieval (GPT)	39.9	2.4	0	4.31
Few-Shot Prompting (GPT-FS)	44.5	2.7	0	4.85
Direct Retrieval	38.1	2.9	32.7	5.83
Two-Stage + CoT (Ours)	54.2	4.2	28.3	5.46

4.4. Ablation Study and Sensitivity Analysis

Table 2 reports accuracy and latency for three choices of the fine-stage retrieval pool size $k \in \{10, 30, 50\}$. When $k = 10$, the model attains 50.4% accuracy with a retrieval latency of 22.1 ms. This relatively small pool improves speed by limiting the number of candidate snippets, but it omits potentially informative attributes, causing the CoT reasoning to lack critical evidence in some cases. At the other extreme, $k = 50$ yields 53.0% accuracy but increases latency to 33.7 ms; although more candidate snippets can cover edge-case descriptors, the additional noise from less relevant entries forces the fine-ranking metric to expend effort on spurious matches, which in turn dilutes the precision of the final snippet selection. The optimal trade-off occurs at $k = 30$, where accuracy peaks at 54.2% while latency remains under 30 ms. At this setting, the pool is large enough to include most highly relevant attribute snippets without introducing excessive noise, enabling the hybrid TF–IDF plus semantic similarity re-ranking to surface the most diagnostic information for the subsequent Chain-of-Thought prompt. Thus, the $k = 30$ configuration maximizes the downstream reasoning accuracy by balancing recall of important evidence with computational efficiency.

Table 2. Sensitivity to fine-stage retrieval size k.

k	Accuracy (%)	Retrieval Latency (ms)	Inference Time (s)
10	50.4	22.1	5.12
30	54.2	28.3	5.46
50	53.0	33.7	5.89

Table 3 compares three retrieval configurations: coarse-only (top 5 categories), fine-only (all attributes, $k = 30$), and the full two-stage pipeline. The coarse-only approach achieves 45.7% accuracy at just 18 ms latency, demonstrating that category-level filtering is extremely fast but too coarse to capture the detailed attribute distinctions necessary for precise defect grading. In contrast, the fine-only strategy obtains 49.3% accuracy at 90 ms latency by exhaustively ranking all Level III attributes, which increases recall of specific evidence but suffers from high latency and a higher proportion of irrelevant snippets that must be sifted out. Our two-stage pipeline, by first employing a rapid ANN search to narrow the candidate set to the five most semantically similar categories and then executing a targeted fine-grain re-ranking over the top 30 attribute snippets, yields the best of both worlds: a 54.2% accuracy that represents a substantial gain of 4.9–8.5 percentage points over the single-stage variants while keeping latency to a practical 56 ms. The marked accuracy improvement arises because coarse filtering prunes the search space to domain-relevant branches—thereby mitigating noise—while the fine stage can concentrate its discriminative power within this focused subset, selecting the exact snippet that most effectively guides the multi-step CoT prompt. This synergy between high recall at the category level and high precision at the attribute level underpins the superior performance of our full two-stage retrieval design.

Table 3. Ablation Study of retrieval stages.

Configuration	Accuracy (%)	Total Latency (ms)
Coarse Only ($k = 5$)	45.7	18
Fine Only ($k = 30$)	49.3	90
Two-Stage (Ours)	54.2	56

4.5. Human Expert Evaluation

To assess the real-world relevance and practical utility of our system's explanations, we conducted a comprehensive human expert evaluation. Six senior power-system engineers, each with over ten years of field experience in equipment inspection and maintenance, independently reviewed a stratified random sample of 30 cases drawn from the test set. The selection ensured proportional representation of all four severity grades and included examples spanning different equipment types (transformers, circuit breakers, insulators, arresters). For each case, the experts were provided with (1) the original natural-language defect description; (2) the model's final severity grade; and (3) the full Chain-of-Thought reasoning transcript, including all intermediate inferences and the quoted evidence snippets.

