

## Dear SIGMOD Meta-Reviewer and Reviewers,

Thank you for your encouraging comments, based on which we have made a substantial revision, with the updated contents highlighted in blue. Below, we first summarize the **main changes** in our revision, followed by **detailed responses** to each reviewer.

**M1 [More Baselines]:** We have added more baselines. (1) For PLMs, we augment the training data for the PLM using our proposed techniques in Section 4.2, and introduce a new baseline *DA + PLM*.

(2) For LLMs, we add the following new baselines.

- We replace GPT with other LLMs, *i.e.*, using Qwen2 and Llama-3 at two different scale levels (less than 10B and larger than 70B) for the new baselines, *Qwen*, *Llama-3*, *Qwen (all knobs)*, *Llama-3 (all knobs)*, *ANDROMEDA-Qwen*, and *ANDROMEDA-Llama-3*.
- We consider *Prompt Engineering* baselines, such as chain-of-thought and task decomposition.
- We consider an *SFT* baseline that fine-tunes GPT-3.5 using our training data.

We provide details of the above baselines in **Section 6.1**, report the results in **Table 2** and analyze the results **Exp-1 in Section 6.2**. The results show that ANDROMEDA outperforms all the baselines.

**M2 [Larger Datasets]:** We have increased the number of test questions to 100 in the Runnable setting, as shown in **Table 1**.

However, it is challenging to expand the datasets in the NL setting, as we have already gathered nearly all knob-configuration-related questions for MySQL and PostgreSQL from StackOverflow and the MySQL forum.

**M3 [New Experiments: End-to-end Time Reduction]:** We have added a new experiment to evaluate whether ANDROMEDA effectively reduces the end-to-end time. Specifically, we employ experienced DBAs and provide them with historical questions, database manuals, and telemetry data. We then consider three strategies: (1) using *manuals*: DBAs manually search DBMS manuals to provide diagnostic suggestions for tuning knobs; (2) using *search engines*: DBAs use search engine to provide diagnostic suggestions for tuning knobs; (3) using *Andromeda*: DBAs follow ANDROMEDA’s recommendations to address the issues. We compare the average time taken by the DBAs to troubleshoot configuration issues, thereby demonstrating the effectiveness of ANDROMEDA.

The results are reported in **Table 9**, and the analysis is presented in **Exp-7 in Section 6.5**. We find that ANDROMEDA can effectively reduce the end-to-end time (*i.e.*, the operational burden) for DBAs, while still preserving higher SuccessRate.

**M4 [New Experiments: Sensitivity to Noise Sources]:** We have conducted experiments to examine whether inaccuracies in the retrieval sources (*e.g.*, historical questions, troubleshooting manuals and telemetry data) impact the overall performance of ANDROMEDA. We manually inject noise into these sources by replacing the ground-truth labels in historical questions or database manuals with incorrect ones at a various probability (*e.g.*, 10%).

The results are reported in **Exp-4 in Section 6.3**, which show that inaccuracies in the retrieval sources impact the performance, highlighting the importance of the quality of retrieval sources.

**M5 [New Experiments: Scalability of Telemetry Data Analysis]:** We have added new experiments to examine the scalability

of our approach for detecting relevant telemetry data. Specifically, we vary the timestamp interval of each metric and the number of metrics, and report the time required for telemetry data analysis.

As reported in **Exp-8 in Section 6.5**, ANDROMEDA performs well in terms of scalability for telemetry data analysis.

**M6 [New Experiments: Cost Efficiency]:** We have reported the cost efficiency, which is the average cost of calling LLMs for each NL debugging question in **Exp-9 in Section 6.5**. We find that ANDROMEDA has good performance on cost efficiency.

**M7 [More In-Depth Experimental Analysis]:**

(1) We have provided more in-depth analysis by categorizing questions by their types in **Exp-5 in Section 6.4** for evaluating telemetry data analysis, and by their frequencies in **Exp-3 in Section 6.3** for evaluating document retrieval.

(2) We have reported precision and recall in **Table 2** to make the comparison more comprehensive in **Exp-1 in Section 6.2**.

(3) We have further discussed the coverage of the evaluation datasets, with a specific focus on the percentage of MySQL/PostgreSQL knobs covered in the documentation, manuals, and test datasets in **Table 1 in Section 6.1**.

**M8 [More Clarifications on Experimental Setup]:** We have provided more clarifications on the experimental setup in **Section 6.1**, with a particular focus on explaining that the training labels can be automatically annotated.

**M9 [Additional Related Work]:** We have revised the Related Work in **Section 2.2** to include a detailed discussion of traditional IR techniques to analyze the unique challenge of IR approaches.

**M10 [Clarifications on Design Choices]:** We further clarify our design choices, particularly focusing on the following aspects.

(1) There could be multiple ways, such as different tuning knobs, to fix the same issue. ANDROMEDA uses an LLM to reason about the various tuning knobs and select the most appropriate set. Specifically, ANDROMEDA generates a prompt that includes all the retrieved data and then calls the LLM to recommend the configurations. It is clarified in **Section 3 Step 3: Configuration Reasoning**.

