TRACE BEFORE YOU CONCLUDE: LEARNING EVIDENTIAL REASONING FOR INDUSTRIAL MAINTENANCE REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The signals-to-semantics (S2S) paradigm bridges the modality gap between industrial time series and large language models (LLMs) by translating sensor dynamics into textual descriptions for maintenance report generation. This semantic interface, however, leaves a distinct learning problem: how to trace report claims to supporting observations and how to leverage that evidence to improve root-cause analysis and maintenance recommendations. We introduce TRACE-S2S, a framework that grounds a domain-structured reasoning chain in temporal semantic evidence. The reasoning chain consists of coarse root-cause identification, fine-grained cause analysis, and maintenance action recommendation. Guided by an external maintenance knowledge graph, an Evidence Tracing Policy constructs stage-specific evidence packages that directly condition a frozen reasoner. The resulting reports preserve explicit links from their conclusions to supporting observations and knowledge. Evidence tracing therefore controls how conclusions are formed, rather than merely explaining them afterward. We optimize only the evidence policy through downstream reinforcement learning, leveraging fault labels and sparse expert report annotations to reward accurate root-cause diagnosis and appropriate maintenance actions. Training requires no annotations of evidence links, and both the S2S front-end and report reasoner remain frozen. The policy can additionally inspect previously uninspected signals when evidence is insufficient. We evaluate TRACE-S2S on a multiphase-flow process and a real-world thermal-power process, assessing diagnostic accuracy, maintenance-action quality, report factuality, and evidence traceability. On average, TRACE-S2S achieves the evidence tracing F1 of 71.69%, outperforming the strongest baseline by 31.50%.

1 INTRODUCTION

In process industries such as power generation and chemical production, maintenance requires more than identifying a fault. Engineers need reports that make the reasoning process interpretable, explaining how observations support fault root-cause conclusions and why the recommended maintenance actions follow from them. Recent work incorporates time series as an input modality for large language models (LLMs), enabling language-based understanding and reasoning (Xie et al., 2025; Hou et al., 2026; Jin et al., 2024). The signals-to-semantics (S2S) framework instead bridges the modality gap through language: it translates sensor dynamics into multi-level temporal descriptions and uses knowledge-guided, multi-turn reasoning to generate maintenance reports (Yue et al., 2026).

However, making observations understandable to an LLM does not establish how they justify a maintenance conclusion. A human expert report usually links observed abnormalities, coarse fault cause, fine fault cause, and maintenance actions through an evidence-supported reasoning chain. The temporal semantic units produced by S2S, termed *Effects*, describe signal behavior. *In fact, the same descriptions can have different evidential value at different stages of this chain*. Evidence useful for coarse diagnosis may be insufficient for distinguishing fine causes or selecting maintenance actions. As the diagnosis narrows, the evidence needed to justify the next decision changes. The challenge is to learn which Effects to leverage at each stage in light of preceding diagnostic results, so that evidence tracing improves fault root-cause analysis and maintenance recommendations while keeping the reasoning process traceable.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Time-series models such as TimesNet and MOMENT address classification, forecasting, and related tasks (Wu et al., 2023; Goswami et al., 2024; Chen & Zhao, 2025), rather than maintenance report generation. Reasoning agents and adaptive retrieval methods obtain external information during inference (Yao et al., 2023; Jiang et al., 2023; Asai et al., 2024), while Search-R1 learns search behavior from downstream rewards (Jin et al., 2025). Industrial maintenance poses a further challenge: domain knowledge constrains the decision chain, and prior decisions affect the evidence needed for subsequent analysis. Learning how to leverage temporal evidence across these dependent decisions, with the quality of the final report as the objective, is the focus of this work.

We introduce TRACE-S2S, which learns *evidential reasoning* through temporal semantic evidence tracing. The evidence path is a learned intermediate representation between signal understanding and report generation. Unlike open-ended text generation, maintenance diagnosis and recommendations are constrained by equipment mechanisms and established procedures. We therefore use an external maintenance knowledge graph to regulate coarse root-cause identification, fine-grained cause analysis, and maintenance action recommendation. Within this structure, a shared Evidence Tracing Policy (ETP) first selects Effects for comparing coarse root-cause candidates, then revisits the Effect pool in the context of the selected coarse cause to support fine-cause analysis and maintenance decisions. These selections form two evidence packages, one for each reasoning stage. The frozen reasoner receives event observations only through these packages, making the learned evidence path part of how conclusions are formed. Reports retain links to the selected Effects, allowing conclusions to be traced to the corresponding signal intervals. The policy can additionally query uninspected signals when the available evidence is insufficient.

We learn these evidence paths from the quality of the conclusions they produce, without annotated Effect–claim links. Group Relative Policy Optimization (Shao et al., 2024) compares alternative paths for the same event using fault labels and sparse expert annotations of fine causes and maintenance actions. A hierarchical reward credits fine-cause and action quality only when the coarse diagnosis is correct, so that plausible downstream text cannot compensate for an incorrect coarse diagnosis. Only the ETP is updated; the S2S front-end and report reasoner remain frozen. An accompanying error-decomposition analysis distinguishes the evidence exposed to the reasoner from its sufficiency for the target conclusions. In fact, our focus is evidence-grounded maintenance reporting rather than fault classification alone.

Contributions. *(1) Formulation.* We formulate industrial maintenance report generation as *temporal semantic evidence tracing*: temporal Effects are organized along a domain-structured chain of coarse-to-fine cause analysis and maintenance recommendation. This makes the evidence path an intermediate representation that both supports report generation and preserves the connection between conclusions and their underlying observations.

(2) Framework. We propose TRACE-S2S, which learns to select evidence according to the changing needs of a diagnostic and maintenance reasoning chain. Its ETP allows different reasoning stages to draw on distinct yet effective evidence from temporal observations. These evidence choices directly guide the frozen reasoner and are optimized to improve the root-cause conclusions and maintenance recommendations in the final report.

(3) Training. We optimize evidence tracing across the complete reasoning chain with fault labels and sparse expert reports. This requires no annotations of individual evidence links and updates only the ETP, with the S2S front-end and report reasoner remaining frozen.

2 Related Work

Temporal understanding and knowledge-grounded reasoning. Time-series models develop representations for classification and forecasting (Wu et al., 2023; Goswami et al., 2024). Time-LLM reprograms inputs for forecasting with a frozen LLM, ChatTS aligns temporal inputs with language, and S2S translates signals for knowledge-guided maintenance reporting (Jin et al., 2024; Xie et al., 2025; Yue et al., 2026). These interfaces make temporal information accessible; they do not by themselves determine its evidential role at each decision stage. CoT introduces intermediate reasoning steps (Wei et al., 2022), while RoG and Think-on-Graph use KG paths to guide reasoning (Luo et al., 2024; Sun et al., 2024). Our focus is not constructing another semantic interface, but learning which event observations should support successive diagnostic and maintenance decisions.

