No More Labelled Examples? An Unsupervised Log Parser with LLMs

JUNJIE HUANG, Chinese University of Hong Kong, Hong Kong
ZHIHAN JIANG, Chinese University of Hong Kong, Hong Kong
ZHUANGBIN CHEN[*], Sun Yat-sen University, China
MICHAEL LYU, Chinese University of Hong Kong, Hong Kong

Log parsing serves as an essential prerequisite for various log analysis tasks. Recent advancements in this field have improved parsing accuracy by leveraging the semantics in logs through fine-tuning large language models (LLMs) or learning from in-context demonstrations. However, these methods heavily depend on high-quality labeled examples to achieve optimal performance. In practice, continuously collecting high-quality labeled data is challenging since logs are huge in volume and under frequent evolution, leading to performance degradation or heavy maintenance efforts for existing log parsers after deployment. To address this issue, we propose LUNAR, an unsupervised LLM-based method for efficient and off-the-shelf log parsing. Our key insight is that while LLMs may struggle with direct log parsing, their performance can be significantly enhanced through comparative analysis across multiple logs that differ only in their parameter parts. We refer to such groups of logs as *Log Contrastive Units (LCUs)*. Given the vast volume of logs, obtaining LCUs is difficult. Therefore, LUNAR introduces a hybrid ranking scheme to effectively search for LCUs by jointly considering the *commonality* and *variability* among logs. Additionally, LUNAR crafts a novel parsing prompt for LLMs to identify contrastive patterns and extract meaningful log structures from LCUs. Experiments on large-scale public and industrial log datasets demonstrate that LUNAR significantly outperforms state-of-the-art log parsers in terms of accuracy and efficiency, providing an effective and practical solution for real-world deployment.

CCS Concepts: • **Software and its engineering → Software creation and management**.

Additional Key Words and Phrases: log parsing, log analysis, unsupervised learning, large language models

ACM Reference Format:
Junjie Huang, Zhihan Jiang, Zhuangbin Chen, and Michael Lyu. 2025. No More Labelled Examples? An Unsupervised Log Parser with LLMs. *Proc. ACM Softw. Eng.* 2, FSE, Article FSE107 (July 2025), 24 pages. https://doi.org/10.1145/3729377

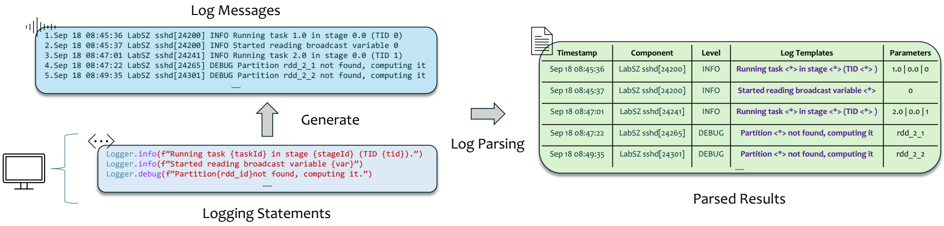
1 Introduction

Log messages record events, transactions, or activities generated by software applications and operating systems at runtime [19, 34, 35]. They provide valuable insights for system performance monitoring and reliability assurance. Various tools have been developed to conduct automated log analysis tasks, including anomaly detection [6, 38, 69–71], root cause analysis [2, 36, 50, 60], and

[*] Zhuangbin Chen is the corresponding author.

Authors' Contact Information: Junjie Huang, Chinese University of Hong Kong, Shatin, Hong Kong, jjhuang23@cse.cuhk.edu.hk; Zhihan Jiang, Chinese University of Hong Kong, Shatin, Hong Kong, zhjiang22@cse.cuhk.edu.hk; Zhuangbin Chen, Sun Yat-sen University, Zhuhai, China, chenzhb36@mail.sysu.edu.cn; Michael Lyu, Chinese University of Hong Kong, Shatin, Hong Kong, lyu@cse.cuhk.edu.hk.

ACM 2994-970X/2025/7-ARTFSE107
https://doi.org/10.1145/3729377



```
1.Sep 18 08:45:36 LabSZ sshd[24200] INFO Running task 1.0 in stage 0.0 (TID 0)
2.Sep 18 08:45:37 LabSZ sshd[24200] INFO Started reading broadcast variable 0
3.Sep 18 08:47:01 LabSZ sshd[24241] INFO Running task 2.0 in stage 0.0 (TID 1)
4.Sep 18 08:47:22 LabSZ sshd[24265] DEBUG Partition rdd_2_1 not found, computing it
5.Sep 18 08:49:35 LabSZ sshd[24301] DEBUG Partition rdd_2_2 not found, computing it
......
```

```
Logger.info(f“Running task {taskId} in stage {stageId} (TID {tid}).”)
Logger.info(f“Started reading broadcast variable {var}”)
Logger.debug(f“Partition{rdd_id}not found, computing it.”)
......
```

Timestamp	Component	Level	Log Templates	Parameters
Sep 18 08:45:36	LabSZ sshd[24200]	INFO	Running task <*> in stage <*> (TID <*>)	1.0 | 0.0 | 0
Sep 18 08:45:37	LabSZ sshd[24200]	INFO	Started reading broadcast variable <*>	0
Sep 18 08:47:01	LabSZ sshd[24241]	INFO	Running task <*> in stage <*> (TID <*>)	2.0 | 0.0 | 1
Sep 18 08:47:22	LabSZ sshd[24265]	DEBUG	Partition <*> not found, computing it	rdd_2_1
Sep 18 08:49:35	LabSZ sshd[24301]	DEBUG	Partition <*> not found, computing it	rdd_2_2
			......	

Fig. 1. An example of log parsing procedure.

failure diagnosis [5, 21, 25, 65]. *Log parsing*, which transforms semi-structured log messages into structured formats [30], serves as a critical preliminary step in log analysis. Typically, a raw log message contains two parts: 1) *log templates*: constant parts that describe the main content of the logged event; 2) *log parameters*: dynamic parts that contain the parameters (determined at runtime) associated with the event. Figure 1 demonstrates some log messages generated by their logging statements and parsed into structured data.

A straightforward approach to log parsing involves matching raw log messages with their corresponding logging statements in the source code [4, 54, 55]. However, in practice, source code is not always accessible, particularly for commercial software and third-party libraries. Consequently, various data-driven log parsers have been proposed to directly extract templates and parameters from raw logs. They can be generally categorized into two types: *syntax-based parsers* and *semantic-based parsers*. Syntax-based parsers [17, 59, 63, 67] resort to statistical or heuristic rules (*e.g.*, log length, word frequency) to identify common parts among logs as templates. However, these approaches often struggle to accurately recognize templates when log messages deviate from the handcrafted features, resulting in limited parsing accuracy in practice. To address these limitations, deep learning models have been introduced in this field to leverage more advanced features of log data, *i.e.*, their semantics. For example, LogPPT [31] fine-tunes a pre-trained language model (*e.g.*, RoBERTa [40]) based on manually labelled log templates, aiming for better performance. With the increasing popularity of large language models (LLMs) for log analysis, recent studies [26, 64] have started employing LLMs for log parsing. These parsers, which are based on LLMs, leverage the comprehensive pre-trained knowledge these models possess and apply the in-context learning (ICL) paradigm [10, 14]. By providing labelled examples as demonstrations, they specialize LLMs for the task of log parsing.

Despite effective, existing semantic-based log parsers largely depend on high-quality labelled examples to achieve optimal performance [44]. However, collecting high-quality labelled log data is challenging in real-world scenarios, which hinders their applicability and scalability in practice. On one hand, production log data are huge in volume, *e.g.*, hundreds of millions of logs per hour according to recent studies [61]. Manually annotating log messages of this scale from diverse systems is labor-intensive and error-prone, requiring substantial domain expertise to ensure accuracy and consistency [27]. On the other hand, the log data from real-world software are continuously evolving over time [61, 64]. New log messages and log templates can emerge frequently, reflecting changes in system behavior, updates, and new feature deployments. This dynamic nature requires constant re-annotation and adaptation of parsers, making it difficult to preserving high accuracy. Consequently, the performance of existing semantic-based log parsers may degrade significantly without continuous and significant manual intervention. As demonstrated in our empirical study (Section 2.1), when the proportion of labelled examples decreases by 70%, the F1 score of template accuracy (FTA) of the state-of-the-art semantic-based log parsers, LogPPT [31] and LILAC [26], drops by 33% and 15%, respectively.

To address this label-demanding problem, we propose LUNAR, an LLM-based unsupervised log parser, which can generalize to any new log dataset without manual annotations or prior knowledge of the specific log formats. The core idea behind LUNAR lies in leveraging LLMs' ability to perform comparative analysis on multiple log messages that vary in their parameter parts. While LLMs may struggle with direct log parsing, the comparison can derive valuable insights by identifying and interpreting patterns across these variations. For instance, by contrasting logs such as "`session opened for user news`" and "`session opened for user test`," LLMs can easily infer that the tokens "news" and "test" are likely to represent a username parameter. We refer to such group of logs as a Log Contrastive Unit (LCU), which are similar enough to facilitate meaningful comparisons, yet diverse enough to highlight the variable parameters. To find LCUs, we introduce a hybrid ranking scheme by jointly considering the *commonality* and *variability* among logs. However, given the vast volume of logs, evaluating every possible combination of logs is impractical. Thus, we introduce a hierarchical sharder to first divide logs into different buckets based on log length and top-k frequent tokens. The LCU selection and subsequent log parsing can then be efficiently performed in each bucket in parallel. Using the selected LCU, LUNAR instructs LLMs with our novel specialized parsing prompt, leveraging LLMs' advanced textual understanding to identify templates. Unlike ICL adopted by existing methods that demand labelled demonstrations [26, 64], the prompt specifies task intention and output constraints in detail, and provides representative parameter examples to inform LLMs of parameter characteristics. Guided by the prompt, the LLM can differentiate between the variant and invariant parts among the given logs, thereby generating more accurate templates.

We evaluate the performance of LUNAR against a range of label-free (unsupervised) parsers and label-dependent parsers. The experiments are conducted on 14 large-scale log parsing datasets in Loghub-2.0 [27] from LogPAI [72]. The results show that LUNAR substantially outperforms the unsupervised baselines, surpassing Brain [67] and LILAC w/o ICL [26] by 43.5% and 19.6% in FTA, respectively. Compared with label-dependent methods, LUNAR achieves performance on par with the current state-of-the-art parser, LILAC. Moreover, with parallelization capability, LUNAR attains a parsing speed comparable to most syntax-based baselines and superior than semantic-based baselines, facilitating efficient parsing of large-scale log data. Our evaluation shows the potential of LUNAR for deployment in real-world production systems, where the efficiency, accuracy and generalizability are critical concerns.

To sum up, the main contributions of this work are as follows:

- To the best of our knowledge, we propose the first LLM-based unsupervised log parser LUNAR, which instructs LLMs to identify the invariant parts of the given log contrastive units (LCUs).
- To enable efficient LCU sampling, we introduce a hierarchical sharding scheme, which reduces sample overhead and allows parallel parsing. Moreover, we propose a hybrid method to measure the LCU by balancing the commonality and variability.
- We evaluate LUNAR on large-scale public datasets. The results show that LUNAR significantly outperforms unsupervised parsers by 19.6% and achieves comparable performance with state-of-the-art label-depended method, *i.e.*, LILAC, in terms of accuracy and efficiency.

2 Motivation

2.1 Limitations of Semantic-based Log Parsers

Recently, semantic-based log parsers [31, 42] have gained significant attention, outperforming traditional syntax-based methods [15, 17, 67] that use specially designed features or heuristics (e.g. ngram [7], prefix tree [11, 17]) by a considerable margin [27, 31]. The improvement mainly stems from the use of language models (LMs) to comprehend the semantics within log messages without the reliance of handcrafted features [31, 32]. LMs are pre-trained on a large-scale corpus of textual data and can take advantage of the learned language patterns to perform downstream tasks [9, 40]. For example, LogPPT [31] fine-tunes a small language model (*e.g.*, RoBERTa [40] with 110M parameters) to predict templates and parameters in log messages. Recent works [26, 64] also utilize large language models (LLMs) (*e.g.*, GPT-3.5 with 175B parameters) for log parsing. With the increasing size of training corpus and model parameters, LLMs have learned more pre-trained knowledge related to textual logs. This equips them with the ability to make more accurate differentiations between static templates and varying parameters within log messages.