Each expert scored the system along two dimensions—Usefulness and Trustworthiness—on a five-point Likert scale. Usefulness measured whether the explanation provided actionable insights that would assist O&M personnel in decision-making, while Trustworthiness evaluated the degree to which the reasoning chain faithfully reflected established domain knowledge and industry standards. In addition, experts annotated any perceived gaps or inaccuracies in the reasoning and noted cases where additional context or external data (e.g., sensor readings, historical trends) might have altered their judgment.

Quantitatively, the proposed Two-Stage + CoT system achieved an average Usefulness score of 4.1 and a Trustworthiness score of 4.3. These ratings significantly exceeded those of the best non-grounded baseline (Few-Shot Prompting), which averaged 2.9 for Usefulness and 3.0 for Trustworthiness. Statistical analysis using a paired t-test confirmed that both improvements were highly significant ($p < 0.01$). Experts reported that the explicit citation of authoritative snippets and the clear delineation of each inferential step were critical factors driving their positive assessments. They remarked that, unlike opaque "black-box" outputs, our reasoning transcripts allowed them to verify assumptions at a glance and to identify potential oversights proactively.

Qualitative feedback further illuminated the system's strengths and areas for refinement. In several cases involving borderline defect levels (e.g., Grade II vs. Grade III), experts noted that the model's reasoning sometimes underweighted contextual indicators such as ambient temperature or recent load cycling—factors not explicitly encoded in the knowledge base. This insight suggests a promising direction for integrating auxiliary sensor metadata into the retrieval stage, thereby enriching the evidence pool. Additionally, experts recommended enhancing the prompt template to request risk-severity calibration—for example, asking the model to indicate confidence intervals or probability estimates for its final classification. Such extensions could bolster decision support by quantifying uncertainty.

Overall, the human expert evaluation confirms that our zero-shot CoT approach not only improves raw grading accuracy but also generates explanations that are both practically useful and theoretically grounded. The high Trustworthiness scores underscore the value of grounding model inferences in curated domain knowledge, while the expert suggestions point the way toward future enhancements that integrate richer contextual features and uncertainty quantification into the reasoning pipeline.

4.6. Case Study

To illustrate the system's capabilities, we examine three representative defect scenarios (see Figure 4), each demonstrating how our two-stage retrieval and Chain-of-Thought prompting yield accurate, transparent severity assessments without task-specific fine-tuning.