(2) We justify two main reasons of using LLMs to predict knob values in **Section 3 Step 3: Configuration Reasoning**.

(3) We agree that we need more techniques to optimize the configuration value reasoning using LLM such as dealing with numeric value more smartly. We have explained it in our future work (see **Section 7**).

**M11 [Discussion on Generalizability]:** We have added a discussion in **Section 3** to justify that ANDROMEDA can be easily generalized to other DBMSes.

**M12 [Case Studies]:** We have added *good cases* to demonstrate the user questions that ANDROMEDA successfully addressed, and the *failure cases* where ANDROMEDA does not perform well, followed by an analysis of limitations of ANDROMEDA and how ANDROMEDA can be further improved.

Due to space limit, we have to put the case studies in **Appendix A** in our technical report [10].

## RESPONSES TO REVIEWER #1

**R1O1:** Experimental eval should include a few other baselines that use LLMs for solving the same problem...

**RESPONSE:** Thanks for the suggestion. To the best of our knowledge, we are the first to develop an LLM-based DBMS diagnosis system designed to accept NL questions on DBMS configuration issues and recommend specific configurations to address the issues. In response to your concern, we have added more baselines by substituting GPT with other LLMs, e.g., Qwen and Llama-3. Please refer to **M1** in our revision plan for details.

**R1O2:** There seems to be a lot of human effort required in collecting ground labels for training and evaluation (including confirmation by experts). Human in the loop in ML requires non-trivial amount of time and dollar cost. More discussion is needed.

**RESPONSE:** We apologize for the confusion. ANDROMEDA does not require significant human effort to collect ground truth labels for training and evaluation, thanks to its automated labeling process. We have revised the experimental setup (**Section 6.1**) for greater clarity. Please refer to **M8** for further details.

**R1O3:** Related to the above point, the goal of ANDROMEDA is to reduce the time it takes to manually arrive at a solution by expert DBAs. How could the paper quantify this claim?

**RESPONSE:** We added new experiments to evaluate **end-to-end time** of DBAs to solve issues with / without ANDROMEDA, to qualify how ANDROMEDA reduces the time. Please see **M3** for details.

**R1O4:** the overall F-1 scores from Table 2 are still not very high for ANDROMEDA. There seems to be a lot of room for improvement.

**RESPONSE:** First, we would like to explain why the overall F-1 scores for ANDROMEDA, as shown in Table 2, are not particularly high. As discussed in **Section 6.1**, the evaluation in the NL setting requires that ANDROMEDA’s output knobs strictly match the automatically labeled answers, while disregarding other potentially valid knobs. This constraint can lead to an underestimation of the F1 score, which is why we proposed the Runnable setting. In this setting, DBMS experts assess whether the recommended configurations resolve the issues. As shown in **Table 2**, ANDROMEDA achieves a SuccessRate of over 0.7.

We also acknowledge that “there seems to be a lot of room for improvement”. We encourage a call to action to significantly advance this core database problem. To support this effort, we have conducted a thorough analysis of the less favorable cases produced by ANDROMEDA. Please refer to **M12** for further details.

**R1O5:** What was the actual prompt submitted to the LLM? It wasn’t presented in the main body of the paper.

**RESPONSE:** Please refer to **Figure 2(b)** as an illustrative prompt submitted to the LLM. To make it clearer, we have added more real prompts in experiments in our case studies (see **M12**). Moreover, we have already included all the prompts in our experiments in the repo: <https://anonymous.4open.science/r/Andromeda-D150/>

**R1O6:** it would also be particularly insightful to discuss some loss cases for ANDROMEDA for which it does not do well and how ANDROMEDA could be improved to address those losses?

**RESPONSE:** We have included a detailed analysis of the bad cases to highlight the limitations of ANDROMEDA and discuss potential improvements. Please refer to **M12** for more details.

**R1O7:** How generalizable is ANDROMEDA to other DBMSes? Would it require training data generation, model training, and evaluation for each DBMS?

**RESPONSE:** ANDROMEDA can be generalized to other DBMSes. Please see **M11** for more details.

**R1O8:** The paper talks about the time efficiency of ANDROMEDA but a discussion on the cost efficiency is missing.

**RESPONSE:** As suggested, we have reported the cost efficiency, which is the average cost of calling LLMs for each NL debugging question  $q$ . Please see **M6** for more details.

### Minor Remarks:

**R1C1:** It appears all telemetry data is always analyzed in real-time? There seems to be a lot of room for optimization here.

**RESPONSE:** Yes, the telemetry data is analyzed in real-time. We have conducted a latency analysis, as reported in **Table 8**. The result shows that the time used for telemetry analysis is 2.85 seconds, which we believe is acceptable for users. However, we agree that there is room for optimization, such as filtering out unnecessary telemetry data. We will address this in future work.

**R1C2:** How does ANDROMEDA (and related approaches) deal with cases where a small variation of a question compared to a historical question leads to a very different solution?