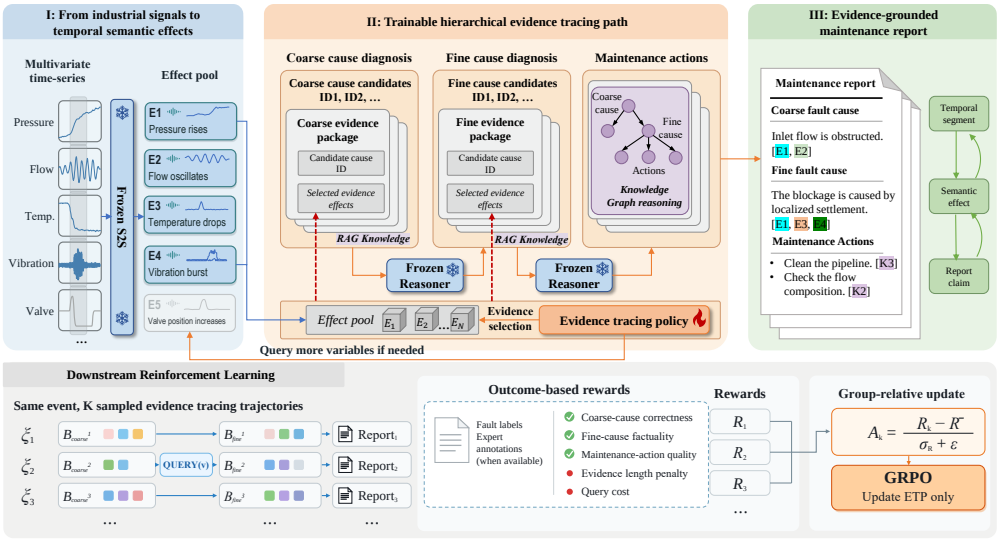

Figure 1: **Overview of TRACE-S2S.** Frozen S2S translates industrial signals into temporal Effects. The ETP constructs coarse- and fine-level evidence packages along a prescribed root-cause-to-action reasoning chain. An external maintenance KG specifies admissible fine causes and actions, while a frozen LLM generates an evidence-grounded report. GRPO updates only the ETP using downstream diagnosis and report feedback. Additional variables can be queried when the current evidence is insufficient.

Learned evidence selection and supported generation. Neural rationalization learns predictive subsets without rationale labels (Lei et al., 2016). RECOMP optimizes compressed context, and query rewriting trains an upstream rewriter from frozen-reader feedback (Xu et al., 2024; Ma et al., 2023). ReAct, FLARE, Self-RAG, and Search-R1 adapt or learn information access during reasoning (Yao et al., 2023; Jiang et al., 2023; Asai et al., 2024; Jin et al., 2025). ALCE, RARR, and fine-grained citation rewards address answer support and source references (Gao et al., 2023b;a; Huang et al., 2024). The distinction between our work and them is not learning to retrieve or filter information, but learning how temporal observations support different, dependent maintenance conclusions and decisions.

3 TRACE-S2S

3.1 Problem Setup and Temporal Effects

Let $X \in \mathbb{R}^{T\times V}$ represent an industrial event with T time steps and V variables. Denote the variable index set by $\mathcal{V} = \{1, \dots, V\}$ and the initially inspected subset by $\mathcal{V}_0 \subseteq \mathcal{V}$. A frozen S2S module (Yue et al., 2026) translates these variables into N_0 temporal semantic units, forming the initial Effect pool $\mathcal{E}_0 = \{E_i\}_{i=1}^{N_0}$, where

$$E_i = (v_i, z_i, \tau_i), \qquad \tau_i \in \{[t_i^{\mathrm{s}}, t_i^{\mathrm{e}}], \varnothing\}. \tag{1}$$

Here $v_i \in \mathcal{V}$ identifies the variable, z_i is its natural-language temporal description, and τ_i gives the interval from start time t_i^{s} to end time t_i^{e}. For a global description, $\tau_i = \varnothing$, and the Effect is associated with the corresponding variable over the event. Each Effect retains a unique ID for subsequent selection and citation.

The task provides J coarse root-cause candidates $\mathcal{C}^c = \{C_j^c\}_{j=1}^{J}$ and an external, fixed maintenance knowledge graph $\mathcal{K}$. For each candidate C_j^c, the subtree $\mathcal{T}(C_j^c)$ specifies admissible fine causes and maintenance actions, together with their references. We consider maintenance events within this supplied cause-and-action space. The output $(\widehat{C}^c, \widehat{C}^f, \widehat{\mathcal{A}}, \widehat{R}, \widehat{\mathcal{G}})$ comprises a coarse cause, a fine cause, an action set, a maintenance report, and its evidence trace graph. Hats denote predictions; stars denote reference annotations. Training uses a coarse-cause label C^{c*} for each event and, when available, expert fine-cause annotations C^{f*} and maintenance actions $\mathcal{A}^*$.

3.2 CANDIDATE-CONDITIONED EVIDENCE TRACING

Evidence requirements change along the reasoning chain: observations that distinguish coarse causes may differ from those needed to resolve a fine cause or justify an action. We therefore use a shared Evidence Tracing Policy (ETP), π_θ, to construct two stage-specific evidence packages. At interaction step t, its input contains the current Effect pool $\mathcal{E}_t$, candidate causes, retrieved KG context, preceding decisions, and the remaining query budget b_t. The policy assigns existing Effect IDs to the relevant candidates, making the evidence relationships explicit before downstream reasoning. The detailed framework is illustrated in Fig. 1.

Coarse evidence. For candidate C_j^c, let $S_j \subseteq \mathcal{E}_t$ denote the selected Effects and $\mathcal{K}_j$ the knowledge context retrieved from $\mathcal{K}$ for that candidate. The coarse package is

$$B_{\text{coarse}} = \left\{(C_j^c, S_j, \mathcal{K}_j)\right\}_{j=1}^{J}, \qquad \widehat{C}^c = F_\phi^c(B_{\text{coarse}}). \tag{2}$$

Here F_ϕ^c is a call to a frozen LLM with parameters ϕ, driven by a diagnosis prompt that compares the supplied candidates and returns one coarse-cause ID. Candidate-conditioned selection allows the same Effect to participate in different diagnostic comparisons rather than imposing a single undifferentiated evidence subset.