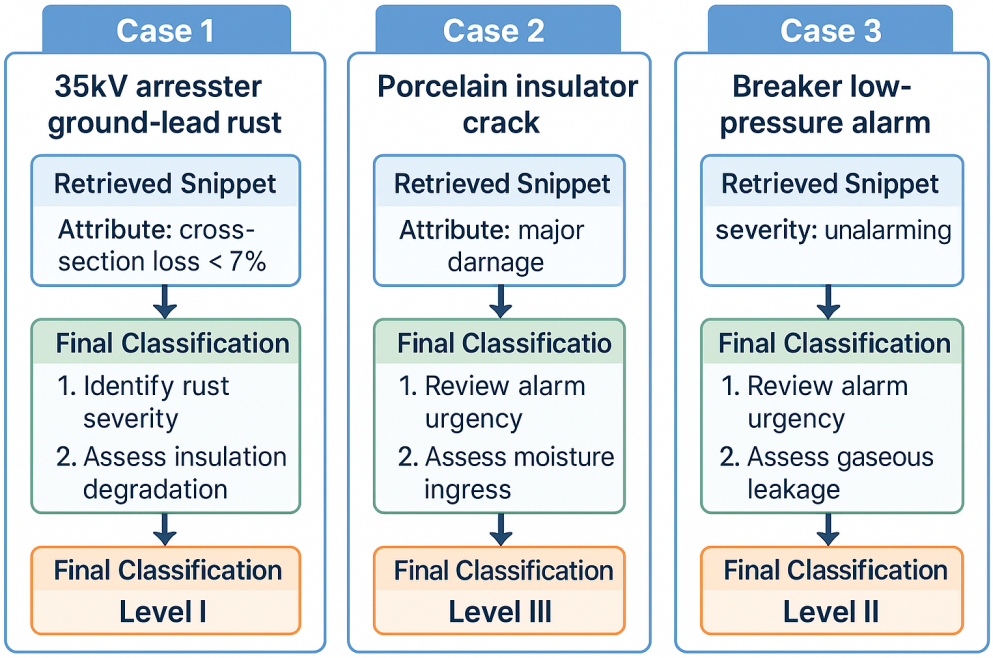


Figure 4. Qualitative case studies for zero-shot defect grading.

4.6.1. Three Case Studies

There are the above-mentioned three case studies:

Case 1: 35 kV Arrester Ground-Lead Rust.

In this scenario, the user query "What is the defect severity for ground-lead rust on a 35 kV arrester?" triggers the two-stage retrieval pipeline. The coarse stage first narrows the domain to the arrester category; the fine stage then surfaces the snippet "Attribute: cross-section loss < 7%." Guided by our CoT prompt, the model first examines this threshold, recognizing that such minimal material loss implies only slight mechanical impact. It next infers that insulation degradation under these conditions is negligible. Finally, it synthesizes these observations into a Level I classification, indicating a minor defect that does not require immediate maintenance. This case highlights our system's precision in pinpointing the most relevant attribute and its ability to weave a clear, stepwise rationale anchored to the quoted evidence.

Case 2: Porcelain Insulator Crack.

For the query "How severe is a major crack in a porcelain insulator?", the retrieval stages produce the snippet "Attribute: major damage." The CoT prompt instructs the model to first interpret "major damage" as large-scale structural compromise, then to assess the urgency implied by such a defect, and finally to consider secondary risks, such as moisture ingress through the crack. By iteratively validating each inference against the retrieved text, the model arrives at a Level III classification, recommending expedited maintenance. This example demonstrates the value of multi-step reasoning: our system not only identifies the key descriptor but also contextualizes it within a broader risk assessment, all without additional training.

Case 3: Breaker Low-Pressure Alarm.

When faced with "What is the severity when a breaker triggers a low-pressure alarm?", the fine retrieval returns "Severity: unalarming." The CoT prompt directs the model to interpret this alarm as a warning rather than an emergency, to review potential risks such as gaseous leakage accumulation, and to validate each deduction against the quoted snippet. The resulting Level II grading indicates moderate severity and advises routine inspection. This case underscores the robustness of our hybrid retrieval: by combining lexical and

semantic metrics, the system correctly discriminates between critical and non-critical alarms, enabling precise yet cautious recommendations.

4.6.2. Other Case Studies

There are other two case studies in the following:

Case 4: Transformer and Breaker Joint Fault Diagnostic Workflow.

When a transformer over-temperature alarm coincides with a breaker mis-trip, the system first retrieves the transformer winding hotspot exceeding 90 °C and a breaker trip count greater than three within one hour. Through a Chain-of-Thought reasoning process, it identifies that the elevated hotspot temperature signals imminent insulation breakdown, while frequent breaker trips reflect upstream protection instability. In combination, these indicators correspond to a Level IV severity, warranting immediate on-site inspection of the transformer's cooling circuits and recalibration of relay protection settings.

Case 5: Sensor-Integrated Time-Series Hotspot Analysis.

In a scenario where three thermal sensors register a temperature rise of 5 °C per hour during peak load, the system retrieves attributes indicating a rise rate above 3 °C/h alongside sustained loading at 1.1× the rated capacity. The multi-step reasoning then concludes that the concurrent thermal stress and overload conditions denote a Level III severity, recommending scheduled downtime and targeted maintenance of cooling fans and redistribution of load to mitigate hotspot formation.