**RESPONSE:** ANDROMEDA deals with these cases by matching domain-specific contexts from other sources, including troubleshooting manuals and DBMS telemetry data. This is one of the key reasons we use multiple sources in our RAG approach. Specifically, when relying solely on historical questions is not sufficient, prioritizing relevant manuals or telemetry data over potentially confusing historical questions can provide more accurate solutions.

**R1C3:** there could be multiple ways to fix the same issue? It’s unclear how ANDROMEDA and related approaches are reasoning about and ranking between different tuning knobs?

**RESPONSE:** Yes, there could be multiple ways, such as different tuning knobs, to fix the same issue, as illustrated in **Figure 2(b)**. ANDROMEDA uses an LLM to reason about the various tuning knobs and selects the most appropriate set. Please see **M10** for details.

**R1C4:** why manual-manual samples are prioritized over others?

**RESPONSE:** We conducted an experiment to evaluate different prioritization strategies, and the results show that prioritizing manual-manual training samples yields the best performance. The primary reason is that aligning manual-manual samples is “easier” than aligning question-question or manual-question samples due to the formality and consistency of manuals. By training our Alignment

component on these easier cases first, we ensure it is well-prepared before gradually improving it on more challenging cases.

**R1C5:** *what kind of vector DB and what kind of and size of indexes are needed to meet the runtime SLA of ANDROMEDA?*

**RESPONSE:** We use Meta Faiss as our vector database to index the embeddings of historical questions and manuals. In the current implementation of ANDROMEDA, we have indexed 5,138 vectors for MySQL and 1,158 vectors for PostgreSQL, with each vector having a dimension of 768. We have provided further clarification on this further in **Section 6.1**.

**R1C6:** *Analyzing a large amount of telemetry data and zooming in on the sections that might be correlated with the issue being debugged does not seem to be a new challenge and there may already be existing approaches.*

**RESPONSE:** There are two key challenges that distinguish our work. First, detecting anomalous telemetry data from vast amounts of time-series information is particularly difficult due to the influence of seasonality and trends, which can obscure what is truly anomalous. Second, existing approaches do not address the semantic correlation between natural language queries and telemetry data, which is an essential aspect of our solution.

## RESPONSES TO REVIEWER #2

**R2O1:** *The experimental setup and findings lack sufficient details ... (1) the paper should explore which types of queries LLMs perform well on, where RAG helps, and whether historical queries can be categorized by type and frequency. It should also examine how the system performs for queries with less frequent questions. (2) Including results on precision and recall in Table 2 would also help.*

**RESPONSE:** We have revised the paper by providing more details about experimental setup and findings.

(1) We have already categorized queries by considering the frequency and types of their ground-truth labels (i.e., the tuning knobs) in **Section 6.3** and **Section 6.4**. Specifically, as shown in **Figures 8 and 9**, we classify queries into high-, medium-, and low-frequency categories. We also explore the questions raised by the reviewer. For instance, in most cases, LLMs perform better on high-frequency queries compared with medium- and low-frequency ones, while RAG consistently improves performance across all frequency levels. We examine ANDROMEDA’s performance on less frequent queries. In **Figure 10**, we further categorize queries by the type of their corresponding knobs (i.e., numerical, categorical, and boolean). Our findings show that ANDROMEDA achieves better performance on numerical and boolean queries compared to categorical ones.

(2) As suggested, we have included precision and recall results in **Table 2** to make the comparison more comprehensive.

We have revised the paper to include detailed explanations of these experimental findings, as indicated in **M7**.

**R2O2:** *Traditional IR techniques have long addressed issues of query-document matching, but the paper does not discuss these prior works in detail. The paper should highlight the unique challenges in DBMS configuration debugging that necessitate novel approaches.*

**RESPONSE:** As suggested, we have revised the Related Work section to include a detailed discussion of traditional IR techniques and highlight the unique challenges. Please see **M9** for details.

**R2O3:** *(1) Inaccuracies in these sources could lead to incorrect or suboptimal debugging suggestions. (2) Achieving high accuracy in document retrieval among such diverse data types is inherently challenging and can affect overall performance. The paper should discuss (if possible empirically) the sensitivity of the approach to these factors.*

**RESPONSE:** Thanks for pinpointing these crucial factors.

(1) We agree that inaccuracies in sources of historical questions, troubleshooting manuals and telemetries can significantly impact overall performance. As suggested, we have added new experiments to empirically investigate these factors. Please see **M4** for details.

(2) We have already explored the impact of document retrieval on overall performance in **Section 6.3**. Our results show that the effectiveness of document retrieval significantly affects the overall performance. For example, our proposed retrieval method, which combines different types of documents, not only achieves better retrieval performance (i.e., higher recall) but also enhances overall results compared to using a single type of document.

**R2O4:** *(1) The paper should provide more details on the scalability of the approach (both in number and types) for detecting relevant telemetries. (2) it should address whether telemetries are “always” necessary, what queries benefit more telemetries, and what happens when there is ambiguity.*

**RESPONSE:** (1) As suggested, we have added a new experiment to examine the scalability of our approach for detecting relevant telemetry data in **Section 6.4**. Specifically, we have varied the number of input metrics and differentiate between metric types, and report the time required for telemetry data analysis. Please refer to **M5**.