Despite the promising results reported, semantic-based parsers require labelled examples (*i.e.*, log messages and their corresponding ground-truth log templates) to achieve their full effectiveness on log parsing tasks. For instance, the state-of-the-art log parser, LILAC [26], employs labelled samples to perform in-context learning (ICL) to guide language models in producing templates, which is a crucial component to ensure its effectiveness. However, collecting sufficient labelled data is difficult in practice. The reasons behind this are two-fold. Firstly, manually collecting labelled examples from log data is labor-intensive and error-prone, which requires extensive domain expertise. Secondly, software frequently undergoes changes [61, 64, 70], resulting in new semantics and patterns in the log data (*i.e.*, concept drift [13]). Consequently, the performance of these log parsers tends to degrade over time.

To quantitatively understand the impact of labelled examples on parsing performance, we re-evaluated two representative semantic-based log parsers, *i.e.*, LogPPT [31] and LILAC [26], using varying label proportions. Specifically, our study utilized large-scale log parsing datasets from Loghub-2.0 [27], which contain 50.4 million log messages from 14 real-world software systems. For each dataset, we randomly sampled different proportions of labelled log templates as the model-accessible oracles, which are then applied for fine-tuning or ICL. The proportion ranges among 75%, 50%, 25%, 10%, and 5%, simulating real-world scenarios where oracle labelled templates are scarce due to insufficient manual labeling or system evolution. We measured their effectiveness using the widely adopted F1 score of template accuracy (FTA) [30]. To reduce the bias introduced by randomness, we performed the experiments five times for each setting, following previous studies [27, 31, 64, 72]. We reported the average FTA scores, as shown in Figure 2.

We observe that the performance of both LogPPT and LILAC significantly declines as the label proportion decreases. For example, when the label proportion is 75%, LogPPT achieves a performance of over 50%. However, when the label proportion drops to 5%, its performance falls to about 17%, indicating a substantial decrease of 33%. Similarly, with a label proportion of 75%, LILAC attains an average FTA score of approximately 81%. When the proportion decreases to 25%, the score drops to about 75%. At a label proportion of just 5%, the score further declines to around 66%, representing a reduction of over 15% compared to the highest score. These results suggest that the proportion of labelled examples can significantly impact the performance of semantic-based log parsers, posing challenges to apply them into real-world software systems. Therefore, we aim to develop a label-free semantic-based log parser, which can avoid label insufficiency problems and generalize to unseen software without labelled templates.

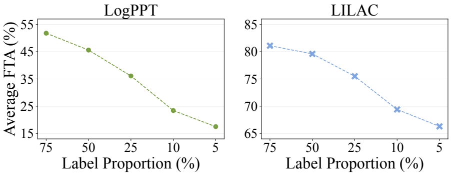
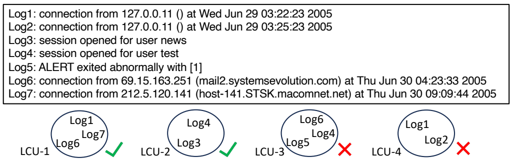


Fig. 2. Empirical study on the influences of label proportion on semantic-based log parsers.



```
Log1: connection from 127.0.0.11 () at Wed Jun 29 03:22:23 2005
Log2: connection from 127.0.0.11 () at Wed Jun 29 03:25:23 2005
Log3: session opened for user news
Log4: session opened for user test
Log5: ALERT exited abnormally with [1]
Log6: connection from 69.15.163.251 (mail2.systemsevolution.com) at Thu Jun 30 04:23:33 2005
Log7: connection from 212.5.120.141 (host-141.STSK.macomnet.net) at Thu Jun 30 09:09:44 2005
```

Fig. 3. Examples of log contrastive units (LCUs).

2.2 Log Contrastive Unit for LLM-based Parsing

Here, we present our motivation of developing a label-free, LLM-based log parsing approach.

Existing solutions rely on labelled data to guide language models in inferring the parameters from log messages. In contrast, *our core insight is that the LLM itself can derive sufficient hints by comparing multiple logs.* For example, by comparing the logs: "session opened for user news" and "session opened for user test", LLMs can easily infer that "news" and "test" are likely username parameters, because these segments vary while the rest of the log message remains consistent. We aim to leverage such comparisons to guide LLM without the need for labelled data.

We define such grouped log messages as a *log contrastive unit (LCU)*, which consists of multiple log messages potentially sharing the same template, allowing LLMs to parse them through comparisons. Constructing an effective LCU that guides the LLM to produce accurate parsing results is non-trivial. Specifically, log messages within the same LCU need to exhibit both commonality and variability:

- Commonality: These log messages should share some common tokens. If they differ significantly, they are likely generated by different logging statements and are not different templates. For example, logs in LCU-3 differ greatly, thus failing to provide proper hints to LLMs.
- Variability: These log messages should differ in some of their tokens. If they are all identical, the LLM might mistakenly interpret all tokens as constants. For instance, the IP address in logs of LCU-4 is identical, which can mislead the LLM to regard the parameter parts as constants.

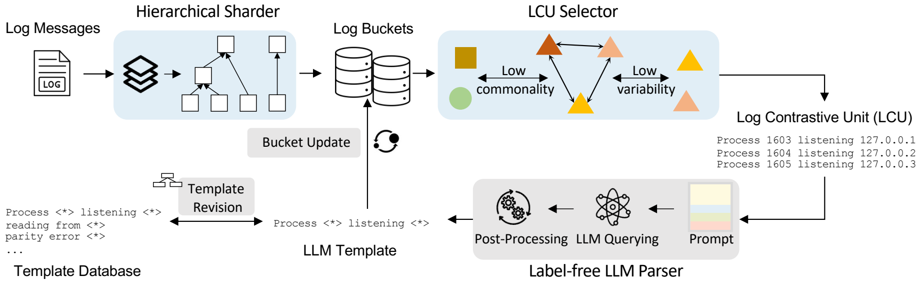
LCU-1 and LCU-2 are examples of well-structured LCUs, as log messages within them exhibit both commonality and variability, guiding the LLM to infer the templates and parameters through comparison. However, we observe the following challenges in practice:

- **Challenge 1: LCU Volume Explosion**: The sheer volume of log data generated by modern software systems can lead to an exponential increase in the number of potential LCUs. The explosion in combinations complicates the efficient scaling of the LCU-based approach.
- **Challenge 2: Balancing Commonality and Variability**: Log messages often exhibit significant diversity, with templates comprising vastly different tokens. The diversity makes it difficult to identify effective LCUs that balance commonality and variability. Finding suitable combinations of log messages that share enough common tokens while still exhibiting variability is a complex task, exacerbated by the dynamic nature of log data and the need for continuous adaptation.

3 Methodology

3.1 Overview

Figure 4 illustrates the overall framework of LUNAR, which consists of three main components: *hierarchical sharder*, *generative LCU ranker*, and *label-free LLM parser*. To address the first challenge, we propose the hierarchical sharder, which divides raw log messages into different buckets based on log length and top-k ranked tokens. By separating logs with low similarity into different buckets, this component reduces the sampling overhead and enables parallelization for efficient parsing (§3.2). To address the second challenge, we propose the generative LCU ranker, which operates within each bucket to continuously sample LCUs for the LLM to parse. This module computes a hybrid LCU score that jointly considers both the commonality and variability of the LCUs, guiding the sampling process to ensure effective LCUs (§3.3). Then, the label-free LLM parser constructs an organized prompt for the mined LCUs to query the LLM and obtain the templates from the LLM's response. We introduce a novel format to organize the parsing prompt, which specifies the task intention and output constraints, and includes several representative parameter examples to inform the LLM of the parameter characteristics (§3.4). Finally, the template is passed to a template database, which can validate the template and merge similar ones to ensure accurate and efficient parsing(§3.5). The combination of all components ensures that the LLM can effectively parse the logs without the need for labelled data, addressing both scalability and diversity challenges.



```
Process 1603 listening 127.0.0.1
Process 1604 listening 127.0.0.2
Process 1605 listening 127.0.0.3
```

```
Process <*> listening <*>
reading from <*>
parity error <*>
...
```

```
Process <*> listening <*>
```

Fig. 4. The overall workflow of LUNAR.

3.2 Hierarchical Log Sharder

Directly sampling LCUs from the whole log messages is overwhelming due to the large volume of log data, which often contain million of log messages. Hence, LUNAR first groups the log messages into multiple log buckets, where each bucket contain logs that share some similarity. By doing so, extremely different logs, which have less potentials to belong to the same template [17], are placed to different buckets. Specifically, we propose a hierarchical algorithm to shard log: 1) firstly divide the log messages by log length, and 2) progressively divide logs by top-k frequent tokens. Figure. 5 demonstrates the workflow of hierarchical log sharder.

3.2.1 Length-based Sharding. LUNAR first shards the log messages by the log length. Logs with the same number of tokens are grouped in a bucket, which is a widely adopted heuristic to conduct initial grouping in log analysis [17, 33]. Here we simply adopt the whitespace as the delimiter. We do not use other delimiters such as ':' because most of these delimiters appear in parameters and introducing them can produce over-fragmented logs, leading to a noisy sharding result.

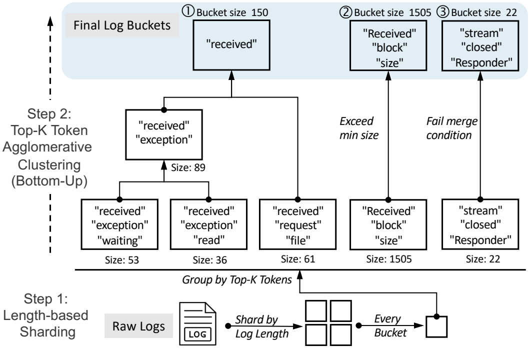

Final Log Buckets
① Bucket size 150
"received"
② Bucket size 1505
"Received" "block" "size"
③ Bucket size 22
"stream" "closed" "Responder"
Step 2: Top-K Token Agglomerative Clustering (Bottom-Up)
"received" "exception"
Size: 89
Exceed min size
Fail merge condition
"received" "exception" "waiting"
Size: 53
"received" "exception" "read"
Size: 36
"received" "request" "file"
Size: 61
"Received" "block" "size"
Size: 1505
"stream" "closed" "Responder"
Size: 22
Group by Top-K Tokens
Step 1: Length-based Sharding
Raw Logs
LOG
Shard by Log Length
Every Bucket

Fig. 5. Hierarchical log sharding in LUNAR.

3.2.2 Top-k Token Agglomerative Clustering. However, simply sharding by log length is too coarse-grained since logs belonging to different templates could have the same length [26]. Thus LUNAR continuously shards each bucket into more purified ones where less templates are included in one bucket. Intuitively, the static parts of a log generally have higher occurrences than its parameter part, therefore logs that share the most frequent tokens have more potential to belong to the same template [26, 27, 39]. To achieve this, we introduce a *hierarchical agglomerative clustering* algorithm based on top-k frequent tokens, which builds the sub-buckets in a bottom-up way.

Specifically, for each bucket obtained from the first step, we first group logs into singleton clusters based on the same top-k tokens in each log. The top-k tokens of a log is an ordered list extracted based on the token frequency and position. The frequency is counted among logs in the bucket. The more frequent tokens takes a more preceding place in the list. However, some log messages can have multiple tokens with the same frequency, which leads to confusion in ranking; hence we additionally rank tokens with the same frequency by their positions in the logs, with the preceding token being ranked before the following token.

Then, we iteratively merge the singleton clusters to obtain final log buckets until a stopping criterion is met. In each iteration, if a cluster contains more than N logs, it will be directly nominated as a standalone bucket and no further merge operations will be conducted on it. For the remaining clusters, which contain less than N logs, we merge those with the same top-$(k-1)$ tokens to form a parent cluster. The iteration stops if all clusters exceed N logs, or no any two clusters satisfy the merge condition. Finally, we obtain a collection of buckets containing a smaller number of logs with higher relevance.