Across all these case studies, our unified framework excels in speed and accuracy: optimized indexes yield sub-100 ms retrieval latencies, while CoT reasoning consistently produces expert-grade classifications and coherent explanations. The zero-shot design obviates the need for labeled data or fine-tuning, yet maintains high interpretability by quoting authoritative snippets and laying out each inference step in natural-language paragraphs. Furthermore, the lightweight JSON-based knowledge base supports incremental updates, ensuring the system remains aligned with evolving standards and field data. Together, these features deliver an intelligent QA tool that is both reliable and transparent, directly addressing the key challenges of specialized defect diagnosis in power equipment maintenance.

5. Conclusions

We present a zero-shot grading method for power equipment defects that combines a hierarchical domain knowledge base, a two-step retrieval process, and Chain-of-Thought prompting to enhance both transparency and accuracy—all without any model fine-tuning. On a held-out test set of 218 inspection records, the system achieves a classification accuracy of 54.2%, outperforming a fine-tuned BERT model by 6.0 percentage points; an Explanation Coherence Score of 4.2; a retrieval latency of 28.3 ms; and an inference time of 5.46 s. Ablation and sensitivity studies confirm that setting $k = 30$ in the fine retrieval stage yields the best balance between accuracy and responsiveness; human expert evaluations average 4.1 for Usefulness and 4.3 for Trustworthiness.

Nevertheless, our study has several limitations. First, evaluation on a relatively small, text-only dataset may not fully capture the diversity of real-world defect scenarios, potentially limiting generalizability. Second, the hierarchical knowledge base requires extensive manual curation and may struggle to accommodate novel or rare defect types without further expansion. Third, although our current retrieval and inference times meet offline analysis requirements, stricter real-time monitoring applications may demand additional latency optimizations.

To address these gaps, future research will (1) enlarge and diversify our test corpus by incorporating inspection logs from multiple grid operators; (2) integrate multimodal

sensor metadata—such as thermal readings, acoustic signals, and imagery—into both retrieval and reasoning pipelines to improve situational awareness; (3) explore active and continual learning strategies that adapt to emerging failure modes; (4) develop lightweight, automated metrics for assessing explanation validity to reduce dependence on expert annotation; and (5) further refine system architecture and indexing strategies to meet the stringent low-latency requirements of live maintenance environments.

Author Contributions: Conceptualization, J.D. and Z.C.; Methodology, J.D. and Z.C.; Validation, J.D. and Z.C.; Data curation, B.L., Z.C., L.S., P.L. and Z.R.; Writing—original draft, Z.C.; Writing—review and editing, Z.C.; Funding acquisition, B.L., L.S., P.L. and Z.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a Science and Technology Project from the Big Data Center of State Grid Corporation of China (No. SGSJ0000SJJS2400022).

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: All authors were employed by the company State Grid Corporation of China. The authors declare no conflicts of interest. The authors declare that this study received funding from State Grid Corporation of China. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