(2) We have already analyzed the impact of telemetry data in **Section 6.4**. Our experimental results suggest that telemetry data may not “always” be necessary, as shown in **Figure 10**. Specifically, compared to ANDROMEDA-docs-only (i.e., without telemetry data), ANDROMEDA does not always demonstrate superior performance. This is because some anomalies are not caused by anomalous resource usage, such as the error “column ‘DateReceived’ cannot be null” shown in **Figure 1**. Additionally, we found that queries involving tuning numerical knobs tend to benefit more from telemetry data. We have revised **Section 6.4** to provide a more detailed analysis.

**R2O5:** *The paper should explain why these results are statistically significant (e.g., how does it significantly reduce the operational burden) compared to the option of having a DBA search discussion or community forums. Furthermore, discuss how the provided results by ANDROMEDA can be trusted w/o significant troubleshooting.*

**RESPONSE:** We have conducted a new experiment to evaluate whether ANDROMEDA offers a statistically significant advantage over DBAs searching through discussion or community forums in **Section 6.5**. Please see **M3** for details.

We have also added a discussion in **Section 3** on how ANDROMEDA can be trusted. The key idea is that ANDROMEDA not

only provides configuration recommendations but also offers *explanations*, grounded in historical questions, troubleshooting manuals, and telemetry data.

## RESPONSES TO REVIEWER #3

**R3O1:** *the intuition why it is a good candidate for recommending knob values is not as clear. LLMs are not particularly good at numerics. I wonder why the problem setup chooses to directly output the knob value, instead of suggesting directional adjustments (e.g., increase/decrease, enable/disable).*

**RESPONSE:** We have included a detailed discussion in **Section 3** to justify the intuition why LLM is good for recommending knob values. Please see **M10** for details. The reason why we directly output the knob value is that it is more convenient for the user to set an exact value than suggesting directional adjustments. For example, if ANDROMEDA recommends the user to increase a knob, user may still not know which value is suitable.

**R3O2:** *it seems like statistics on the sql table, statistics of query workload and machine resources such as memory size could also play a role in debugging.*

**RESPONSE:** We have conducted new experiments to investigate the impact of providing additional contextual information to prompt the LLM, specifically focusing on *DB statistics from the tables*. We utilize the table names, the number of tuples, column names, and column data types as the DB statistics. We added this information to origin ANDROMEDA and use GPT 3.5 to evaluate it. The accuracy of dataset “MySQL Run” in Runnalbe setting is 0.61, which is lower than ANDROMEDA (0.79). These results suggest that the current three sources, *i.e.*, historical questions, DBMS manuals, and telemetry, seem sufficient, and adding more information for the LLM may adversely affect its predictions.

**R3O3:** *This PLM baseline could be made stronger if you augment the training data using examples generated from Section 4.2.*

**RESPONSE:** We have added this baseline in the revised version, as described in **M1**.

**R3O4:** *It would also help to get some idea on the coverage of the evaluation datasets. What percent of MySQL/Postgres knobs were covered in the documentation, manual and test datasets?*

**RESPONSE:** We have discussed the coverage of the evaluation datasets in **Section 6.1**, as described in **M7**.

### Minor remarks:

**R3C1:** *Numeric knobs have higher success rate than categorical ones (Figure 9). Does it mean that the specific values of numeric knobs do not matter that much - multiple values would work.*

**RESPONSE:** Yes, multiple values would work. We have explained this in the case study (see **M12**).

**R3C2:** *It wasn't clear until the evaluation that there are essentially two version of ANDROMEDA. One is only concerned with identifying the correct subset of the knobs and does not predict knob values (the NL setting). The other takes telemetry data and also predicts knob values (the Runnable setting).*

**RESPONSE:** Thanks. We have clarified it when introducing the framework in **Section 3**.

**R3C3:** *In section 6.4 “(2) ANDROMEDA-telemetry-only: ANDROMEDA without telemetry data analysis, (3) ANDROMEDA- docs-only: ANDROMEDA without document retrieval”. Is this a typo?*

**RESPONSE:** We have fixed it. Thanks!

## RESPONSES TO META REVIEWER

**C1:** *(1) The reviewers felt the experiments warranted more specific motivations – characterizing the types of queries for which these techniques are (and are not) effective, and assessing the sensitivity of these methods to question frequency in the training data. (2) Providing experiments to show more specifically where the proposed approach does or does not excel would provide a more nuanced recommendation for practitioners – some results are modest, and limitations are not discussed.*

**RESPONSE:** (1) Please refer to our responses to **R2O1**. (2) See our responses to **R1O4** and **R2O5**.

**C2:** *(1) The reviewers also noted that the experiments are limited to a small dataset, which puts competitive methods at a disadvantage. If this limited dataset is realistic, some additional motivation is warranted. Otherwise, increasing its size to provide a more realistic setting for competitive methods would make the argument more convincing. (2) A related issue is the availability of training labels – the cost in time and training warrants additional discussion. (3) In addition, providing more baselines involving LLMs (e.g., varying prompting, fine-tuning, or other setups). Your shepherd can assist with considering additional experiments and discussion.*

**RESPONSE:** (1) We have increased the dataset size. Please refer to **M2**. (2) Please refer to our responses to **R1O2**. (3) We will add these baselines in the experiments. Please refer to **M1**.