It is worth noting that the collected log buckets have a two-level hierarchy, where each hyper-bucket obtained by length-based sharding is disjointly separated to multiple final buckets. Therefore, buckets under the same level are mutually exclusive with each other, *i.e.*, every log belongs to one bucket on either sharding level. The advantages of this design are two-fold. Firstly, once a template is obtained, it can be leveraged to match the logs in the bucket while not required for other buckets. Secondly, this characteristic enables parallel parsing, where the buckets can be independently allocated on multiple executors. More details will be described in Section 3.6

3.3 Two-stage LCU Selector

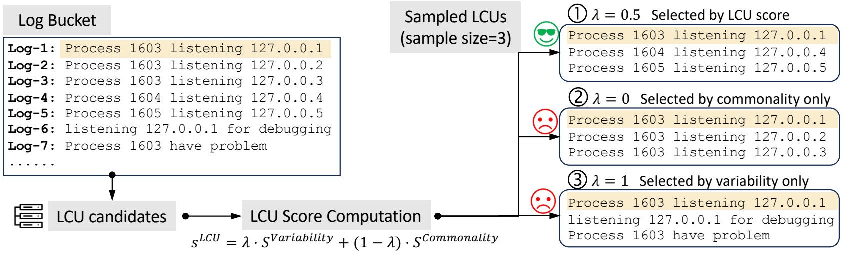
After log sharding, LUNAR iteratively select a group of similar logs for the LLM to make comparisons and extract templates. A suitable group of logs should be similar in tokens to each other, as logs generated by the same logging statement share several identical tokens. On this basis, the logs are expected to have as much variance as possible, enabling LLMs to make cross-log comparisons to infer parameters. For example, in Figure 6, group ① and ② are preferred than group ③ as their logs are more likely to belong to the same template. Moreover, group ① is better than ② due to its higher variance reflected in the diverse parameter value of process index.

We term such a group of logs as a log contrastive unit (LCU). Logs in each LCU have both commonality (*e.g.*, with common words or belonging to the same template) as well as variability (*e.g.*, with different tokens), so that it can be leveraged to infer the parameters. To collect LCUs, we leverage a two-step sampling approach: *stratified LCU generation* and *hybrid LCU nomination.*

3.3.1 *Stratified LCU Generation.*

Given a bucket of logs, LUNAR first generate multiple candidate LCUs that share some similarity. A straightforward approach is to enumerate all possible log combinations. However, this method is computationally expensive due to the sheer volume of logs and the complexity of enumeration. To reduce computation overhead, we apply stratified sampling to collect a limited-sized log pool by sampling from different similarity levels before combination.

Given a log bucket, LUNAR preprocess the logs with simple regular expressions following previous unsupervised parser Drain [17]. It then randomly selects a log message in the bucket as the anchor log. Next, LUNAR computes the pair-wise similarities of the anchor log to the remaining logs. We use Jaccard similarity, which is a widely adopted metric to measure pair-wise log similarity [17, 26]. Specifically, we first split the log l into tokens $T(l)$ with white-space delimiter and then compute the Jaccard similarity $JS(l_1, l_2) = \frac{T(l_1) \cap T(l_2)}{T(l_1) \cup T(l_2)}$. Then, we remove the logs with a similarity below a threshold s or equals to 1.0. The poor similarity indicates the less possibility to belong to the same template, while a similarity of 1.0 indicates duplication. After that, we create a pool of logs by stratified sampling from each similarity level to ensure balanced sampling. Specifically, we set the sample size of each level equal to the LCU size m to involve more diverse logs. Lastly, from the log pool, LUNAR generates multiple LCUs by combinatorial enumeration. Each LCU consists of an anchor log, as well as $m-1$ logs chosen from the pool. Since we sample a small size of LCU (*e.g.*, $m=3$ or $m=4$) from a limited size of log pool, the computational cost of combinatorial enumeration is within an acceptable range (More details about efficiency can be found in Section 5.4). If the log pool contain less than $m-1$ logs, we directly return the anchor with the pool as the candidate LCU. Finally, a candidate set of LCUs are generated and proceeded to hybrid ranking in the next step.


Log Bucket
LCU candidates
LCU Score Computation
$s^{LCU} = \lambda \cdot S^{Variability} + (1 - \lambda) \cdot S^{Commonality}$
Sampled LCUs
(sample size=3)
① $\lambda = 0.5$ Selected by LCU score
② $\lambda = 0$ Selected by commonality only
③ $\lambda = 1$ Selected by variability only

```
Log-1: Process 1603 listening 127.0.0.1
Log-2: Process 1603 listening 127.0.0.2
Log-3: Process 1603 listening 127.0.0.3
Log-4: Process 1604 listening 127.0.0.4
Log-5: Process 1605 listening 127.0.0.5
Log-6: listening 127.0.0.1 for debugging
Log-7: Process 1603 have problem
......
```

```
Process 1603 listening 127.0.0.1
Process 1604 listening 127.0.0.4
Process 1605 listening 127.0.0.5
```

```
Process 1603 listening 127.0.0.1
Process 1603 listening 127.0.0.2
Process 1603 listening 127.0.0.3
```

```
Process 1603 listening 127.0.0.1
listening 127.0.0.1 for debugging
Process 1603 have problem
```

Fig. 6. An example of expected LCU in LCU nomination.

3.3.2 Hybrid LCU Ranking. After obtaining a small number candidate LCUs, LUNAR conducts hybrid ranking to select the most suitable LCU. The selected logs are required to exhibit variability, while maintain some commonality as we discussed in Section 2.2. By doing so, LLMs can be less likely to incorrectly recognize the frequent tokens as parameters, while also benefit from the contrastive information in the LCU. Specifically, we propose to compute the variability score and the commonality score to measure an LCU.

Firstly, the variability score of an LCU is computed as average of pair-wise distances to measure the overall variability of the logs. Suppose an LCU consists of L logs $LCU = \{l_1, l_2, ...l_L\}$, we first compute the distance between any two logs in the LCU based on Jaccard similarity:

$$dist(l_i, l_j) = 1 - JS(l_i, l_j). \tag{1}$$

Then we average the distances of every two logs in the LCU to obtain the variability score:

$$S_{LCU}^{Variability} = \frac{2}{L(L-1)} \sum_{i=1}^{L} \sum_{j=i+1}^{L} dist(l_i, l_j), \tag{2}$$

where the higher the variability score, the more diverse the LCU, indicating more contrastive information is included.

Simply ranking by variability could result in LCUs containing totally different logs (*e.g.*, LCU ③ in Figure 6). As the compensation, the commonality score of an LCU is introduced, which is computed as the average of absolute similarity difference to balance the variance of logs. Specifically, we P log pairs can be obtained from an LCU, then:

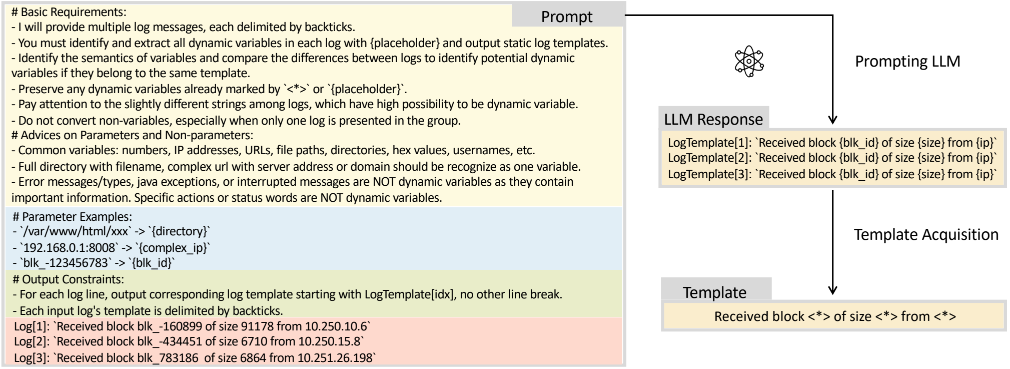
$$S_{LCU}^{Commonality} = \frac{2}{P(P-1)} \sum_{i=1}^{P} \sum_{j=i+1}^{P} (1 - |JS(p_i) - JS(p_j)|), \tag{3}$$

where $JS(p_i)$ is the Jaccard similarity of log pair p_i. The higher the commonality score, the more likely that logs belong to the same template, as the similarity difference between pair to pair is relatively low. Logs that share the same template but differ only in their parameters will exhibit exactly such a pattern. For example, in the LCU② of Figure 6, the similarity between any two logs is the same as any other two logs within the LCU, which implies the highest commonality score of 1.

To jointly consider the variability and commonality scores, we combine them with linear interpolation with a weight of λ, to obtain the hybrid LCU score. The LCU with maximum interpolation score selected as the final LCU, which is computed as follows:

$$S_{LCU} = \lambda \cdot S_{LCU}^{Variability} + (1 - \lambda) \cdot S_{LCU}^{Commonality}. \tag{4}$$

To present the LCU score in a straightforward way, we provide an example with a sample size of 3 in Figure 6. In this example, by selecting with variability score only, LCU ③ will be sampled. However, LCU ③ is sub-optimal as the logs belong to three different templates. LLMs might misclassify the parameter "1603" as this group fails to provide useful contrasts towards it. With the commonality score only, LCU ② will be sampled. However, this is also sub-optimal due to the



```
# Basic Requirements:
- I will provide multiple log messages, each delimited by backticks.
- You must identify and extract all dynamic variables in each log with {placeholder} and output static log templates.
- Identify the semantics of variables and compare the differences between logs to identify potential dynamic variables if they belong to the same template.
- Preserve any dynamic variables already marked by `<*>` or `{placeholder}`.
- Pay attention to the slightly different strings among logs, which have high possibility to be dynamic variable.
- Do not convert non-variables, especially when only one log is presented in the group.
# Advices on Parameters and Non-parameters:
- Common variables: numbers, IP addresses, URLs, file paths, directories, hex values, usernames, etc.
- Full directory with filename, complex url with server address or domain should be recognize as one variable.
- Error messages/types, java exceptions, or interrupted messages are NOT dynamic variables as they contain important information. Specific actions or status words are NOT dynamic variables.
# Parameter Examples:
- `/var/www/html/xxx` -> `{directory}`
- `192.168.0.1:8008` -> `{complex_ip}`
- `blk_-123456783` -> `{blk_id}`
# Output Constraints:
- For each log line, output corresponding log template starting with LogTemplate[idx], no other line break.
- Each input log's template is delimited by backticks.
Log[1]: `Received block blk_-160899 of size 91178 from 10.250.10.6`
Log[2]: `Received block blk_-434451 of size 6710 from 10.250.15.8`
Log[3]: `Received block blk_783186  of size 6864 from 10.251.26.198`
```

```
LogTemplate[1]: `Received block {blk_id} of size {size} from {ip}`
LogTemplate[2]: `Received block {blk_id} of size {size} from {ip}`
LogTemplate[3]: `Received block {blk_id} of size {size} from {ip}`
```

```
Received block <*> of size <*> from <*>
```

Fig. 7. Demonstration-free parsing prompt in LUNAR.

identical parameter “1603”. This LCU could mislead LLMs to incorrectly recognize parameters due the identical tokens. In contrast, by combining the variability score and commonality score, we can obtain an optimal LCU, where each parameter has at least two values present. Using this LCU can be maximumly inform LLM to consider the changing tokens, thereby leading to an improved parsing accuracy.

3.4 Label-free LLM Parser

Upon obtaining an LCU, LUNAR applies an LLM-based parser to extract the template without any labelled examples. The parser utilizes the advanced text comprehension ability of LLMs, which have demonstrated remarkable performance in various unsupervised information extraction tasks (*e.g.*,, named entity recognition [62] and text classification [43]) using a task-specific prompt instruction. Hence, we believe that LLMs have the potential to identify templates from LCUs in an unsupervised manner with an optimized prompt instruction. Specifically, our LLM parser first creates a well-designed prompt for each LCU to instruct the LLM (§3.4.1) and then extract the template from the LLM response (§3.4.2).

3.4.1 Prompt Design. As LLMs are not specifically tuned for log parsing, existing LLM-based log parsers leverage exemplars of log and its labelled template to instruct the task [26, 64]. These demonstrations could specify task intents, output format, and parameter information of the domain, which are useful to enhance log parsing [44, 47]. However, in this work, we focus on developing an unsupervised parser where no labelled templates are provided. Therefore, we need to design a more concrete prompt covering the aforementioned demanding information to facilitate accurate parsing. Specifically, our prompt contains four parts: task instructions, parameter examples, output constraints, and input LCUs. Figure 7 shows a prompt example.