Fine causes, actions, and reporting. After coarse diagnosis, $\mathcal{K}$ exposes $\mathcal{T}(\widehat{C}^c)$. The ETP selects available Effects for fine-cause comparison and action recommendation, recording their candidate associations in an indexed evidence collection S_f. The fine package and output are

$$\begin{aligned} B_{\text{fine}} &= \big(\widehat{C}^c, \mathcal{T}(\widehat{C}^c), S_f\big), \\ (\widehat{C}^f, \widehat{\mathcal{A}}, \widehat{R}) &= F_\phi^f(B_{\text{fine}}). \end{aligned} \tag{3}$$

F_ϕ^f uses the same frozen LLM under a fine-cause and report-generation prompt, selecting causes and actions from the supplied subtree. Diagnostic claims cite Effects; procedure-based recommendations cite KG nodes, with event-specific applicability supported by observations. Both packages, selected IDs, knowledge references, query records, and report citations are stored in $\widehat{\mathcal{G}}$. Each reasoning call receives only its serialized package and fixed task instructions. The resulting claim-to-Effect paths continue to signal intervals when available, or to the corresponding variables for global Effects.

On-demand variable inspection. When evidence is insufficient, the ETP additionally selects $a_t^q \in \{\text{STOP}\} \cup \{\text{QUERY}(v) : v \in \mathcal{V}_t^{\text{uninspected}}\}$, where a_t^q is its inspection decision and $\mathcal{V}_t^{\text{uninspected}}$ contains uninspected variables associated with the current candidate causes in $\mathcal{K}$. A query reveals $\mathcal{E}(v) = F_{\text{S2S}}^E(X_{:,v})$, the Effects produced by frozen S2S from the recorded sequence of variable v, and updates

$$\mathcal{E}_{t+1} = \mathcal{E}_t \cup \mathcal{E}(v), \qquad b_{t+1} = b_t - 1. \tag{4}$$

The current package is then reconstructed. The initial budget $b_0 = Q_{\max}$ is shared across both stages. A STOP completes the current stage, leaving unused budget available to the next. During training, queries reveal precomputed Effects; unrevealed entries remain outside all policy and reasoner inputs.

3.3 LEARNING THE EVIDENCE TRACING POLICY

The ETP is an instruction-tuned LLM with trainable LoRA parameters θ (Hu et al., 2022). Its output follows a prescribed schema for Effect IDs and inspection decisions. S2S, the external KG, the downstream reasoner, and the reward evaluator remain fixed.

Outcome-based rewards. A trajectory ξ_k records the kth sequence of policy decisions, evidence packages, diagnoses, and report. Its coarse reward is $r_{c,k} = \mathbb{I}[\widehat{C}_k^c = C^{c*}]$, where $\mathbb{I}[\cdot]$ is the indicator function. A fixed evaluator compares the generated fine-cause statements and maintenance recommendations with the expert annotations. The fine-cause score $r_{f,k} \in [0, 1]$ is based on precision and recall over semantically matched atomic facts, with a penalty for contradictions. The action score $r_{a,k} \in [0, 1]$ evaluates coverage of the reference actions and applicability to the diagnosed fault,

with penalties for conflicting recommendations. Each score is evaluated only when its reference annotation is available and is set to zero otherwise. The reward is

$$\mathcal{R}_k = r_{c,k} + r_{c,k}(r_{f,k} + r_{a,k}) - \lambda_E \frac{N_{E,k}}{\max(1, N_{E,k}^{\text{avail}})} - \lambda_Q N_{Q,k}. \tag{5}$$

We count distinct selected Effect IDs across the two completed packages as $N_{E,k}$; $N_{E,k}^{\text{avail}}$ is the number of distinct Effects in the final visible pool, and $N_{Q,k}$ is the number of executed queries. Gating the fine-cause and action rewards by $r_{c,k}$ prevents plausible downstream text from compensating for an incorrect coarse diagnosis. Expert report supervision enters through its root-cause and action content, while the last two terms discourage unnecessary evidence accumulation.

Group-relative optimization. For each event, we sample K trajectories using a fixed rollout snapshot $\pi_{\theta_{\text{old}}}$. Their group-relative advantages are $A_k = (\mathcal{R}_k - \overline{\mathcal{R}})/(\sigma_{\mathcal{R}} + \epsilon)$, where $\overline{\mathcal{R}}$ and $\sigma_{\mathcal{R}}$ are the group reward mean and standard deviation, and $\epsilon > 0$ ensures numerical stability. Let L_k be the number of ETP-generated tokens in ξ_k, $u_{k,\ell}$ its ℓth such token, and $h_{k,\ell}$ the preceding interaction history. Define the token probability ratio

$$\rho_{k,\ell} = \frac{\pi_\theta(u_{k,\ell} \mid h_{k,\ell})}{\pi_{\theta_{\text{old}}}(u_{k,\ell} \mid h_{k,\ell})}. \tag{6}$$

We optimize the clipped GRPO objective

$$\begin{aligned}\mathcal{L}_{\text{GRPO}} = & -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{L_k}\sum_{\ell=1}^{L_k}\min[\rho_{k,\ell}A_k, \text{clip}(\rho_{k,\ell}, 1-\delta, 1+\delta)A_k] \\ & + \beta\,\overline{D_{\text{KL}}}(\pi_\theta \| \pi_{\text{ref}}),\end{aligned} \tag{7}$$

where $\delta > 0$ is the clipping threshold, π_{ref} is the fixed initial policy, and $\beta \geq 0$ weights the KL divergence averaged over policy-token contexts. Only ETP-generated tokens contribute to the policy loss; tool responses and frozen-reasoner outputs provide subsequent context (Shao et al., 2024; Jin et al., 2025). Invalid structured actions receive a fixed negative reward and terminate the rollout. At test time, the policy generates one deterministic trajectory without reward evaluation or parameter updates. Algorithmic details appear in Appendix B.

4 A Theoretical Error-Decomposition View of Evidence Tracing

Evidence selection should retain decision-relevant information while making it usable by a frozen reasoner. We separate these two requirements to explain when selection can help and why its value depends on the reasoning stage.