**C3:** *(1) There was also consensus that the paper could be significantly improved by highlighting the unique challenges in this context of configuration debugging relative to more general query-document matching in the IR literature, and (2) the generalizability of the system to other DBMS. (3) Also, some design decisions warrant more discussion: The output is numeric settings, but LLMs can struggle with reasoning about numeric values. Therefore some discussion about alternatives, including qualitative adjustments (direction, categorical magnitudes) could produce different results.*

**RESPONSE:** (1) Please refer to our responses to **R2O2**. (2) Please refer to our responses to **R1O7** or **M11**. (3) Please refer to our responses to **R3O1** or **M10**.

**C4:** *... the results could be improved by providing additional contextual information to the LLM: machine resources, DB statistics from the tables, statistics of the query workload when available.*

**RESPONSE:** Please refer to our responses to **R3O2**

**Required Changes:** Overall: R1O1, R1O2, R1O4, R1O7, R2O1, R2O2, R2O4, R3O1, R3O3; Optional changes include R1O3, R1O5, R1O6, R1O8, R2O3, R3O2

**RESPONSE:** We have addressed all the concerns.

# Automatic Database Configuration Debugging using Retrieval-Augmented Language Models

Anonymous Author(s)

## ABSTRACT

Database management system (DBMS) configuration debugging, e.g., diagnosing poorly configured DBMS knobs and generating troubleshooting recommendations, is crucial in optimizing DBMS performance. However, the configuration debugging process is tedious and, sometimes challenging, even for seasoned database administrators (DBAs) with sufficient experience in DBMS configurations and good understandings of the DBMS internals (e.g., MySQL or Oracle). To address this difficulty, we propose ANDROMEDA, a framework that utilizes large language models (LLMs) to enable automatic DBMS configuration debugging. ANDROMEDA serves as a natural surrogate of DBAs to answer a wide range of natural language (NL) questions on DBMS configuration issues, and to generate diagnostic suggestions to fix these issues. Nevertheless, directly prompting LLMs with these professional questions may result in overly generic and often unsatisfying answers. To this end, we propose a retrieval-augmented generation (RAG) strategy that effectively provides matched *domain-specific contexts* for the question from multiple sources. They come from related historical questions, troubleshooting manuals and DBMS telemetries, which significantly improve the performance of configuration debugging. To support the RAG strategy, we develop a document retrieval mechanism addressing heterogeneous documents and design an effective method for telemetry analysis. Extensive experiments on real-world DBMS configuration debugging datasets show that ANDROMEDA significantly outperforms existing solutions.

## 1 INTRODUCTION

Database management system (DBMS) performance diagnosis is a recurring theme, which constantly baffles cloud database customers. Poorly configured DBMSs may suffer from unexpected performance pains. There has been a growing focus on *automatic DBMS configuration debugging*. Traditional approaches [11, 15, 17, 18, 26, 37, 43] focus on building complex models from performance metrics (such as runtime, resource usages, etc.) to analyze the root cause of a performance issue, and then identifying potential configurations (e.g., problematic knobs) to tune.

**Human-based Debugging.** In real-world scenarios, users often pose natural language (NL) questions regarding configuration issues. Database administrators (DBAs) then perform configuration debugging, as illustrated in Figure 1. Specifically, database users can

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGMOD '20, June 14–19, 2020, Portland, OR

© 2024 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

query in NL on *diverse* configuration debugging requirements, including diagnosing performance issues (e.g., “*Executing an INSERT statement is very slow*”), execution errors (e.g., “*Importing data into RDS encounters issues*”) and sub-optimal configurations (e.g., “*The number of decimals does not meet my application requirements*”), or even requesting for troubleshooting recommendations (e.g., “*How to fix the issue*” and “*How to specify proper configurations*”). Given such NL debugging questions, a DBA typically analyzes information from multiple sources, such as related historical questions, troubleshooting manuals and DBMS telemetry data (*i.e.*, performance metrics, like CPU utilization and query latency). Based on this analysis, the DBA provides recommendations on configuration settings to resolve the issues (as shown in the right part of Figure 1).

Clearly, the above process is challenging and time-consuming, as it requires the DBA to be highly experienced in configuration debugging and have a comprehensive understanding of the DBMS’s internal working mechanisms. This raises a crucial question: *can we replace the DBA with an LLM-based agent to enable automatic DBMS configuration debugging?*

**Our Proposal: LLM-based Debugging.** To address the aforementioned problem, we propose ANDROMEDA, a framework that harnesses large language models (LLMs) to support automatic DBMS configuration debugging. ANDROMEDA serves as a natural surrogate of DBAs to answer a wide range of NL questions regarding DBMS configuration issues, as shown in Figure 1. A straightforward approach to support ANDROMEDA is to directly prompt an existing LLM (e.g., GPT-4) with NL debugging questions and return the answer of the LLM. However, this approach often yields *overly-generic yet useless* recommendations. This limitation stems from the fact that LLMs are trained on general-purpose datasets and lack the specific domain knowledge related to DBMS configuration. To solve this problem, ANDROMEDA employs a retrieval-augmented generation (RAG) strategy that enriches the NL debugging questions with domain-specific context drawn from multiple sources, including historical questions, troubleshooting manuals and DBMS telemetry data. We further illustrate the strategy in the following example.