Task Instruction. A clear and useful prompt should provide comprehensive instructions for LLMs to understand the task [16]. In LUNAR, we create the instruction of log parsing with two parts: basic requirements and general advice on parameters. For basic requirements, the task input/output and the mapping mechanism from input to output are firstly specified. The instruction also guides the LLM to consider the similarity between the grouped input logs. For general advice on parameters, high-level descriptions of parameter types are provided. We also highlight error-prone non-parameters, such as error messages as they contain important information.

Parameter Examples. Inspired by the entity examples used for unsupervised NER [24], we involve parameter examples (PEs) to the prompt. These examples are used to provide domain knowledge of parameters and instruct extraction formats of the task, which consist of a parameter type and a parameter value. For instance, “`/var/www/html/xxx`” is the value of a parameter “directory”. We connect the parameter value and type by an arrow in the prompt, aiming to indicate the task intent of converting the value to a placeholder with a bracket.

In our paper, we introduce six types of most frequent PEs and fix them among all evaluation datasets to demonstrate PEs' effectiveness and ensure fair comparison in an unsupervised setting. To obtain these PEs, we first manually examine the parameter types and values identified in previous study [37], and collect three most frequent non-digit parameter types: location indicator, object ID, and time. We then manually examine the corresponding parameter values and summarize six representative and concrete types, including directory, file, IP for location, timestamp for time, and block id, service api for object ID. Finally, we randomly sample a value for each type.

Output Constraints. In output constraints, we explain the desired output formats. The prompt instructs LLMs to produce a template for every log message. Each generated template should be appended with a prefix "LogTemplate[idx]" and delimited with backticks to facilitate template aggregation and extraction.

Queried LCUs. Finally, the LCU of multiple logs are incorporated to the prompt. We arrange the logs in sequential order and apply a prefix schema (*i.e.*, "Log[]:") and index indicator to indicate the order. In this way, the LLM can make straightforward comparisons and produce corresponding template for each log.

Guided by the task instruction, parameter examples, and output constraints, the LLM could more accurately generate the templates of the queried logs in the LCU.

3.4.2 Template Acquisition. After constructing the prompt, we use the LLM to identify the templates in the LCU. Specifically, we first leverage prompts to query an LLM to get a response. After that, we perform post-processing to extract valid templates from the response. Specifically, we extract all backticked strings with a "LogTemplate" prefix, thanks to clear output constraints in the prompt. Then we aggregate the extracted templates into one template by selecting the most frequent template. Finally, we replace the bracketed parameters by the placeholder $< * >$.

3.5 Template Revision and Bucket Updating

After extracting a template from LLM, LUNAR compares it with existing templates and examines the logs that match the template. As LCU selector weights static words and parameters equally in commonality scores, it can produce suboptimal LCUs with several identical parameters across logs. The repetition can mislead the LLM parser to incorrectly classify parameters as static parts in templates. To mitigate the issue, we follow LILAC [26] and compare the newly extracted template with existing ones, revising it to improve accuracy. Once obtaining the revised template, LUNAR examines the matchable logs to update the bucket. To improve efficiency and minimize redundancy, LUNAR uses a template database to store both templates and indexes of matched logs.

3.5.1 Template Revision. When an LLM-derived template is obtained, LUNAR first selects a similar template and subsequently generate a revised version. The revision process is skipped if no similar template can be found. Specifically, the selection is based on the token difference, where a template with equal token count or with only 1 or 2 different tokens is chosen as the candidate. The strict criterion maximizes the probability that the templates belong to a single hyper template.

Subsequently, LUNAR performs a pairwise, token-level merge to generate a revised template. Specifically, for tokens at the same position, identical ones are retained, while differing ones are modified. We employ a bi-directional editing strategy to merge tokens, retaining matching characters across tokens on both sides, and replacing the middle positions with a wildcard symbol "<∗>". For example, the token "mesos-slave-<∗>" and "mesos-master-<∗>" are edited to "mesos-<∗>". It is worth-noting that to avoid over-merging meaningful content, LUNAR only edits tokens that resemble parameters, specifically those with non-alphabetic symbols.

3.5.2 Bucket Updating. After template revision, LUNAR matches logs to the template using regular expressions and writes the updating history to a template database. The template is assigned as the final prediction to these successfully matched logs, which are then removed from the buckets. To reduce redundant matching, the template and corresponding log indexes are stored in template database. This allows LUNAR to directly assign the revised template to logs that match a pre-merged template without repeating regular expression matching. Remaining logs in the buckets proceeded to next parsing iteration. Parsing terminates once all buckets are empty.

It is worth-noting that hierarchical sharding produces buckets at two levels, *i.e.*, sharded by length and further clustered by top-k tokens. A straightforward approach for bucket updating is to only examine and update the final bucket. However, logs of the same template can be distributed in different buckets of equal length, potentially causing redundant iterations for one template, leading to prolonged parsing time and increased LLM querying expenses. To balance the parsing overhead, LUNAR conducts template matching and updates buckets at the first sharding level, meaning that buckets with identical log lengths are all examined once a template is obtained. Additionally, template revision is also consequently conducted at the first sharding level, where templates of equal-length logs are stored in one template database.

3.6 Parallelization

In previous sections, we have discussed the building components of LUNAR. In practice, scalability is a crucial issue when parsing vast amounts of log messages [61]. Therefore, we introduce the parallelization mechanism of LUNAR to improve the parsing speed.

As mentioned in §3.2, hierarchical sharding produces mutually exclusive log buckets at two levels. Consequently, logs in different buckets can be processed in a completely parallel fashion. However, since template revision and bucket updating is performed among all the buckets of the same length at the first level. Simply allocating buckets at the second level to produce template and update buckets can lead to *template conflict* and *revision conflict*, *i.e.*, *different templates being assigned to the same log* and *one template being simultaneously modified with different LLM templates.* To address the problem, the allocation is conducted at the first level. Specifically, we aim to allocate n executors for all first-level log buckets. Within each first-level log bucket, LUNAR iteratively selects the largest second-level bucket, selects an LCU, queries the LLM to obtain the template, revise it, and updates buckets under this level.

3.7 Online Parsing

LUNAR can be seamlessly extended to an online setting, allowing it to consistently parse streaming logs without the need for manually labeled templates. After obtaining a collection of templates with LUNAR during offline phase, we can leverage the extracted templates for online parsing. Specifically, when new logs arrive, they are firstly matched against existing templates. The unmatched logs are then passed to LUNAR for LLM-based parsing. To construct LCUs for LUNAR, we introduce a log pool to collect the unparsed logs. When the pool exceeds a number of logs, LUNAR is invoked to iteratively parse the logs by performing hierarchical sharding, LCU selection, label-free LLM parsing, template revision, and bucket updating. After that, the parsed templates are added to the template database for future matching.

4 Experiment Setup

We conduct experiments to evaluate LUNAR by answering following research questions (RQs):

- **RQ1:** How effective is LUNAR in log parsing?
- **RQ2:** How do different components affect LUNAR?
- **RQ3:** How effective is LUNAR integrated with different LLM and prompts?
- **RQ4:** How efficient and cost-effective is LUNAR in parsing large-scale log data?

4.1 Datasets

Table 1. Statistics of Log Parsing Datasets

	Dataset	# Logs	# Templates
Open-source Logs [27, 72]	Proxifier	21,320	11
	Apache	51,977	29
	OpenSSH	638,946	38
	HDFS	11,167,740	46
	OpenStack	207,632	48
	HPC	429,987	74
	Zookeeper	74,273	89
	HealthApp	212,394	156
	Hadoop	179,993	236
	Spark	16,075,117	236
	BGL	4,631,261	320
	Linux	23,921	338
	Mac	100,314	626
	Thunderbird	16,601,745	1,241
Industrial Logs [21, 63]	CTS [63]	264	78
	HiBench [63]	1,879	93
	LoFI [21]	2,080	453
Average		2,965,931.9	230.1

To comprehensively evaluate LUNAR, we select 17 log parsing benchmark datasets with diverse log sources and template availability, including 14 datasets from Loghub-2.0 [20, 27], CTS [63], HiBench [63], and LoFI [21]. **Loghub-2.0** [20, 27] is a collection of 14 large-scale benchmark datasets for log parsing. The logs are publicly released by the LogPAI team [72] and annotated with ground-truth templates in Loghub-2.0, covering a variety of systems such as distributed systems, supercomputer systems, and server-side applications. Compared with Loghub [20], Loghub-2.0 is equipped with 3× templates and $1,800\times$ log messages on average, which can support more comprehensive evaluation of accuracy and efficiency. Apart from Loghub-2.0, we also introduce three industry log datasets released in recent two years. **CTS** [63] logs are collected from a cloud testing system. **HiBench** [63] logs are collected from a commercial cloud benchmark system. Both are annotated with ground-truth templates and are publicly available. We only keep the single-line log messages for evaluation. **LoFI** [21] logs are collected from a cloud service system with no available ground-truth templates. We manually annotated the templates following the previous guidelines [27, 30]. Table 1 shows the statistics of these datasets.

4.2 Baselines

We compare LUNAR with seven open-sourced state-of-the-art log parsers, including four unsupervised syntax-based parsers and three label-required semantic-based parsers. For unsupervised label-free parsers, we adopt four syntax-based methods and one LLM-based method. The syntax-based methods include AEL [28], Drain [17], and Brain [67], selected for their superior performance compared to other syntax-based methods [27, 30, 72]. Additionally, LogMine [15] is included due to its method that initially clusters and then merges, akin to LUNAR. As there are currently no LLM-based log parsers proposed for label-free parsing, we adopt a label-free variant of LILAC [26] (LILAC w/o ICL) by removing the in-context learning (ICL) module. In terms of label-required log parsers, we select three state-of-the-art semantic-based log parsers: UniParser [42], LogPPT [31], and LILAC [26]. UniParser trains a long short-term memory (LSTM) model on labelled log data for log parsing. LogPPT uses labelled log data to perform prompt-based fine-tuning based on RoBERTa [40]. LILAC samples similar logs from a small set of labelled logs (*e.g.*, 32 logs) for each log, utilizing an ICL paradigm for LLMs to parse logs.

4.3 Evaluation Metrics

Following previous works [27, 30, 72], we evaluate LUNAR with the following four metrics.

- *Grouping Accuracy* (GA) [72] is a log-level metric that measures the the amount of log messages of a same template are grouped together by the parser. It is computed as the ratio of correctly grouped log messages over all log messages, where a log message is regarded as correctly grouped if and only if its predicted template have the same group of log messages as the oracle.
- Parsing Accuracy (PA) [7] is a log-level metric that measures the correctness of extracted templates and variables. It is defined as the ratio of correctly parsed log messages over all log messages,

where a log message is considered to be correctly parsed if and only if all its static text and dynamic variables are identical with the oracle.

- F1 score of Grouping Accuracy (FGA) [27] is a template-level metric that measures the ratio of correctly grouped templates. It is computed as the harmonic mean of precision and recall of grouping accuracy, where the template is considered as correct if log messages of the predicted template have the same group of log messages as the oracle.
- F1 score of Template Accuracy (FTA) [30] is a template-level accuracy computed as the harmonic mean of precision and recall of Template Accuracy. A template is regarded as correct if and only if it satisfies two requirements: log messages of the predicted template have the same group of log messages as the oracle, and all tokens of the template is the same as the oracle template.

4.4 Environment and Implementation

We conduct all the experiments on a Ubuntu 20.04.4 LTS server with 256RAM and an NVIDIA A100 40G GPU. The LLM used in LUNAR is GPT-3.5 due to its popularity in recent log analysis studies [26, 44, 64]. We use the latest available LLM *gpt-3.5-turbo-0125* and invoke the official API provided by OpenAI [1]. To minimize the randomness introduced by token sampling, we set the temperature to 0 and report the average performance of three runs. By default, we set the a top-k token number to 3 and minimum cluster size to 100 in our hierarchical log sharder. For the LCU selector, we use an LCU sample size of 3, a interpolation factor λ of 0.6, and a minimum similarity of 0.5, respectively. On average, 0.9 regular expressions per dataset are used at the beginning of LCU selection. For parallelization, we distribute the jobs to 8 executors. We also conduct experiments of LUNAR with different LCU sample sizes and interpolation factors. As for baseline methods, we directly use the accuracy reported in previous work [27] and rerun them with their default parameters on the same environment to fairly compare the efficiency.