For stage $s \in \{c, f\}$, let $Y_c^* = C^{c*}$ and $Y_f^* = (C^{f*}, \mathcal{A}^*, F_R^*)$, where F_R^* denotes reference report facts. Write $W_s = (\mathcal{K}, H_s)$, with $H_c = \varnothing$ and $H_f = C^{c*}$; the fine-stage analysis considers a correct coarse parent. Let $\mathcal{E}$ contain all obtainable Effects and B be a complete recorded evidence package. For the frozen predictive distribution $q_{\phi,s}$, define $\mathcal{L}_{\phi,s}(B) = \mathbb{E}[-\log q_{\phi,s}(Y_s^* \mid B, W_s)]$. Assuming finite relevant quantities and $I(Y_s^*; B \mid \mathcal{E}, W_s) = 0$, this expected log loss decomposes as

$$\mathcal{L}_{\phi,s}(B) = \underbrace{H(Y_s^* \mid \mathcal{E}, W_s)}_{\text{full-observation uncertainty}} + \underbrace{\Gamma_s(B)}_{\text{omitted information}} + \underbrace{\mathcal{D}_{\phi,s}(B)}_{\text{reasoner mismatch}}, \tag{8}$$

where $\Gamma_s(B) = I(Y_s^*; \mathcal{E} \mid B, W_s)$ measures omitted task information, and

$$\mathcal{D}_{\phi,s}(B) = \mathbb{E}_{B,W_s}[D_{\text{KL}}(p(\cdot \mid B, W_s) \| q_{\phi,s}(\cdot \mid B, W_s))] \tag{9}$$

measures the discrepancy between the true conditional target distribution p and the frozen reasoner. Here H, I, and D_{KL} denote conditional entropy, conditional mutual information, and KL divergence, respectively.

For a matched package B_{full} containing all Effects under the same task context, $\Gamma_s(B_{\text{full}}) = 0$, giving

$$\begin{aligned}\mathcal{L}_{\phi,s}(B) - \mathcal{L}_{\phi,s}(B_{\text{full}}) = \Gamma_s(B) + \mathcal{D}_{\phi,s}(B) \\ - \mathcal{D}_{\phi,s}(B_{\text{full}}).\end{aligned} \tag{10}$$

Thus selection reduces expected log loss when its reduction in reasoner mismatch exceeds the information it omits. Selection cannot create event information, but organizing retained evidence can make it more usable by an imperfect reasoner. Both terms depend on the target Y_s^* and diagnostic context W_s: evidence adequate for coarse diagnosis may omit distinctions needed for fine causes or maintenance actions. This motivates revisiting the Effect pool after coarse diagnosis instead of reusing one relevance ranking throughout. The diagnosis-and-action reward is a task-level surrogate for downstream quality, not an estimator of these information quantities. Appendix A provides the derivation.

5 EXPERIMENTS

We evaluate TRACE-S2S on a real-world thermal-power dataset and a multiphase-flow dataset.

5.1 DATASETS

Thermal-power process. The dataset contains eight real fault cases provided by our industrial partner, together with multivariate operating signals and expert maintenance reports. The cases cover gas-turbine systems, pumps, and heat-recovery steam generators. Detailed fault descriptions are provided in Appendix C.1.

Multiphase-flow process. The Cranfield multiphase-flow dataset contains multivariate process measurements and fault labels collected from a controlled experimental facility (Ruiz-Cárcel et al., 2015).

5.2 EVALUATION PROTOCOLS, BASELINES, AND METRICS

Evaluation protocols. For thermal power, *Protocol-S* evaluates seen faults, whose categories are represented in the training data. Full sequences are partitioned into training, validation, and test sets before reporting windows are extracted. *Protocol-U* simulates generalized zero-shot diagnosis by withholding a fault category's signals and reports from training while retaining both seen and unseen categories in the test candidate set. The maintenance KG provides knowledge for both groups of candidates. Data preparation and split details appear in Appendix C.2.

Baselines. We compare three families with progressively broader output capabilities, from fault diagnosis to report generation and explicit evidence tracing. (i) *Traditional diagnosis models*: supervised models TCN, TimesNet, MOMENT, and Transformer–LSTM-DAE are evaluated under Protocol-S (Bai et al., 2018; Wu et al., 2023; Goswami et al., 2024; Guo et al., 2026); zero-shot diagnosis methods FDAT, SSHTN, and TPF are evaluated under Protocol-U (Feng & Zhao, 2021; Yue et al., 2024; Wang et al., 2026). (ii) *Direct report generators*: ChatTS-ZS uses ChatTS through prompting without task-specific fine-tuning, whereas ChatTS-SFT applies supervised fine-tuning using the training signals and expert reports available under each protocol (Xie et al., 2025). Original S2S uses knowledge-guided, multi-turn reasoning to generate maintenance reports (Yue et al., 2026). (iii) *Evidence-based generators*: S2S-KG-Full is a prompt-only variant that asks S2S to generate reports with explicit evidence references, supplying all currently visible Effects and retrieved KG context to the same frozen reasoner. ReAct-Query uses ReAct-style prompts to alternate reasoning with evidence retrieval (Yao et al., 2023). Search-R1-style replaces Search-R1's search engine with the Effect pool and learns evidence retrieval from downstream rewards (Jin et al., 2025). Further adaptations appear in Appendix C.3.

Metrics. Coarse and Fine Accuracy evaluate root-cause correctness; Action F1 measures agreement with reference maintenance actions; Report Fact F1 measures agreement between generated and reference report facts; and Evidence Precision/Recall/F1 evaluates the reported Effect–claim links. Diagnosis models are evaluated using Coarse Accuracy. Direct report generators are evaluated using Coarse Accuracy, Fine Accuracy, Action F1, and Report Fact F1. Evidence-based generators and TRACE-S2S are evaluated using all seven metrics. Dashes indicate metrics outside a method's evaluated output scope. All scores are percentages; details appear in Appendix C.3.

5.3 COMPARATIVE RESULTS

Table 1: Thermal-power results on seen faults (Protocol-S). Scores are percentages. Best values are bold; dashes refer to metrics that are beyond the models' capability.

	Diagnosis and maintenance			Evidence tracing			Report
Method	Coarse Acc.	Fine Acc.	Action F1	P	R	F1	Fact F1
Seen-fault diagnosis							
TCN	92.96	—	—	—	—	—	—
TimesNet	84.82	—	—	—	—	—	—
MOMENT	88.91	—	—	—	—	—	—
Transformer–LSTM-DAE	90.81	—	—	—	—	—	—
Direct report generation							
ChatTS-ZS	45.83	8.33	21.88	—	—	—	19.03
ChatTS-SFT	**100.00**	73.29	67.21	—	—	—	56.27
Original S2S	85.42	58.33	31.18	—	—	—	60.31
Evidence-based report generation							
S2S-KG-Full	74.58	60.42	42.92	32.59	59.17	42.03	59.45
ReAct-Query	79.17	70.83	57.36	45.07	25.12	32.25	57.94
Search-R1-style	85.42	71.42	59.31	41.73	32.17	36.33	60.49
TRACE-S2S	**100.00**	**87.50**	**73.33**	**74.48**	**76.34**	**75.39**	**69.12**

Table 2: Multiphase-flow results on seen faults (Protocol-S). Scores are percentages. Best values are bold; dashes refer to metrics that are beyond the models' capability.