**EXAMPLE 1.** Let’s consider an NL question regarding the slow execution of an *INSERT* statement, as shown in Figure 2. We can see that the results of directly prompting an existing LLM (e.g., GPT-4) are vague and overly generic, although they are “technically correct” (Figure 2 (a)), which are unhelpful for the users to solve the issue. In contrast, as shown in Figure 2 (b), equipped with our RAG strategy, ANDROMEDA finds contextual information from the following sources:

(a) ANDROMEDA finds similar historical debugging questions that another user also encounters slow execution during data insertion, and provides a corresponding configuration solution, i.e., *disabling foreign\_key\_checks and unique\_checks*, for reference.

(b) ANDROMEDA identifies a document from the MySQL Manual 10.5.5: Bulk Data Loading for InnoDB Tables, which describes that *disabling autocommit when importing large amounts of data can achieve speed-up and explains the reasons*.

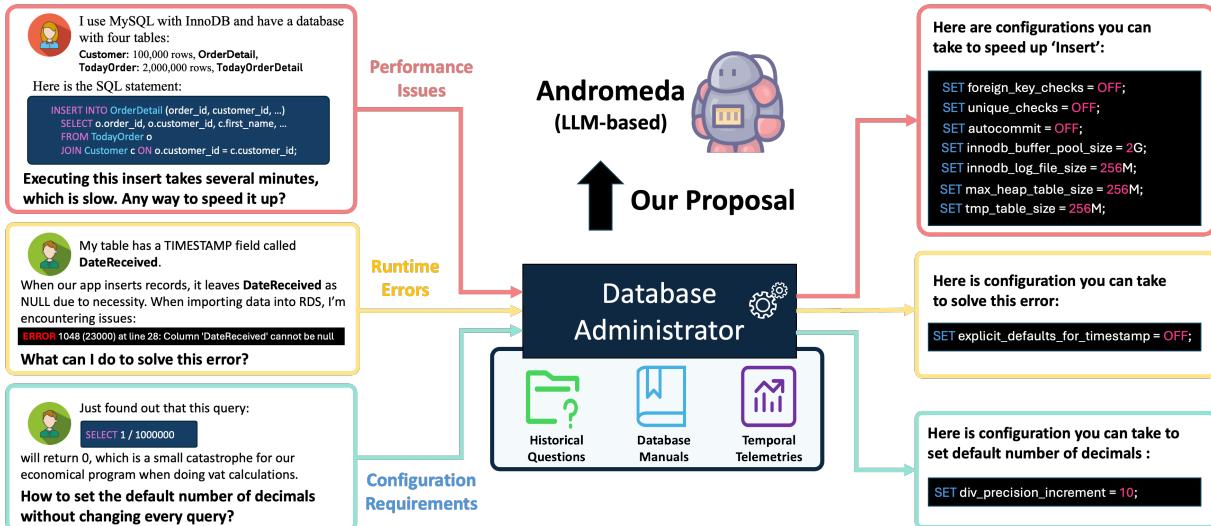


Figure 1: Overview of automatic DBMS configuration debugging, where users directly pose NL debugging questions regarding configuration issues, and there exists a “co-pilot” to diagnose the issues and generate recommendations to fix the issues.

(c) ANDROMEDA considers the telemetry log and detects three troublesome performance metrics, e.g., `innodb_log_write_requests`, and suggests related knobs to tune.

With the above contextual information, ANDROMEDA leverages LLMs to diagnose improper DBMS configurations and generates accurate troubleshooting recommendations.

### Key Technical Challenges.

(C1) *Retrieving heterogeneous documents*. We retrieve documents from multiple sources, such as historical questions and DBMS manuals, where the semantics of documents from different sources may be heterogeneous. For example, as shown in Figure 2, historical questions typically describe “*what is wrong*” (e.g., “*It takes over 15 minutes ...*”), while DBMS manuals focus on “*How to fix it*” (e.g., “*turn off autocommit mode*”). This heterogeneity poses significant challenges on retrieval mechanisms.

(C2) *Detecting troublesome telemetries*. We also extract relevant telemetry data, i.e., time series data of a performance metric (such as CPU utilization) with timestamps, which is related to the user’s NL question. However, there are many types of telemetries in a DBMS, and only a small number of telemetries are relevant to the question. Moreover, even given a relevant telemetry, only a small proportion of timestamps (e.g., the duration of high CPU utilization) are useful. Thus, it is non-trivial to detect such troublesome telemetries that are relevant to a user’s NL question.