5 Evaluation Results

5.1 RQ1: How effective is LUNAR in log parsing?

Setup: In the first research question (RQ), we evaluate the accuracy of LUNAR by comparing it with state-of-the-art log parsers, as accuracy is the most critical factor for log parsers. For comprehensiveness, we compare LUNAR with both unsupervised parsers, which do not require human-labelled log templates, and label-required semantic-based log parsers, which rely on labelled log templates to support training, fine-tuning, or in-context learning. Moreover, we evaluate LUNAR using datasets with both open-source and industry logs to assess its generalizability. Table. 2 presents the overall results on four metrics for LUNAR compared to baseline methods. The best results for each metric on each dataset are highlighted in **bold**, while the second-best results are underlined. If a specific parser could not complete the parsing process within a reasonable timeframe (*i.e.*, 12 hours), as per previous work [26, 27, 30, 72], we denote the score as "-".

Results: Overall, we find that LUNAR achieves the best average grouping accuracy (GA), parsing accuracy (PA), and F1 score of template accuracy (FTA), as well as the second-best F1 score of grouping accuracy (FGA). Additionally, LUNAR demonstrates high accuracy across all 17 datasets, showcasing its robustness in parsing log data from diverse systems. For example, LUNAR achieves the highest FTA on 15 out of 17 datasets and the second-highest FTA on the remaining 2 datasets.

Among unsupervised log parsers, it is evident that LUNAR significantly outperforms the other four baselines across all four evaluation metrics. Among the syntax-based parsers (*i.e.*, AEL, Drain, Brain, and LogMine), Brain achieves the highest scores in template-level metrics, with an average FGA of 71.9% and an average FTA of 36.8%. However, LUNAR exhibits a substantially higher average FGA of 91.3% and FTA of 78.5%, surpassing Brain by 19.4% and 41.7%, respectively. This is

Table 2. Accuracy of LUNAR compared to state-of-the-art baselines on public datasets. (%)

Method	Metric	Proxifier	Apache	OpenSSH	HDFS	OpenStack	HPC	Zookeeper	HealthApp	Hadoop	Spark	BGL	Linux	Mac	Thunderbird	CTS	HiBench	LoFI	Avg.
Unsupervised Log Parsers																			
AEL	GA	97.4	**100.0**	70.5	99.9	74.3	74.8	99.6	72.5	82.3	–	91.5	**91.6**	79.7	78.6	78.8	92.1	84.5	85.5
	PA	67.7	72.7	36.4	62.1	2.9	74.1	84.2	31.1	53.5	–	40.6	8.2	24.5	16.3	62.5	56.7	30.4	45.2
	FGA	66.7	**100.0**	68.9	76.4	68.2	20.1	78.8	0.8	11.7	–	58.7	80.6	79.3	11.6	79.2	87.6	84.4	60.8
	FTA	41.7	51.7	33.3	56.2	16.5	13.6	46.5	0.3	5.8	–	16.5	21.7	20.5	3.5	59.7	48.6	25.0	28.8
Drain	GA	69.2	**100.0**	70.7	99.9	75.2	79.3	99.4	86.2	92.1	88.8	91.9	68.6	76.1	83.1	79.9	61.6	79.0	82.4
	PA	68.8	72.7	58.6	62.1	2.9	72.1	84.3	31.2	54.1	39.4	40.7	11.1	35.7	21.6	66.7	26.8	29.5	45.8
	FGA	20.6	**100.0**	87.2	93.5	0.7	30.9	90.4	1.0	78.5	86.1	62.4	77.8	22.9	23.7	69.3	46.0	78.5	57.0
	FTA	17.6	51.7	48.7	60.9	0.2	15.2	61.4	0.4	38.4	41.2	19.3	25.9	6.9	7.1	57.0	23.6	22.6	29.3
Brain	GA	52.1	99.7	66.3	96.0	**100.0**	80.0	99.3	97.9	56.3	97.2	94.0	79.0	83.4	79.2	80.3	79.1	64.3	82.6
	PA	70.3	28.7	48.1	92.9	14.1	66.3	82.2	17.5	14.3	39.3	40.2	1.0	32.5	26.1	70.1	62.2	25.3	43.0
	FGA	73.7	93.3	75.9	75.9	**100.0**	44.7	79.8	87.2	52.8	20.8	75.6	75.1	75.4	74.8	80.0	68.8	69.3	71.9
	FTA	73.7	46.7	34.5	62.1	29.2	21.3	60.1	33.9	20.0	1.0	19.7	27.5	29.4	27.4	69.3	47.7	21.7	36.8
LogMine	GA	50.4	**100.0**	–	–	–	–	69.7	–	82.7	–	64.4	73.6	85.1	–	87.5	79.7	81.0	77.4
	PA	0.0	26.2	–	–	–	–	45.6	–	52.9	–	9.7	3.5	28.3	–	69.7	43.3	29.6	30.9
	FGA	0.8	**100.0**	–	–	–	–	1.6	–	12.4	–	24.7	75.1	45.1	–	83.5	85.1	81.6	51.0
	FTA	0.0	41.4	–	–	–	–	0.8	–	6.3	–	3.7	19.5	12.2	–	65.8	43.6	24.0	21.7
LILAC w/o ICL	GA	52.1	99.7	74.6	**100.0**	52.4	**87.0**	99.8	**100.0**	85.8	90.0	86.2	77.9	78.8	66.7	86.4	69.4	59.8	80.4
	PA	67.1	97.0	43.4	94.7	37.4	93.3	58.1	56.2	71.0	71.5	78.2	78.6	50.1	43.2	76.9	23.1	33.9	63.2
	FGA	42.4	91.8	78.3	69.3	90.0	**88.9**	95.0	**97.1**	90.0	82.8	83.2	79.0	77.3	35.4	80.2	43.2	65.6	75.9
	FTA	48.5	72.1	49.3	56.0	68.0	75.0	74.0	71.7	66.7	60.0	66.3	57.4	46.7	21.9	68.9	22.2	30.0	56.2
Label-required Log Parsers																			
UniParser	GA	50.9	94.8	27.5	**100.0**	**100.0**	77.7	98.8	46.1	69.1	85.4	91.8	28.5	73.7	57.9	61.7	85.8	71.9	71.9
	PA	63.4	94.2	28.9	94.8	51.6	94.1	**98.8**	81.7	**88.9**	79.5	94.9	16.4	**68.8**	**65.4**	56.1	74.5	57.3	71.1
	FGA	28.6	68.7	0.9	**96.8**	96.9	66.0	66.1	74.5	62.8	2.0	62.4	45.1	69.9	68.2	54.5	73.7	77.8	59.7
	FTA	45.7	26.9	0.5	58.1	28.9	35.1	51.0	46.2	47.6	1.2	21.9	23.2	28.3	29.0	46.5	54.4	35.7	34.1
LogPPT	GA	98.9	78.6	27.7	72.1	53.4	78.2	96.7	99.8	48.3	47.6	24.5	20.5	54.4	56.4	95.1	29.6	8.7	58.3
	PA	**100.0**	94.8	65.4	94.3	40.6	**99.7**	84.5	**99.7**	66.6	95.2	93.8	16.8	39.0	40.1	**86.0**	9.0	6.5	66.6
	FGA	87.0	60.5	8.1	39.1	87.4	78.0	91.8	94.7	52.6	37.4	25.3	71.2	49.3	21.6	92.5	20.9	19.9	55.1
	FTA	95.7	36.8	10.5	31.2	73.8	76.8	80.9	82.2	43.4	29.9	26.1	42.8	27.4	11.7	85.0	9.0	5.4	45.2
LILAC	GA	**100.0**	**100.0**	74.8	**100.0**	**100.0**	86.9	**100.0**	**100.0**	92.6	**99.9**	91.1	82.5	81.4	79.4	85.6	**95.8**	78.2	91.1
	PA	**100.0**	97.2	**99.9**	94.7	**95.2**	93.7	68.5	60.4	73.2	68.9	90.4	**81.4**	56.4	53.3	77.7	**78.9**	48.6	78.7
	FGA	**100.0**	**100.0**	87.7	**96.8**	**100.0**	86.1	**96.7**	**97.1**	**94.1**	**89.7**	**88.5**	86.1	**87.1**	85.9	86.9	84.7	85.2	**91.3**
	FTA	**100.0**	86.2	84.9	71.0	83.3	76.4	**83.5**	77.5	69.5	63.6	72.1	64.4	48.7	55.7	77.2	61.4	53.6	72.3
Our proposed method																			
LUNAR	GA	98.9	**100.0**	**78.0**	**100.0**	**100.0**	86.4	99.3	**100.0**	**94.1**	97.5	**95.5**	83.0	**87.7**	**86.5**	**97.0**	93.0	**90.5**	**93.4**
	PA	**100.0**	**99.9**	72.2	**100.0**	90.4	99.0	85.1	96.2	83.7	**99.5**	**98.4**	73.5	58.3	59.5	**86.0**	73.5	**74.5**	**85.3**
	FGA	87.0	**100.0**	**92.3**	**96.8**	**100.0**	83.3	88.5	**97.1**	92.6	88.8	87.3	**87.4**	86.3	**86.9**	**96.1**	**88.4**	**92.6**	**91.3**
	FTA	95.7	**89.7**	**92.3**	**94.6**	**87.5**	**83.3**	81.2	**84.3**	**70.2**	**65.7**	**78.9**	**72.1**	**52.6**	**57.4**	**87.0**	**67.4**	**73.8**	**78.5**

primarily due to the limitations of syntax-based parsers, which rely solely on manually crafted rules and thus struggle with complex and diverse log data. Additionally, LogMine achieves the lowest scores in all four metrics and fails to complete the parsing process on 7 datasets within the allocated time. Furthermore, leveraging the extensive pre-trained knowledge of LLMs, LILAC w/o ICL demonstrates superior performance than syntax-based methods in PA, FGA, and FTA. However, our label-free LLM-based parser, LUNAR, significantly outperforms LILAC w/o ICL, with an average improvement of 12.4% in FGA and 19.5% in FTA, respectively. These substantial improvements underscore the effectiveness of LUNAR in harnessing the zero-shot capabilities of LLMs without labelled examples.

Compared with supervised baseline parsers, LUNAR demonstrate comparable performance to existing state-of-the-art parser LILAC. Specifically, LUNAR outperformed LILAC by attaining higher average scores in GA (93.4%), PA (85.3%), and FTA (78.5%), and matched the top FGA score of 91.3%. LILAC retrieves similar logs and labelled templates to provide demonstrations during parsing, which are often inaccessible in real-world systems. However, LUNAR achieves a comparable performance to the state-of-the-art while does not need labelled templates, which is more robust and generalizable in practice. Additionally, compared to UniParser and LogPPT, LUNAR has achieved substantial improvements with an average improvements of 27.8% and 30.8% in all four metrics. This result again highlights the advantages of LUNAR, which is not only effective, but also requires no labelled data for training. The comparable performance of LUNAR indicates its potential to be applied in real-world production systems, which can address the label dependency problems of semantic-based log parsers as mentioned in Section 2.1.

On three industry datasets, LUNAR achieves the highest performance in template-level metrics, showing an average improvement of 6.8% in FGA and 12.0% in FTA compared to the label-required

LLM parser LILAC. This indicates that LUNAR is capable of parsing not only open-source logs but also domain-specific logs that are unfamiliar to the LLM. Specifically on the LoFI dataset, LUNAR substantially outperforms all unsupervised and label-required baselines across all four metrics. Since the ground-truth templates in LoFI are manually annotated by us and have not been exposed to the pre-training corpus of LLMs, LLMs might possess limited awareness of this system. By following our carefully designed instructions for comparing LCUs, LUNAR effectively understand the task and correctly generates templates. The outstanding performance on industry logs highlights LUNAR's generalizability, demonstrating its potential for use in industrial domain-specific systems where logs and templates are previously unseen.

5.2 RQ2: How do different components affect LUNAR?

5.2.1 Designs. In this research question, we first assess the individual contributions of the designs in the two components of LUNAR: the hierarchical log sharder and the two-stage LCU selector. To accomplish this, we have implemented several variants of LUNAR by either removing or replacing the designs in these components. Specifically, for the hierarchical log sharder, we created two variants by removing the respective clustering stages. Additionally, for the two-stage LCU selector, we replaced the hybrid LCU ranking algorithm with various alternatives: random selection, simple selection based on minimum or maximum Jaccard similarity, and selection of consecutive log messages for querying the LLMs. To mitigate the impact of randomness, we also repeated experiment three times and calculated the mean scores as final results.