	Diagnosis and maintenance			Evidence tracing		
Method	Coarse Acc.	Fine Acc.	Action F1	P	R	F1
Seen-fault diagnosis						
TCN	87.53	—	—	—	—	—
TimesNet	92.38	—	—	—	—	—
MOMENT	81.69	—	—	—	—	—
Transformer-LSTM-DAE	85.28	—	—	—	—	—
Direct report generation						
ChatTS-ZS	30.77	17.96	0.00	—	—	—
ChatTS-SFT	**100.00**	73.08	63.59	—	—	—
Original-S2S	84.62	23.08	18.33	—	—	—
Evidence-based report generation						
S2S-KG-Full	76.92	38.46	30.26	15.59	59.05	24.67
ReAct-Query	57.69	19.23	16.41	15.36	61.25	24.56
Search-R1-style	80.77	42.39	32.82	27.21	**64.88**	38.34
TRACE-S2S	**100.00**	**84.62**	**76.92**	**82.09**	61.34	**67.98**

From correct diagnosis to reliable maintenance reporting. Tables 1 and 2 show that TRACE-S2S and ChatTS-SFT both achieve 100% coarse accuracy on the two datasets, while their subsequent decisions differ. On thermal power, TRACE-S2S improves Fine Accuracy and Action F1 over ChatTS-SFT by 14.21 and 6.12 percentage points, respectively, and raises Report Fact F1 from 56.27% to 69.12%. On some metrics, S2S-KG-Full is even lower than Original-S2S, which is consistent with the previous theoretical analysis: full, unfiltered information may reduce reasoning accuracy. On multiphase flow, Fine Accuracy increases from 73.08% to 84.62%, and Action F1 from 63.59% to 76.92%. Correct coarse diagnosis therefore leaves substantial room for improving finer maintenance conclusions. For multiphase flow, fine-cause and action metrics use mechanism-based, expert-reviewed structured annotations (Appendix C.1). Report Fact F1 is not evaluated because the original dataset does not provide expert maintenance reports.

Table 3: Thermal-power results on knowledge-covered unseen faults (Protocol-U). Scores are percentages. Best values are bold; dashes refer to metrics that are beyond the models' capability.

	Diagnosis and maintenance			Evidence tracing			Report
Method	Coarse Acc.	Fine Acc.	Action F1	P	R	F1	Fact F1
Zero-shot diagnosis							
FDAT	0.00	—	—	—	—	—	—
SSHTN	35.60	—	—	—	—	—	—
TPF	41.67	—	—	—	—	—	—
Direct report generation							
ChatTS-ZS	32.64	5.36	12.92	—	—	—	18.28
ChatTS-SFT	40.32	36.72	35.43	—	—	—	45.45
Original S2S	83.33	54.17	22.92	—	—	—	46.88
Evidence-based report generation							
S2S-KG-Full	70.83	48.33	18.75	23.92	29.53	26.43	42.69
ReAct-Query	72.31	18.33	17.43	21.56	22.50	22.02	31.84
Search-R1-style	83.33	31.25	25.23	30.23	28.92	29.56	42.32
TRACE-S2S	**100.00**	**68.75**	**65.83**	**45.32**	**78.87**	**57.56**	**68.33**

Evidence use and report quality. On thermal power, TRACE-S2S improves Evidence F1 over S2S-KG-Full from 42.03% to 75.39% and Report Fact F1 from 59.45% to 69.12%. On multiphase flow, Evidence F1 reaches 67.98%, compared with 38.34% for Search-R1-style, the strongest alternative on this metric. The improvement is particularly pronounced in evidence precision, with comparable recall. These results show gains in matching conclusions to their supporting observations alongside improvements in fine-cause and maintenance decisions.

Knowledge-covered unseen faults. Under Protocol-U (Table 3), TRACE-S2S achieves 100% coarse accuracy, 68.75% Fine Accuracy, and 65.83% Action F1. FDAT suffers from a relatively severe domain shift, causing its accuracy on unseen classes to remain zero. Fine Accuracy exceeds Original S2S by 14.58 percentage points, while Action F1 exceeds ChatTS-SFT by 30.40 points. Report Fact F1 rises from 46.88% for Original S2S, the strongest alternative, to 68.33%. Evidence F1 reaches 57.56%, compared with 29.56% for Search-R1-style. The gains extend to fine-grained conclusions and evidence links when the tested fault category is withheld from policy training but described by the supplied maintenance knowledge.

5.4 Ablation Studies

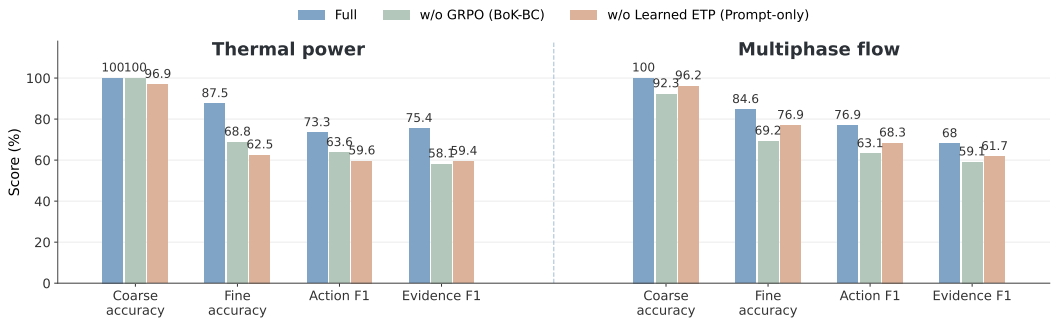

Figure 2: Ablations on thermal power and multiphase flow. BoK-BC replaces GRPO with reward-selected trajectory imitation; Prompt-only uses evidence selection without a learned ETP.

Figure 2 evaluates two key components of TRACE-S2S. Removing GRPO consistently degrades fine-cause diagnosis, maintenance recommendation, and evidence tracing on both datasets, indicating

that downstream reinforcement learning is important for learning effective evidence paths. Replacing the learned ETP with prompt-only selection also causes clear performance drops, particularly in the later reasoning stages. Together, these ablations show that the gains of TRACE-S2S rely on both a trainable evidence policy and outcome-driven optimization, rather than prompting alone.