To address challenge (C1), we propose an effective document retrieval mechanism that addresses the semantic heterogeneity among different sources. We propose a contrastive learning approach that aligns documents from different sources into a unified representation space, and enables accurate document retrieval on the space. We also design a novel data synthesis method to tackle the scarcity of training data. To address challenge (C2), we introduce a telemetry detection method to identify troublesome telemetries that are relevant to a user’s NL question. The method first utilizes a seasonal-trend decomposition mechanism to find troublesome telemetries, and then judiciously selects relevant telemetries.

**Contributions.** Our contributions are summarized as follows.

(1) **LLM-based automatic DBMS configuration debugging framework.** We formally define the problem of automatic DBMS configuration debugging (Section 2) and introduce an overview of the ANDROMEDA framework (Section 3).

(2) **Effective RAG techniques.** We introduce a document retrieval mechanism addressing heterogeneous documents (Section 4) and design an effective method for telemetry analysis (Section 5).

(3) **Extensive experiments.** Experiments using real-world DBMS configuration debugging datasets show that ANDROMEDA significantly outperforms existing alternatives. The code is available at the repo: <https://anonymous.4open.science/r/Andromeda-D150/>.

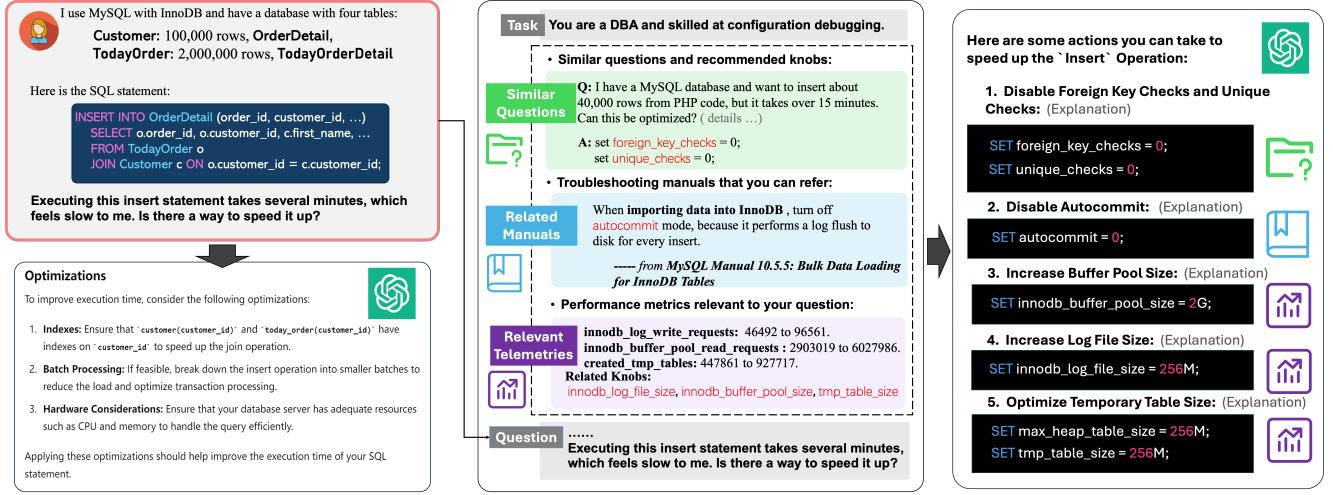
## 2 PRELIMINARIES

### 2.1 Problem Formalization

**NL Debugging Question.** Let  $q$  be a user posed NL debugging question that reports unexpected configuration issues, which require diagnostic procedures over a DBMS  $D$ . Note that the question  $q$  may express a variety of configuration debugging purposes, including but not limited to: (1) *Diagnosing performance issues*, e.g., much longer execution time that a database query or operation typically takes, and significant resource usage that the DBMS is using all available resources such as CPU, memory, or disk I/O. (2) *Fixing runtime errors*, e.g., “*Error: Could not refresh instance*” in MySQL. (3) *Customizing user-specific configurations*, e.g., changing default decimal number and ignoring DBMS runtime warnings.

Moreover, a question  $q$  may also include the description of user’s database (e.g., schema and statistics), query workload and environments. For example, Figure 1 provides three NL debugging questions, which are, respectively, of the types of performance issues, runtime errors, and user-specific configurations.

**Configurable Knobs.** Although there exist different ways to configure the DBMS [15, 37], this paper focuses on diagnosing and



**Figure 2: An example of our RAG strategy in ANDROMEDA. (a) A straightforward strategy that directly prompts an LLM with the NL question results in overly-generic yet useless answers. (b) Our RAG strategy provides domain-specific context of an NL debugging question from multiple sources that improve the inference capabilities of the LLM on configuration debugging.**

tuning the DBMS *configurable knobs* (or *knobs* for short). Specifically, modern DBMSs usually have hundreds of configurable knobs that control their runtime behavior, e.g., 1012 knobs (such as `wait_timeout`, `unique_checks`) in MySQL 8.0 and more than 300 knobs (such as `trace_locks`) in PostgreSQL 12.0, which play crucial roles in DBMS tuning [11, 18].