The average metric scores across all datasets of Loghub-2.0 are presented in Table 3. When the log length-based clustering is removed from the log sharder, the end-to-end accuracy of LUNAR drops from 93.4% to 89.1% and from 79.0% to 78.3%, respectively. Similar decreases are observed for the variant without top-k tokens clustering. These results indicate that both stages within the hierarchical log sharder contribute to the overall performance of LUNAR. On the other hand, for the four variants related to the two-stage LCU selector, all four metrics show decreases across the board. For example, replacing the hybrid LCU ranking algorithm with random selection results in average GA and PA scores dropping by 3.0% and 2.9%, respectively. Notably, selecting the LCU based on maximum similarities among log messages has the most detrimental impact on LUNAR's performance, resulting in a 4.6% decrease in GA and an 4.0% decrease in FTA. This is primarily because selecting only the most similar log messages as an LCU for prompting prevents the LLM from accurately identifying templates and parameters by contrasting the tokens of these log messages. These results demonstrate that our proposed two-stage LCU selector is effective in selecting LCUs, thereby enabling LLMs to accurately parse log messages.

5.2.2 Configurations. In addition to the two key designs of LUNAR mentioned above, we have identified two configurations that could affect LUNAR's performance: the LCU sample size and the λ values for LCU nomination. These configurations directly determine the LCUs for each query provided to the LLMs, and thus, they may significantly impact the accuracy of the LLMs' parsed results. In this part, we explore different settings for the LCU sample size and the λ values to evaluate their impact on LUNAR's performance across all four evaluation metrics. Specifically, we first chose the default λ value (*i.e.*, 0.6) and varied the LCU sample sizes from 1 to 5. Additionally, we fixed the default LCU sample size (*i.e.*, 3) and varied the λ value from 0.0 to 1.0. The average metric scores under different settings across all datasets of Loghub-2.0 are presented in Figure 8.

Based on the results shown in the left sub-figure of Figure 8, we observe that the performance of LUNAR remains consistently high across different LCU sample sizes. Notably, when the LCU sample size is set to 3, GA, PA and FTA achieve their highest scores. Conversely, smaller sample sizes result in a significant lower accuracy. For instance, with a sample size of 1, the average GA and

Table 3. Ablation study of components in LUNAR (%).

Metrics	GA	PA	FGA	FTA
LUNAR	93.4	86.8	91.0	79.0
Variants w.r.t Log Sharder				
w/o log length	89.1 (↓ 4.6%)	83.2 (↓ 4.2%)	90.1 (↓ 1.0%)	78.3 (↓ 0.9%)
w/o top-k tokens	89.1 (↓ 4.6%)	82.9 (↓ 4.5%)	90.4 (↓ 0.6%)	78.2 (↓ 0.9%)
Variants w.r.t LCU Selector				
w/ random selection	90.6 (↓ 3.0%)	84.3 (↓ 2.9%)	90.5 (↓ 0.5%)	77.7 (↓ 1.6%)
w/ minimum similarity	88.8 (↓ 4.6%)	83.9 (↓ 4.5%)	89.5 (↓ 1.9%)	80.2 (↓ 1.4%)
w/ maximum similarity	89.1 (↓ 4.6%)	82.8 (↓ 4.6%)	88.9 (↓ 2.3%)	75.8 (↓ 4.0%)
w/ consecutive selection	89.2 (↓ 4.5%)	82.6 (↓ 4.8%)	90.2 (↓ 0.9%)	77.5 (↓ 1.9%)

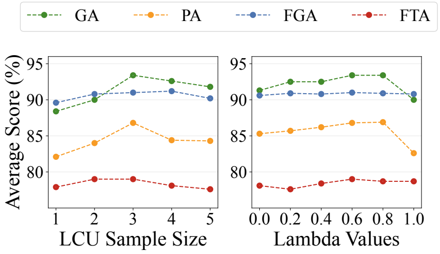


Fig. 8. Sensitivity of configurations in LUNAR.

PA scores drop to approximately 88% and 82%, respectively, which is 5.0% and 4.7% lower than the scores achieved with a sample size of 3. This suggests that the absence of contrastive log messages can impair the parsing ability of LLMs. On the other hand, increasing the LCU sample size to 4 or 5 results in a downward trend in accuracy. For example, the FTA score decreases from 79.0% to 77.6%. This indicates that larger LCU sample sizes may introduce additional noise and thus hurt the accuracy of LUNAR. Therefore, we set the default LCU sample size to 3 in our experiments.

The λ values for LCU nomination take into account both variability and commonality scores when selecting LCUs for prompting. A higher λ value places more emphasis on variability, while a lower λ value prioritizes commonality. As shown in Figure 8, the optimal average accuracy is achieved when the λ value is set to 0.6, indicating a balanced consideration of variability and commonality in log messages within the LCU. When the λ value is low, such as 0.0 or 0.2, all metrics are significantly lower compared to a value of 0.6, underscoring the importance of variability in log messages within LCUs. For instance, at a λ value of 0.0, the average GA and PA scores are 2.1% and 1.5% lower than at 0.6. Conversely, performance also declines when the λ value is too high, particularly when $\lambda = 1.0$. For example, both GA and PA scores drop by 3.4% and 4.2% when the λ value increases from 0.6 to 1.0. These results illustrate that both variability and commonality scores effectively filters log messages belonging to the same template within the LCUs, which enable LLMs to accurately parse the log template.

5.3 RQ3: How effective is LUNAR integrated with different LLM and prompts?

5.3.1 Prompt Design. Parsing prompt is a critical element that instructs LLM to utilize its pretrained knowledge for log parsing. In this research question, we first evaluate the contribution of four key components of our parsing prompt. To this end, we create four prompt variants by removing or substituting specific components, including removing parameter examples, removing output constraints, removing parameter advice, and using simpler basic requirements from LILAC [26].

Table 4 shows the average metrics of LUNAR on Loghub-2.0 with different prompt variants, from which we observe a consistent superior performance of our parsing prompt. In detail, replacing or removing lead to consistent performance degradation, demonstrating that our prompt can effectively instruct LLM to produce accurate log templates. Among the four variants, removing output constraints results in the most significant drop in majority of the four metrics, *e.g.*,, FTA drops by 40.3%. Without a clear instruction on the output format, the LLM tends to generate predictions in a more unstructured format, complicating the process of extracting and aggregating log templates from responses. Parameter Example and parameter advice significantly impact template accuracy metrics (*i.e.*, PA and FTA) but have minimal impacts on grouping accuracy metrics (*i.e.*, GA and FGA). For example, removing parameter advice leads to a 2.7% reduction in FTA while maintaining

Table 4. Ablation study of parsing prompt and LLMs in LUNAR (%)

Metrics	**GA**	**PA**	**FGA**	**FTA**
LUNAR (GPT-3.5)	93.4	86.8	91.0	79.0
Variants w.r.t Parsing Prompt				
w/o parameter examples	91.9 (↓ 1.6%)	84.1 (↓ 3.1%)	90.6 (↓ 0.4%)	77.0 (↓ 2.5%)
w/o output constraints	87.3 (↓ 6.5%)	70.1 (↓ 19.2%)	63.5 (↓ 30.2%)	47.2 (↓ 40.3%)
w/o parameter advice	93.3 (↓ 0.1%)	84.9 (↓ 2.2%)	91.0 (↓ 0.0%)	76.9 (↓ 2.7%)
w/ simpler requirements	89.3 (↓ 4.4%)	82.9 (↓ 4.5%)	87.4 (↓ 4.0%)	76.2 (↓ 3.5%)
Variants w.r.t LLM				
GPT-4	93.4 (↑ 0.0%)	87.0 (↑ 0.2%)	92.1 (↑ 1.2%)	79.8 (↑ 1.0%)
GPT-4o	86.2 (↓ 7.7%)	79.5 (↓ 8.4%)	88.5 (↓ 2.7%)	75.1 (↓ 4.9%)
Claude	90.4 (↓ 3.2%)	84.1 (↓ 3.1%)	90.0 (↓ 1.1%)	78.1 (↓ 1.1%)
Llama-3.1-405B	92.9 (↓ 0.5%)	86.9 (↑ 0.1%)	90.8 (↓ 0.2%)	80.1 (↑ 1.4%)
DeepSeek-V2.5-236B	90.1 (↓ 3.5%)	82.3 (↓ 5.2%)	90.3 (↓ 0.8%)	78.2 (↓ 1.0%)
Qwen2-72B	89.6 (↓ 4.1%)	83.1 (↓ 4.3%)	89.0 (↓ 2.2%)	76.1 (↓ 3.7%)
Llama-3.1-70B	86.6 (↓ 7.3%)	80.6 (↓ 7.1%)	89.0 (↓ 2.2%)	77.4 (↓ 2.0%)
Llama-3.1-8B	90.8 (↓ 2.8%)	75.4 (↓ 13.1%)	88.2 (↓ 3.1%)	65.3 (↓ 17.3%)
Qwen2-7B	86.0 (↓ 7.9%)	68.1 (↓ 21.5%)	87.3 (↓ 4.1%)	59.3 (↓ 24.9%)

the same FGA. The reason could be that parameter examples and advice can help LLMs recognize complex parameters, such as long URLs, which are often partially identical and can otherwise be mistakenly recognized as a static part. Additionally, simpler requirements from LILAC consistently underperform the detailed prompt, resulting a 3.5% to 4.5% drop on four metrics, primarily due to its less informative task instructions.

5.3.2 LLMs. In our experiments, we use GPT-3.5 (*gpt-3.5-turbo-0125*) as the default LLM. In this RQ, we comprehensively evaluate how LUNAR performs when integrated with various LLMs. Specifically, we select three popular closed-source LLMs, *i.e.*, GPT-4 (*gpt-4-0613*) [51], GPT-4o (*gpt-4o-0806*)[52] and Claude (*claude-3-5-sonnet-20240620*)[3], along with six leading open-source LLMs with parameter sizes ranging from 7B to 405B. These open-source LLMs include DeepSeek-V2.5-236B (*DeepSeek-V2.5*) [8], the Llama-3.1 family (*Meta-Llama-3.1-8B-Instruct*, *Meta-Llama-3.1-70B-Instruct*, *Meta-Llama-3.1-405B-Instruct*) [12] and the Qwen2 family (*Qwen2-7B-Instruct* and *Qwen2-72B-Instruct*) [66].

Table 4 shows the average metrics of LUNAR using different LLMs on Loghub-2.0, from which we can find consistent high performance across all LLMs. Specifically, LUNAR with GPT-4 consistently outperforms both GPT-4o and Claude by a substantial margin, while showing a modest 0.6% average accuracy advantage over GPT-3.5. This could be attributed to the enhanced reasoning capabilities of GPT-4, powered by its larger parameter size and extensive training corpus. Despite this, GPT-3.5 is 40 times cheaper [53] while providing performance comparable with GPT-4, making GPT-3.5 our preferred choice for default LLMs due to this cost-efficiency. Among commercial LLMs GPT-4o performs the least effectively with LUNAR, which we link to its smaller parameter size and potentially suboptimal prompt instructions [23]. For open-source LLMs, LUNAR with Llama-3.1-405B exhibits comparable performance to LUNAR with GPT-3.5, with slight improvements in FTA and PA, but marginally lower GA and FGA. These results demonstrate that LUNAR is adaptable to various LLMs while preserving high accuracy. Additionally, we observe a clear trend where larger models (i.e., Llama-3.1-405B and DeepSeek-V2.5-236B) outperform medium-sized models (i.e., Qwen2-72B and Llama-3.1-70B), which in turn outperform smaller models (i.e., Llama-3.1-8B and Qwen2-7B) across all four metrics. This performance enhancement suggests that LLM-based log parsing may adhere to the scaling law [29], where larger size of models and pre-training datasets leads to enhanced performance.