5.5 Evidence-Traced Case Study

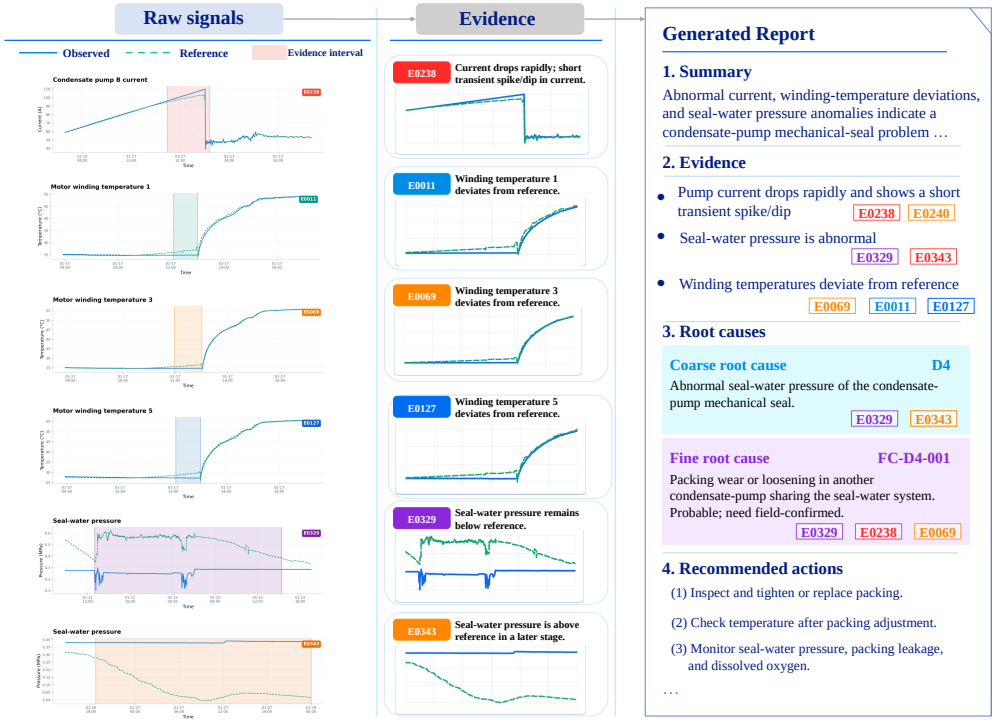

Figure 3: Evidence-traced condensate-pump report. The coarse cause cites seal-water pressure Effects E0329/E0343; the fine-cause interpretation uses E0329 together with current Effect E0238 and temperature Effect E0069. Colored identifiers link report statements, Effects, and signal intervals.

Figure 3 makes stage-dependent evidence use visible. The coarse diagnosis, abnormal seal-water pressure of the condensate-pump mechanical seal, cites pressure Effects E0329/E0343. The fine-cause interpretation retains E0329 and adds the current transient E0238 and winding-temperature deviation E0069. It proposes packing wear or loosening in another pump sharing the seal-water system, subject to field confirmation. Recommendations include packing inspection and adjustment or replacement, followed by temperature and seal-water monitoring. Thus, the report draws on different evidence combinations as its diagnosis becomes more specific, with each citation connected to the recorded signals.

6 Conclusion

We presented TRACE-S2S, which learns how temporal semantic evidence should support successive decisions in industrial maintenance reporting. By constructing stage-specific evidence paths for coarse diagnosis, fine-cause analysis, and maintenance recommendation, TRACE-S2S improves report generation while preserving explicit links between conclusions and their supporting observations. Experiments on a multiphase-flow process and a real-world thermal-power process demonstrate consistent gains in diagnostic, maintenance, report, and evidence-tracing quality. The accompanying analysis interprets evidence selection as a trade-off between omitted task information and the frozen reasoner's ability to use retained evidence effectively.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work to assist with language polishing and to improve the organization, presentation, wording, and readability of the manuscript. In addition, API calls were made to large language models in certain parts of the experiments; the specific models used are stated in the corresponding sections of the manuscript. The authors take full responsibility for the final content of this work.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Jiawei Chen and Chunhui Zhao. Addressing information asymmetry: Deep temporal causality discovery for mixed time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7), 2025.

Liangjun Feng and Chunhui Zhao. Fault description based attribute transfer for zero-sample industrial fault diagnosis. *IEEE Transactions on Industrial Informatics*, 17(3):1852–1862, 2021.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508. Association for Computational Linguistics, 2023a.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023b.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152, 2024.

Lijie Guo, Jiaqi Shi, Jianxin Kang, and Ao Li. An interpretable transformer–lstm denoising autoencoder for semi-supervised fault diagnosis in chemical processes. *Engineering Applications of Artificial Intelligence*, 172:114358, 2026. ISSN 0952-1976.

Difei Hou, Jiaqi Yue, and Chunhui Zhao. Words towards explainability: Caption label-free learning via dual loop agentic time series captioning. In *Forty-third International Conference on Machine Learning*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. Training language models to generate text with citations via fine-grained rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2926–2949. Association for Computational Linguistics, 2024.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1011.

Linhao Luo, Yuan-Fang Li, Reza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*, 2024.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5303–5315, 2023.

C. Ruiz-Cárcel, Y. Cao, D. Mba, L. Lao, and R. T. Samuel. Statistical process monitoring of a multiphase flow facility. *Control Engineering Practice*, 42:74–88, 2015.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. Think-on-Graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*, 2024.

Lv Wang, Junyu Qi, and Yi Qin. Beyond single prototype: A multi-prototype fusion framework with contrastive learning and diffusion model for zero-shot fault diagnosis. *Information Fusion*, 135: 104399, 2026. ISSN 1566-2535.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. ChatTS: Aligning time series with LLMs via synthetic data for enhanced understanding and reasoning. *Proceedings of the VLDB Endowment*, 18(8):2385–2398, 2025.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *International Conference on Learning Representations*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Jiaqi Yue, Jiancheng Zhao, and Chunhui Zhao. Similarity makes difference: SSHTN for generalized zero-shot industrial fault diagnosis by leveraging auxiliary set. *IEEE Transactions on Industrial Informatics*, 20(5):7598–7607, 2024.

Jiaqi Yue, Chunhui Zhao, and Cheng Xu. An on-the-fly signals-to-semantics storytelling framework for generating explainable industrial maintenance decisions. *IEEE Transactions on Automation Science and Engineering*, 23:13573–13584, 2026.