References

1. Yang, T.; Chang, L.; Yan, J.; Li, J.; Wang, Z.; Zhang, K. A survey on foundation-model-based industrial defect detection. *arXiv* **2025**, arXiv:2502.19106.
2. Gao, X.; Wang, X.; Chen, Z.; Zhou, W.; Hoi, S.C.H. Knowledge Enhanced Vision and Language Model for Multi-Modal Fake News Detection. *IEEE Trans. Multim.* **2024**, *26*, 8312–8322. [CrossRef]
3. Jia, J.; Fu, H.; Zhang, Z.; Yang, J. Diagnosis of power operation and maintenance records based on pre-training model and prompt learning. In Proceedings of the 2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), Chizhou, China, 14–18 October 2022; pp. 58–61.
4. Wang, C.; Gao, X.; Wu, M.; Lam, S.K.; He, S.; Tiwari, P. Looking Clearer with Text: A Hierarchical Context Blending Network for Occluded Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 4296–4307. [CrossRef]
5. Gao, X.; Chen, Z.; Wei, J.; Wang, R.; Zhao, Z. Deep Mutual Distillation for Unsupervised Domain Adaptation Person Re-Identification. *IEEE Trans. Multim.* **2025**, *27*, 1059–1071. [CrossRef]
6. Wang, C.; Cao, R.; Wang, R. Learning discriminative topological structure information representation for 2D shape and social network classification via persistent homology. *Knowl.-Based Syst.* **2025**, *311*, 113125. [CrossRef]
7. Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceedings of the 10th European Conference on Machine Learning (ECML), Chemnitz, Germany, 21–23 April 1998; pp. 137–142.
8. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
9. Gao, X.; Li, Z.; Shi, H.; Chen, Z.; Zhao, P. Scribble-Supervised Video Object Segmentation via Scribble Enhancement. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 2999–3012. [CrossRef]
10. Xu, Y.; Ishii, E.; Cahyawijaya, S.; Liu, Z.; Winata, G.I.; Madotto, A.; Su, D.; Fung, P. Retrieval-free knowledge-grounded dialogue response generation with adapters. *arXiv* **2021**, arXiv:2105.06232.
11. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst. NeurIPS* **2013**, *26*, 3111–3119.
12. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
13. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1746–1751.
14. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

16. Gururangan, S.; Marasovic, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; Smith, N.A. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 8342–8360.
17. 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 American Chapter of the Association for Computational Linguistics (NAACL), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
18. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
19. 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. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–12 December 2020; pp. 1877–1901.
20. Ye, J.; Gao, J.; Li, Q.; Xu, H.; Feng, J.; Wu, Z.; Yu, T.; Kong, L. ZeroGen: Efficient Zero-Shot Learning via Dataset Generation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 11653–11669.
21. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-Efficient Transfer Learning for NLP. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 2790–2799.
22. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]
23. Bordes, A.; Chopra, S.; Weston, J. Question Answering with Subgraph Embeddings. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 24–29 October 2014; Moschitti, A., Pang, B., Daelemans, W., Eds.; pp. 615–620. [CrossRef]
24. Ren, H.; Dai, H.; Dai, B.; Chen, X.; Yasunaga, M.; Sun, H.; Schuurmans, D.; Leskovec, J.; Zhou, D. LEGO: Latent Execution-Guided Reasoning for Multi-Hop Question Answering on Knowledge Graphs. In Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR, Online, 18–24 July 2021; Volume 139, pp. 8959–8970.
25. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–12 December 2020; Volume 33, pp. 9459–9474.
26. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M.W. Retrieval-Augmented Language Model Pretraining. In Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, Online, 13–18 July 2020; Volume 19, pp. 3929–3938.
27. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.T. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 6769–6781.
28. Yates, A.; Nogueira, R.; Lin, J. Pretrained Transformers for Text Ranking: BERT and Beyond. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), New York, NY, USA, 11–15 July 2021; pp. 2666–2668. [CrossRef]
29. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Online, 25–30 July 2020; pp. 39–48.
30. Genesis, J. Retrieval-Augmented Text Generation: Methods, Challenges, and Applications. *Preprints* **2025**. [CrossRef]
31. Yuan, X.; Shen, C.; Yan, S.; Zhang, X.; Xie, L.; Wang, W.; Guan, R.; Wang, Y.; Ye, J. Instance-adaptive zero-shot chain-of-thought prompting. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 125469–125486.
32. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LO, USA, 28 November–9 December 2022; Volume 35, pp. 24824–24837.
33. Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; Narasimhan, K. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023; Article No.: 517, pp. 11809–11822.
34. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023; pp. 1–33.

35. *GB 50059-2011*; Code for Design of 35 kV ~110 kV Substation. China Planning Press: Beijing, China, 2011.
36. Lu, M.; Chai, Y.; Xu, K.; Chen, W.; Ao, F.; Ji, W. Multimodal fusion and knowledge distillation for improved anomaly detection. *The Visual Computer.* **2025**, *41*, 5311–5322. [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.