5.4 RQ4: How efficient and cost-effective is LUNAR in parsing large-scale log data?

5.4.1 Efficiency Analysis. Efficiency is an essential factor for log parsers in real-world applications due to the substantial volume of logs produced [61, 72]. In this RQ, we first evaluate the time efficiency of LUNAR along with baseline parsers by applying them on the large-scale datasets from Loghub-2.0, each comprising an average of 3.6 million log messages. Specifically, we recorded the execution times for each log parser with default parameters when parsing the all log datasets in Loghub-2.0, subsequently calculating the mean parsing time for all datasets. LogMine is excluded in this RQ as its execution time significantly exceeds 12 hours, which is slower than the other baseline methods by a large margin. For LUNAR, we use the default parameters introduced in Section 4.4 for evaluation. Since LUNAR is designed to enable parallelism for processing log buckets, we compute the average parsing time for both serial mode and parallel mode of LUNAR. The efficiency results are demonstrated in the Figure 9.

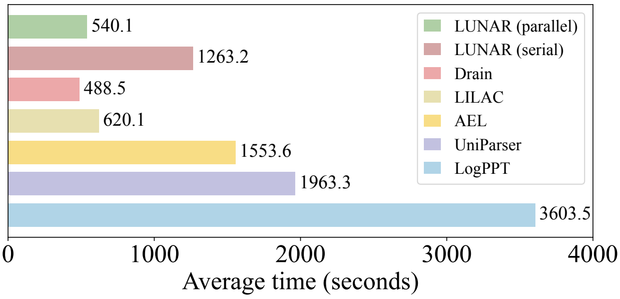


Fig. 9. Efficiency of LUNAR and baselines (second).

The results demonstrate that LUNAR achieves higher efficiency compared to the most efficient semantic-based log parser, LILAC, and comparable efficiency to the most efficient syntax-based log parser, Drain. Specifically, the average parsing time for 3.6 million log messages in serial mode using LUNAR is 1263.2 seconds. In practice, we can leverage the parallel mode of LUNAR to achieve better efficiency, reducing the parsing time to just 540.1 seconds. For the most efficient syntax-based parser, Drain, which only takes an average of 488.5 seconds to parse each dataset, LUNAR achieve comparable efficiency, only slower by 10.6%. In contrast, the most efficient semantic-based parser, LILAC, takes an average of 620.1 seconds, which is 14.8% slower than LUNAR. Notably, even with GPU acceleration, other semantic-based parsers such as UniParser and LogPPT require significantly more time to parse large-scale log data. LUNAR outperforms them by 3.64 times and 6.67 times in efficiency, respectively. These results indicate that LUNAR is efficient in parsing large-scale log data, making it suitable for application in real-world production systems.

5.4.2 Cost Analysis. LUNAR utilizes LLMs to read prompt instructions and generate templates. Unlike traditional syntax-based and semantic-based methods, which can run efficiently on CPUs or low-memory GPUs with low computational costs, LLM-based approaches incur additional costs due to their reliance on cloud-based inference. Given the substantial GPU demands of LLMs, LLM service providers, like OpenAI, offer these as remote services, providing APIs to handle input prompts and return outputs based on token usage charges. As a result, each LLM query involves a financial cost tied to the number of input and output tokens. In this research question, we conduct a financial cost analysis of LUNAR compared to baseline methods to evaluate its practical feasibility. Specifically, we compare the cost and accuracy of LUNAR with four methods, including two LLM-based baselines (*i.e.*, LILAC and LILAC w/o ICL) and two top-performing syntax-based baselines (*i.e.*, Drain and Brain). We report the (1) average accuracy of four metrics, (2) number of input and output tokens consumed per dataset, (3) number of LLM invocations, and (4) total dollar cost per dataset [53].

Table 5. Cost analysis

	Avg. Accuracy	# Input Tokens	# Output Tokens	# Invocation	$ Cost
Drain	53.6	-	-	0	0
Brain	58.6	-	-	0	0
LILAC w/o ICL	68.9	116.3K	26.8K	726.3	$0.098
LILAC	83.3	68.4K	4.5K	219.0	$0.041
LUNAR	87.1	131.7K	13.9K	269.4	$0.086

Table 5 shows the costs of different methods. We can observe that: (1) compared to top-performing syntax-based methods, LUNAR achieves a substantial improvement in average accuracy, with gains of 33.5% and 28.5% respectively. This improvement comes at an additional cost of $ 0.045 per dataset for LLM API usages. Given the significant accuracy boost, this additional expense represents a reasonable trade-off, making LUNAR a viable choice where accuracy is a priority. (2) Compared to LILAC, LUNAR incurs an additional cost of $0.30 per dataset while achieving an average accuracy improvement of 3.8%. The increased cost primarily stems from the use of more detailed instructions and multiple logs in an LCU, indicated by the approximately 1.5 times higher input token consumption. However, LUNAR eliminates the costs of human annotation and demonstration selection required by LILAC, making it a fully label-free alternative with reduced manual effort. (3) Compared to LILAC w/o ICL, LUNAR achieves a notably higher accuracy (+18.2%) while reducing cost by $0.012 and using significantly fewer LLM calls. This underscores LUNAR's efficacy in unsupervised log parsing, where incorporating detailed prompts and LCUs boosts performance without introducing excessive costs. Overall, these findings demonstrate that LUNAR offers a cost-effective and fully label-free solution for log parsing. By balancing cost and accuracy, LUNAR provides a practical alternative for practitioners seeking high-performance unsupervised methods with minimal financial overhead.

6 THREATS TO VALIDITY

We have identified the following major threats to validity:

Data leakage Given that large language models (LLMs) are trained on extensive datasets, one potential risk is data leakage. Specifically, the LLM used in LUNAR might have been trained on open-source log datasets, which could lead to memorizing ground-truth templates instead of performing genuine inference. However, experimental results show that simply using LLMs (LILAC w/o ICL) yields much lower performance than LUNAR, suggesting a low chance of direct memorization. Moreover, we present a new parsing dataset, LoFI, sourced from an industrial online service system, with ground-truth templates manually annotated by us. The experimental results on LoFI indicate that LUNAR can still outperform baselines by a large margin, even with previously unseen logs.

Randomness Randomness can influence the performance of LUNAR and other baseline methods. To mitigate this issue, we minimized the randomness of the LLM by setting the temperature to 0, ensuring consistent outputs for the same input text. Additionally, we conducted each experiment three times for every experimental setting and used the average of these results as the final outcome.

Implementation and settings To mitigate the bias of implementation and settings, in our evaluation, we compared our LUNAR with state-of-the-art approaches within the same evaluation framework. We adopted the implementations from their replication packages and benchmarks, using the parameters and settings (*e.g.*, number of log templates and similarity threshold) optimized by previous work [27, 72]. Moreover, the results of the baseline approaches are consistent with the best results in recent benchmarks.

7 Related Works

Log parsing is a critical preliminary step for various log analysis tasks [18, 30], including anomaly detection and root cause analysis. Therefore, numerous efforts have been made to achieve accurate and efficient log parsing [7, 11, 45, 46, 48, 49, 56, 57, 59, 67]. These log parsers can be categorized into two types: unsupervised syntax-based and supervised semantic log parsers. Unsupervised syntax-based log parsers leverage predefined rules or heuristics to extract the constant parts of log messages as log templates. For instance, SLCT [58] was the first approach to use token frequencies to determine log templates and parameters. Drain [17] utilizes a fixed-depth prefix tree structure to effectively extract commonly occurring templates based on specific heuristics (*i.e.*, prefix tokens and log length). These syntax-based log parsers do not rely on manually labelled examples. However, their parsing accuracy can significantly decline when log data do not conform to predefined rules or handcrafted features, limiting the adaptability and usability [26, 64]. Among these unsupervised methods, LogMine [15] applies a similar parsing paradigm to LUNAR, which first partitions log messages by a bottom-up clustering algorithm and then extract log templates from each cluster. However, its extraction relies on preset parameter rules, restricting it to a narrow range of parameter types and resulting in suboptimal accuracy.

On the other hand, semantic-based log parsers utilize neural networks [22, 37, 42, 68] or language models [31] to identify log templates and parameters by understanding the semantics of log messages. They require human-labelled log templates to train or tune the models by learning the semantics and patterns in the labelled log data. UniParser [42] is one of the pioneering work in parsing logs with a focus on their semantic meaning. It integrates a BiLSTM-based semantics miner with a joint parser to identify log templates. Additionally, LogPPT [31] proposed identifying log templates and parameters using prompt-based few-shot learning, based on the RoBERTa model. Recently, with the rise of large language models (LLMs), a series of LLM-based log parsers have been developed to achieve more effective log parsing. These LLM-based log parsers leverage fine-tuning [41] or in-context learning [26, 64] to specialize LLMs for log parsing tasks, thereby achieving remarkable performance. However, these semantic-based log parsers heavily rely on labelled data, which limits their ability to generalize to different types of log data or evolving log data [27]. In contrast, our proposed unsupervised LLM-based log parser, LUNAR, does not require labelled examples, allowing it to generalize to diverse and evolving log data.

8 Conclusion

In this work, we propose an LLM-based unsupervised log parser named LUNAR, which leverages log contrastive units (LCUs) to facilitate effective comparisons by the LLM. To efficiently identify effective LCUs from large-scale log data, LUNAR employs a hierarchical sharder to divide logs into buckets, thereby reducing sampling overhead and enabling parallel computation. Additionally, LUNAR incorporates a hybrid LCU selector that jointly measures the commonality and variability of LCUs, which is crucial for prompting LLMs. Furthermore, LUNAR introduces an improved prompt format to guide LLMs in a zero-shot setting and a template revision strategy to globally improve templates. Experimental results on 14 large-scale public datasets show that LUNAR achieves high parsing accuracy, significantly outperforming other unsupervised parsers and being comparable to state-of-the-art parsers that require labelled data. In terms of efficiency, LUNAR is also comparable to the fastest baseline parser, highlighting its potential for application in real-world systems.

Acknowledgment

We extend our sincere gratitude to the anonymous reviewers for their constructive feedback. We also thank Jinyang Liu and Jiazhen Gu for their valuable discussions. The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. SRFS2425-4S03 of the Senior Research Fellow Scheme and No. CUHK 14209124 of the General Research Fund) and supported by the National Natural Science Foundation of China (No. 62402536).

Data Avalability

The code and data are available at: https://github.com/Jun-jie-Huang/LUNAR.

References

[1] 2023. OpenAI API. https://openai.com/blog/openai-api [Online; accessed 5 April 2025].

[2] Anunay Amar and Peter C Rigby. 2019. Mining historical test logs to predict bugs and localize faults in the test logs. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 140–151.

[3] AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card* (2024).

[4] Vincent Bushong, Russell Sanders, Jacob Curtis, Mark Du, Tomas Cerny, Karel Frajtak, Miroslav Bures, Pavel Tisnovsky, and Dongwan Shin. 2020. On matching log analysis to source code: A systematic mapping study. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*. 181–187.

[5] An Ran Chen, Tse-Hsun Chen, and Shaowei Wang. 2021. Pathidea: Improving information retrieval-based bug localization by re-constructing execution paths using logs. *IEEE Transactions on Software Engineering (TSE)* 48, 8 (2021), 2905–2919.

[6] Zhuangbin Chen, Jinyang Liu, Wenwei Gu, Yuxin Su, and Michael R Lyu. 2021. Experience report: Deep learning-based system log analysis for anomaly detection. *arXiv preprint arXiv:2107.05908* (2021).

[7] Hetong Dai, Heng Li, Che-Shao Chen, Weiyi Shang, and Tse-Hsun Chen. 2020. Logram: Efficient log parsing using n n-gram dictionaries. *IEEE Transactions on Software Engineering* 48, 3 (2020), 879–892.

[8] DeepSeek-AI. 2024. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv:2405.04434 [cs.CL]

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[10] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 1107–1128.

[11] Min Du and Feifei Li. 2016. Spell: Streaming parsing of system event logs. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 859–864.

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[13] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* 46, 4 (2014), 1–37.

[14] Shuzheng Gao, Xin-Cheng Wen, Cuiyun Gao, Wenxuan Wang, and Michael R Lyu. 2023. Constructing Effective In-Context Demonstration for Code Intelligence Tasks: An Empirical Study. *arXiv preprint arXiv:2304.07575* (2023).

[15] Hossein Hamooni, Biplob Debnath, Jianwu Xu, Hui Zhang, Guofei Jiang, and Abdullah Mueen. 2016. Logmine: Fast pattern recognition for log analytics. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM)*. 1573–1582.