A Derivation of the Evidence-Risk Decomposition

Fix a reasoning stage and write $Y = Y_s^*$ and $W = W_s$. Under the assumptions in Section 4, the conditional cross-entropy identity gives

$$\begin{aligned}\mathcal{L}_{\phi,s}(B) &= H(Y \mid B, W) + \mathbb{E}_{B,W} D_{\mathrm{KL}}(p(\cdot \mid B, W) \| q_{\phi,s}(\cdot \mid B, W)) \\ &= H(Y \mid B, W) + \mathcal{D}_{\phi,s}(B). \end{aligned} \tag{11}$$

Since $I(Y; B \mid \mathcal{E}, W) = 0$,

$$\Gamma_s(B) = I(Y; \mathcal{E} \mid B, W) = H(Y \mid B, W) - H(Y \mid \mathcal{E}, W). \tag{12}$$

Substituting the second identity into the first yields Equation (8). For a complete package B_{full} from which $\mathcal{E}$ is recoverable, $\Gamma_s(B_{\mathrm{full}}) = 0$. Comparing the packages under the same stage and knowledge context therefore gives Equation (10). Evidence selection lowers expected log loss when

$$\mathcal{D}_{\phi,s}(B_{\mathrm{full}}) - \mathcal{D}_{\phi,s}(B) > \Gamma_s(B).$$

The reduction in reasoner mismatch then exceeds the information omitted by selection. The fine-stage comparison conditions on the correct coarse cause, as specified in the main text.

B Algorithmic Details

Algorithms 1 and 2 formalize the evidence-tracing procedure and training objective in Section 3. An Effect ID resolves to its variable, description, and temporal interval. The two evidence packages retain candidate associations and knowledge references. The coarse and fine reasoning calls use the same frozen LLM with different task instructions.

B.1 Evidence Tracing and Report Generation

A query is eligible when its variable is associated with the current candidates, has not been inspected, and $b > 0$. A STOP completes the current stage; unused budget carries over to the fine stage. Each reasoning call receives its own evidence package and fixed task instructions. At inference, the policy uses deterministic decoding and produces one trajectory. Temporal citations resolve to intervals for local Effects and to the corresponding variables for global Effects.

B.2 Policy Optimization

For the formulation in Section 3.3, coarse-cause correctness gates the fine-cause and action rewards. The evidence-length term counts distinct selected IDs across both completed packages, and the query term counts executed inspections. Missing fine-cause or action annotations give a zero value for the corresponding reward. Algorithm 2 uses the reward and clipped GRPO loss defined in the main text.

Tool responses and frozen-reasoner outputs enter subsequent context but do not contribute ETP-token loss terms. A group with identical total rewards has zero centered advantages. The reference KL term is retained in the objective.

Model and optimization settings. The policy backbone is Qwen3-8B. LoRA is applied to its attention projections, with the configuration summarized in Table 4.

C Experimental Details

C.1 Datasets and Maintenance Knowledge

Thermal-power process. The collaborating thermal-power plant provides eight fault cases, their operating signals and expert maintenance reports, and a maintenance knowledge graph (KG). The complete knowledge graph provided by the plant has 903,790 nodes and 1,115,785 edges. From it, we extracted a subgraph related to operations and maintenance, consisting of 1,030 nodes and 2,590 edges. The length corresponding to each case varies, and each includes a fault sampling time series of approximately 10 days in length, with a sampling period of 2 minutes. Table 5 summarizes the coarse and fine causes of these cases.

Algorithm 1 Two-stage evidence tracing as defined in Section 3

Require: Event X, initial variables $\mathcal{V}_0$, candidates $\mathcal{C}^c$, KG $\mathcal{K}$, policy π, budget $Q_{\max}$, decoding mode
$\mathcal{E}_{\text{vis}} \leftarrow F^{E}_{\text{S2S}}(X_{:,\mathcal{V}_0})$; $\mathcal{V}_{\text{seen}} \leftarrow \mathcal{V}_0$; $b \leftarrow Q_{\max}$
Initialize trajectory record ξ
for $s \in \{c, f\}$ **do**
  if $s = c$ **then**
    Retrieve candidate contexts $\{\mathcal{K}_j\}$ for $\mathcal{C}^c$
  else
    Retrieve fine-cause and action candidates from $\mathcal{T}(\widehat{C}^c)$
  end if
  loop
    Form the policy state from $\mathcal{E}_{\text{vis}}$, candidates, knowledge, preceding decision, and b
    Sample or decode selected Effect IDs S_s and inspection decision a^q; record policy tokens
    if IDs, candidate associations, or the inspection decision are invalid **then**
      return ξ with an invalid-action flag
    end if
    if $a^q = \textsc{Query}(v)$ **then**
      Reveal $\mathcal{E}(v)$; $\mathcal{E}_{\text{vis}} \leftarrow \mathcal{E}_{\text{vis}} \cup \mathcal{E}(v)$
      $\mathcal{V}_{\text{seen}} \leftarrow \mathcal{V}_{\text{seen}} \cup \{v\}$; $b \leftarrow b - 1$; record the query
    else
      Construct B_s using Equation (2) or (3); **break**
    end if
  end loop
  if $s = c$ **then**
    $\widehat{C}^c \leftarrow F^c_\phi(B_{\text{coarse}})$
  else
    $(\widehat{C}^f, \widehat{\mathcal{A}}, \widehat{R}) \leftarrow F^f_\phi(B_{\text{fine}})$
  end if
  Record B_s and the corresponding reasoning output in ξ
end for
Build $\widehat{\mathcal{G}}$ from the recorded packages, decisions, and report citations
return ξ, predictions, $\widehat{R}$, and $\widehat{\mathcal{G}}$

Algorithm 2 GRPO optimization of the evidence policy

Require: Training events and annotations, initial policy π_{ref}, group size K, reward and optimizer settings
Initialize π_θ from π_{ref}; make only its LoRA parameters trainable
for each training update **do**
  Fix the rollout snapshot $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$; select a training event x
  for $k = 1, \ldots, K$ **do**
    $\xi_k \leftarrow \textsc{Rollout}(x, \pi_{\theta_{\text{old}}}, \text{sampling})$ using Algorithm 1
    Evaluate $\mathcal{R}_k$ by Equation (5), or assign r_{invalid} to an invalid action
    Retain ETP-generated tokens and their rollout log probabilities
  end for
  Compute $\overline{\mathcal{R}}$, $\sigma_{\mathcal{R}}$ and $A_k = (\mathcal{R}_k - \overline{\mathcal{R}})/(\sigma_{\mathcal{R}} + \epsilon)$
  Update only θ with Equation (7); keep the reference policy and downstream modules fixed
end for
Select a checkpoint on the validation split; freeze it for testing

Table 4: Policy and optimization settings.