[16] Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303* (2022).

[17] Pinjia He, Jieming Zhu, Zibin Zheng, and Michael R Lyu. 2017. Drain: An online log parsing approach with fixed depth tree. In *2017 IEEE international conference on web services (ICWS)*. IEEE, 33–40.

[18] Shilin He, Pinjia He, Zhuangbin Chen, Tianyi Yang, Yuxin Su, and Michael R Lyu. 2021. A survey on automated log analysis for reliability engineering. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–37.

[19] Shilin He, Xu Zhang, Pinjia He, Yong Xu, Liqun Li, Yu Kang, Minghua Ma, Yining Wei, Yingnong Dang, Saravanakumar Rajmohan, et al. 2022. An empirical study of log analysis at Microsoft. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE)*. 1465–1476.

[20] Shilin He, Jieming Zhu, Pinjia He, and Michael R Lyu. 2020. Loghub: A large collection of system log datasets towards automated log analytics. *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)* (2020).

[21] Junjie Huang, Zhihan Jiang, Jinyang Liu, Yintong Huo, Jiazhen Gu, Zhuangbin Chen, Cong Feng, Hui Dong, Zengyin Yang, and Michael R Lyu. 2024. Demystifying and Extracting Fault-indicating Information from Logs for Failure Diagnosis. In *2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 511–522.

[22] Yintong Huo, Yuxin Su, Cheryl Lee, and Michael R Lyu. 2023. Semparser: A semantic parser for log analytics. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 881–893.

[23] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[24] Guochao Jiang, Zepeng Ding, Yuchen Shi, and Deqing Yang. 2024. P-ICL: Point In-Context Learning for Named Entity Recognition with Large Language Models. *arXiv preprint arXiv:2405.04960* (2024).

[25] Zhihan Jiang, Junjie Huang, Zhuangbin Chen, Yichen Li, Guangba Yu, Cong Feng, Yongqiang Yang, Zengyin Yang, and Michael R Lyu. 2025. L4: Diagnosing Large-scale LLM Training Failures via Automated Log Analysis. *arXiv preprint*

arXiv:2503.20263 (2025).

[26] Zhihan Jiang, Jinyang Liu, Zhuangbin Chen, Yichen Li, Junjie Huang, Yintong Huo, Pinjia He, Jiazhen Gu, and Michael R Lyu. 2024. Lilac: Log parsing using llms with adaptive parsing cache. *Proceedings of the ACM on Software Engineering* 1, FSE (2024), 137–160.

[27] Zhihan Jiang, Jinyang Liu, Junjie Huang, Yichen Li, Yintong Huo, Jiazhen Gu, Zhuangbin Chen, Jieming Zhu, and Michael R Lyu. 2024. A large-scale evaluation for log parsing techniques: How far are we?. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*. 223–234.

[28] Zhen Ming Jiang, Ahmed E Hassan, Parminder Flora, and Gilbert Hamann. 2008. Abstracting execution logs to execution events for enterprise applications (short paper). In *2008 The Eighth International Conference on Quality Software*. IEEE, 181–186.

[29] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[30] Zanis Ali Khan, Donghwan Shin, Domenico Bianculli, and Lionel Briand. 2022. Guidelines for assessing the accuracy of log message template identification techniques. In *Proceedings of the 44th International Conference on Software Engineering (ICSE)*. 1095–1106.

[31] Van-Hoang Le and Hongyu Zhang. 2023. Log parsing with prompt-based few-shot learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2438–2449.

[32] Van-Hoang Le and Hongyu Zhang. 2024. PreLog: A Pre-trained Model for Log Analytics. *Proceedings of the ACM on Management of Data* 2, 3 (2024), 1–28.

[33] Xiaoyun Li, Hongyu Zhang, Van-Hoang Le, and Pengfei Chen. 2024. Logshrink: Effective log compression by leveraging commonality and variability of log data. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–12.

[34] Yichen Li, Yintong Huo, Zhihan Jiang, Renyi Zhong, Pinjia He, Yuxin Su, Lionel C. Briand, and Michael R. Lyu. 2024. Exploring the Effectiveness of LLMs in Automated Logging Statement Generation: An Empirical Study. *IEEE Transactions on Software Engineering (TSE)* 50, 12 (2024), 3188–3207.

[35] Yichen Li, Yintong Huo, Renyi Zhong, Zhihan Jiang, Jinyang Liu, Junjie Huang, Jiazhen Gu, Pinjia He, and Michael R Lyu. 2024. Go static: Contextualized logging statement generation. *Proceedings of the ACM on Software Engineering* 1, FSE (2024), 609–630.

[36] Yichen Li, Yulun Wu, Jinyang Liu, Zhihan Jiang, Zhuangbin Chen, Guangba Yu, and Michael R Lyu. 2025. COCA: Generative Root Cause Analysis for Distributed Systems with Code Knowledge. *arXiv preprint arXiv:2503.23051* (2025).

[37] Zhenhao Li, Chuan Luo, Tse-Hsun Chen, Weiyi Shang, Shilin He, Qingwei Lin, and Dongmei Zhang. 2023. Did we miss something important? studying and exploring variable-aware log abstraction. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 830–842.

[38] Jinyang Liu, Junjie Huang, Yintong Huo, Zhihan Jiang, Jiazhen Gu, Zhuangbin Chen, Cong Feng, Minzhi Yan, and Michael R Lyu. 2023. Scalable and Adaptive Log-based Anomaly Detection with Expert in the Loop. *arXiv preprint arXiv:2306.05032* (2023).

[39] Jinyang Liu, Jieming Zhu, Shilin He, Pinjia He, Zibin Zheng, and Michael R Lyu. 2019. Logzip: Extracting hidden structures via iterative clustering for log compression. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 863–873.

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[41] Yilun Liu, Shimin Tao, Weibin Meng, Feiyu Yao, Xiaofeng Zhao, and Hao Yang. 2024. Logprompt: Prompt engineering towards zero-shot and interpretable log analysis. In *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings*. 364–365.

[42] Yudong Liu, Xu Zhang, Shilin He, Hongyu Zhang, Liqun Li, Yu Kang, Yong Xu, Minghua Ma, Qingwei Lin, Yingnong Dang, et al. 2022. Uniparser: A unified log parser for heterogeneous log data. In *Proceedings of the ACM Web Conference 2022 (WWW)*. 1893–1901.

[43] Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Z-ICL: Zero-Shot In-Context Learning with Pseudo-Demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 2304–2317.

[44] Zeyang Ma, An Ran Chen, Dong Jae Kim, Tse-Hsun Chen, and Shaowei Wang. 2024. LLMParser: An Exploratory Study on Using Large Language Models for Log Parsing. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–13.

[45] Adetokunbo AO Makanju, A Nur Zincir-Heywood, and Evangelos E Milios. 2009. Clustering event logs using iterative partitioning. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*. 1255–1264.

[46] Salma Messaoudi, Annibale Panichella, Domenico Bianculli, Lionel Briand, and Raimondas Sasnauskas. 2018. A search-based approach for accurate identification of log message formats. In *Proceedings of the 26th Conference on Program Comprehension*. 167–177.

[47] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 11048–11064.

[48] Masayoshi Mizutani. 2013. Incremental mining of system log format. In *2013 IEEE International Conference on Services Computing*. IEEE, 595–602.

[49] Meiyappan Nagappan and Mladen A Vouk. 2010. Abstracting log lines to log event types for mining software system logs. In *2010 7th IEEE Working Conference on Mining Software Repositories (MSR)*. IEEE, 114–117.

[50] Paolo Notaro, Soroush Haeri, Jorge Cardoso, and Michael Gerndt. 2023. LogRule: Efficient Structured Log Mining for Root Cause Analysis. *IEEE Transactions on Network and Service Management* (2023).

[51] OpenAI. 2024. GPT-4. *OpenAI Blog* (2024). https://openai.com/index/gpt-4/

[52] OpenAI. 2024. Hello GPT-4o. *OpenAI Blog* (2024). https://openai.com/index/hello-gpt-4o/

[53] OpenAI. 2024. OpenAI API Pricing. *OpenAI Blog* (2024). https://platform.openai.com/docs/pricing

[54] Daan Schipper, Maurício Aniche, and Arie van Deursen. 2019. Tracing back log data to its log statement: from research to practice. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 545–549.

[55] Weiyi Shang. 2012. Bridging the divide between software developers and operators using logs. In *2012 34th international conference on software engineering (ICSE)*. IEEE, 1583–1586.

[56] Keiichi Shima. 2016. Length matters: Clustering system log messages using length of words. *arXiv preprint arXiv:1611.03213* (2016).

[57] Liang Tang, Tao Li, and Chang-Shing Perng. 2011. LogSig: Generating system events from raw textual logs. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM)*. 785–794.

[58] Risto Vaarandi. 2003. A data clustering algorithm for mining patterns from event logs. In *Proceedings of the 3rd IEEE Workshop on IP Operations & Management (IPOM)(IEEE Cat. No. 03EX764)*. Ieee, 119–126.

[59] Risto Vaarandi and Mauno Pihelgas. 2015. Logcluster-a data clustering and pattern mining algorithm for event logs. In *2015 11th International conference on network and service management (CNSM)*. IEEE, 1–7.

[60] Lingzhi Wang, Nengwen Zhao, Junjie Chen, Pinnong Li, Wenchi Zhang, and Kaixin Sui. 2020. Root-cause metric location for microservice systems via log anomaly detection. In *2020 IEEE international conference on web services (ICWS)*. IEEE, 142–150.

[61] Xuheng Wang, Xu Zhang, Liqun Li, Shilin He, Hongyu Zhang, Yudong Liu, Lingling Zheng, Yu Kang, Qingwei Lin, Yingnong Dang, et al. 2022. SPINE: a scalable log parser with feedback guidance. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE)*. 1198–1208.

[62] Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. Self-Improving for Zero-Shot Named Entity Recognition with Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. 583–593.

[63] Junjielong Xu, Qiuai Fu, Zhouruixing Zhu, Yutong Cheng, Zhijing Li, Yuchi Ma, and Pinjia He. 2023. Hue: A user-adaptive parser for hybrid logs. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 413–424.

[64] Junjielong Xu, Ruichun Yang, Yintong Huo, Chengyu Zhang, and Pinjia He. 2024. DivLog: Log Parsing with Prompt Enhanced In-Context Learning. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–12.

[65] Wei Xu, Ling Huang, Armando Fox, David Patterson, and Michael Jordan. 2009. Largescale system problem detection by mining console logs. *Proceedings of SOSP'09* (2009).

[66] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).

[67] Siyu Yu, Pinjia He, Ningjiang Chen, and Yifan Wu. 2023. Brain: Log Parsing with Bidirectional Parallel Tree. *IEEE Transactions on Services Computing (TSC)* (2023).

[68] Siyu Yu, Yifan Wu, Zhijing Li, Pinjia He, Ningjiang Chen, and Changjian Liu. 2023. Log Parsing with Generalization Ability under New Log Types. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 425–437.

[69] Chenxi Zhang, Xin Peng, Chaofeng Sha, Ke Zhang, Zhenqing Fu, Xiya Wu, Qingwei Lin, and Dongmei Zhang. 2022. DeepTraLog: Trace-log combined microservice anomaly detection through graph-based deep learning. In *Proceedings of the 44th International Conference on Software Engineering (ICSE)*. 623–634.

[70] Xu Zhang, Yong Xu, Qingwei Lin, Bo Qiao, Hongyu Zhang, Yingnong Dang, Chunyu Xie, Xinsheng Yang, Qian Cheng, Ze Li, et al. 2019. Robust log-based anomaly detection on unstable log data. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE)*.

807–817.

[71] Nengwen Zhao, Honglin Wang, Zeyan Li, Xiao Peng, Gang Wang, Zhu Pan, Yong Wu, Zhen Feng, Xidao Wen, Wenchi Zhang, et al. 2021. An empirical investigation of practical log anomaly detection for online service systems. In *Proceedings of the 29th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering (FSE).* 1404–1415.

[72] Jieming Zhu, Shilin He, Jinyang Liu, Pinjia He, Qi Xie, Zibin Zheng, and Michael R Lyu. 2019. Tools and benchmarks for automated log parsing. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP).* IEEE, 121–130.

Received 2025-02-27; accepted 2025-04-01