Component	Setting
ETP backbone / precision	Qwen3-8B / bfloat16
LoRA projections	Query, key, value, and output projections
LoRA rank / scaling / dropout	16 / 32 / 0.05
Policy sampling / test temperature	0.7 / 0
Frozen reasoning	DeepSeek-V4-Flash; temperature 0
Optimizer / learning rate	AdamW / 1×10^{-5}
Gradient clipping	1.0
GRPO clipping / KL weight	$\delta = 0.15$ / $\beta = 0.04$
Loss weights	$\lambda_E = 0.1$, $\lambda_Q = 0.05$
GRPO group size	$K = 4$
Maximum queries	$Q_{\max} = 3$
Advantage stabilization	$\epsilon = 10^{-8}$

Table 5: Coarse and fine causes of the thermal-power fault cases.

Case	**Coarse cause**	**Fine cause**
No. 1	Abnormal cooling-air pressure in the gas-turbine compressor.	Elevated ambient temperature reduces inlet-air density and mass flow, lowering the compressor pressure ratio and discharge pressure.
No. 2	Abnormal pre-combustion natural-gas temperature in the gas turbine.	The fuel-gas performance heater is taken out of service, interrupting normal gas heating before entry into the combustor.
No. 3	High lower-bearing temperature in the condensate-pump motor.	Hardened grease blocks the lubricant discharge passage, creating a false indication of grease shortage and prompting repeated regreasing.
No. 4	Abnormal seal-water pressure of the condensate-pump mechanical seal.	Worn or loose packing in another condensate pump on the shared seal-water system increases leakage and lowers the common header pressure.
No. 5	Abnormal low-pressure steam flow in the heat-recovery steam generator.	Unequal pressure-tap elevations across a vertically installed flow nozzle introduce an uncompensated hydrostatic head from the condensate columns, biasing the flow measurement.
No. 6	High coupling-end bearing temperature in the feedwater pump.	Extensive wear of the lower shell of the drive-end radial bearing causes localized friction and a rapid rise in bearing-shell temperature.
No. 7	High stator temperature in the circulating-water pump motor.	A loose connection in the stator-winding temperature measurement circuit produces an erroneously high reading at one measurement point.
No. 8	Abnormal acceleration in the gas-turbine combustion chamber.	Damage to the ACC1 signal extension cable leaves the acceleration reading frozen at a fixed low value for an extended period.

Multiphase-flow process. The original multiphase-flow dataset (Ruiz-Cárcel et al., 2015) does not provide maintenance reports; we compiled case-specific coarse causes, fine causes, maintenance actions, and knowledge graphs based on process mechanisms and had them reviewed by experts in multiphase-flow systems.

C.2 Data Preparation and Evaluation Protocols

Protocol-S. For each recorded fault case, the sequences derived from it are allocated to training, validation, and test sets in a 60/20/20 ratio. Reporting windows are extracted after partitioning, and all windows from a sequence remain in the same split. Protocol-S evaluates report generation for cases represented in the training data.

Protocol-U. Protocol-U simulates generalized zero-shot diagnosis: the fault category of the held-out case is treated as unseen during training, while both seen and unseen categories remain in the test candidate set. We perform leave-one-case-out evaluation, with each case serving once as the unseen-category test case. In each fold, the held-out case's signals, derived sequences, and report targets are excluded from ETP training and validation, which use the remaining cases. The validation-selected checkpoint is evaluated on the held-out original sequence. Knowledge describing both seen and unseen categories remains available through the plant-provided KG. Across the folds, testing covers original sequences. The final metric is the average of the harmonic means of seen and unseen across all folds.

C.3 Baseline and Evaluation Definitions

Comparison methods. TCN, TimesNet, MOMENT, and Transformer–LSTM-DAE are evaluated on coarse diagnosis; FDAT, SSHTN, and TPF provide the unseen-fault diagnosis comparisons. ChatTS-ZS uses prompted inference without task-specific fine-tuning. ChatTS-SFT is fine-tuned on the permitted training signals and reports (if available) with a learning rate of 10^{-4}. Its trainable parameters account for approximately 0.1869% of the complete model's parameters, with LoRA parameters included in the denominator. Original S2S retains its prompted maintenance reasoning. S2S-KG-Full prompts the frozen reasoner to produce reports with evidence references from all currently visible Effects and retrieved knowledge. ReAct-Query uses deepseek-flash to alternate prompted reasoning with evidence retrieval. Search-R1-style uses Qwen3-8B with LoRA as its evidence-search agent and deepseek-flash as its frozen reasoner. All models requiring iterative parameter optimization are trained to convergence.

Diagnosis and actions. Coarse Accuracy compares the predicted and reference cause IDs. Fine Accuracy counts a prediction as correct when its fine-cause ID belongs to the expert-approved reference set for the case. Action F1 compares deduplicated predicted and reference action sets. For M matched actions, n predicted actions, and n^* reference actions,

$$P_A = \frac{M}{n}, \qquad R_A = \frac{M}{n^*}, \qquad F1_A = \frac{2M}{n + n^*}.$$

A missing prediction against a nonempty reference receives zero F1.

Evidence annotation. Three experts label Effect–claim relations as Support or Irrelevant using the recorded signals and canonical Effects. Majority agreement and adjudication establish the reference. Inter-annotator agreement is Fleiss' $\kappa = 0.951$. Evidence Precision, Recall, and F1 compare the reported supporting Effect–claim links with the expert reference links. Let $\widehat{\mathcal{L}}$ and $\mathcal{L}^*$ denote these predicted and reference link sets, respectively, and let $M_E = |\widehat{\mathcal{L}} \cap \mathcal{L}^*|$. Then

$$P_E = \frac{M_E}{|\widehat{\mathcal{L}}|}, \qquad R_E = \frac{M_E}{|\mathcal{L}^*|}, \qquad F1_E = \frac{2M_E}{|\widehat{\mathcal{L}}| + |\mathcal{L}^*|}.$$

A match requires both the Effect and its associated claim to agree with the reference.

Report facts. Report Fact F1 evaluates factual content using deepseek-v4-pro, with maintenance actions scored separately. The relevant results have been reviewed by human experts.After deduplication, precision P_R counts matched reference fact groups and supported additional facts as correct, and

unsupported or contradictory statements as errors. Recall R_R measures the proportion of reference fact groups covered. Their harmonic mean gives

$$F1_R = \frac{2P_R R_R}{P_R + R_R}.$$

Neutral text is excluded. Precision is one when no factual content is scored, and recall is one when the reference contains no required facts; F1 is zero when both precision and recall are zero. Scores are aggregated at the case level.