Formally, we denote the set of configurable knobs in a DBMS as  $\mathbb{K}$ , where each  $k \in \mathbb{K}$  is a knob with a possible value domain  $\mathbb{D}_k$ . Take the knob `autocommit` with value domain  $\{0, 1\}$  in MySQL as an example: If the `autocommit` mode is enabled (i.e., `autocommit=1`), each SQL statement forms a single transaction on its own. If `autocommit` mode is disabled within a session (i.e., `autocommit = 0`), the session always has a transaction open.

**DBMS Configuration Debugging.** Given an NL debugging question  $q$  over a DBMS  $D$ , the problem of DBMS configuration debugging aims to (1) diagnose a subset  $K \subseteq \mathbb{K}$  of knobs relevant to the corresponding configuration issues in  $q$ , and (2) recommend a proper value  $v \in \mathbb{D}_k$  for tuning each relevant knob  $k \in K$ .

**EXAMPLE 2.** Figure 2 illustrates a running example of DBMS configuration debugging. A database user poses an NL question  $q$  to report a performance issue about a much longer execution time for an `INSERT` SQL statement (the left part of the figure). Given question  $q$ , the problem aims to diagnose a subset  $K \subseteq \mathbb{K}$  of knobs relevant to the corresponding configuration issues in  $q$ , and (2) recommend a proper value  $v \in \mathbb{D}_k$  for tuning each relevant knob  $k \in K$ .

## 2.2 Related Work

**DBMS Diagnosis and Configuration Tuning.** Traditional DBMS diagnosis studies aim to identify and resolve performance and stability issues to ensure efficient, stable, and reliable operation of the underlying DBMS. These studies mainly employ techniques

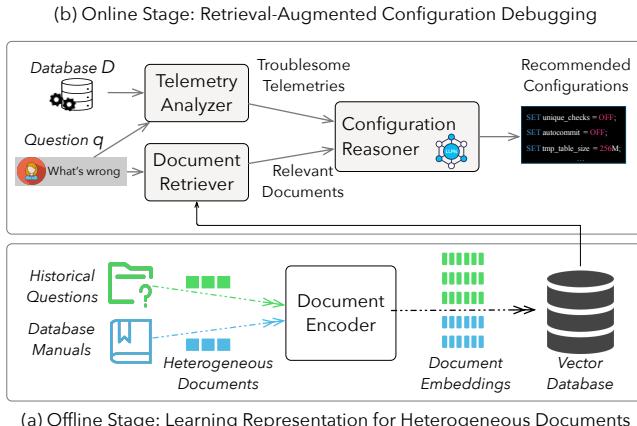
based on empirical rules, graph algorithms, and machine learning models [16, 27, 28, 38].

Recently, modern DBMSs are equipped with automatic configuration tuning tools [11, 17–19, 26, 40, 42, 43]. Some representative studies include the Lasso algorithm in OtterTune [11], Sensitivity Analysis (SA) in Tuneful [18] for automatic knob selection, Bayesian optimization (e.g., iTuned [17] and ResTune [43]), reinforcement learning (e.g., CDBTune [40] and Qtune [26]) for configuration optimization, and RGPE [19] in ResTune [43] for knowledge transformation. Recently, GPTuner [25] employs LLMs for configuration tuning based on DBMS manuals and Bayesian optimization.

However, these studies are not designed for taking NL questions from database users as input and recommend configurations corresponding to the issues, which is a key feature of our ANDROMEDA framework. Moreover, they are constructed relying on pre-defined optimization objectives (such as runtime), making them difficult to address runtime errors and user-specific configurations.

**LLM-based DBMS Diagnosis.** Panda [34] takes the user’s NL question as input, and introduces a context-grounding mechanism to LLMs by retrieving relevant documents and telemetries, so as to generate troubleshooting recommendations for the questions. Note that there are two differences between Panda and ANDROMEDA. First, ANDROMEDA explicitly outputs specific and accurate DBMS configuration settings with knobs and their values, whereas Panda outputs text-based debugging recommendations. Second, compared with Panda, ANDROMEDA develops effective RAG techniques.

D-Bot [46] is also a recent LLM-based DBMS diagnosis tool that acquires knowledge from diagnosis documents, and generates a diagnosis report (i.e., identifying the root causes and solutions) when a DBMS triggers an alert. There are two differences between D-Bot and ANDROMEDA. The first is that the input of ANDROMEDA is NL questions while that of D-Bot is triggered by alerts based on rules during DBMS execution. The second is that ANDROMEDA outputs specific DBMS configuration settings with knob types and



**Figure 3: An overview of ANDROMEDA. (a) Offline:** ANDROMEDA learns representation for heterogeneous documents and stores the document embeddings in a vector database. **(b) Online:** ANDROMEDA utilizes an RAG-based configuration debugging strategy to recommend configurations for an NL question  $q$  and a database  $D$ .

values, whereas D-Bot outputs analysis report texts that may not explicitly recommend specific knob configurations.

**Retrieval-Augmented Generation (RAG).** RAG has been proved to excel in many tasks, including open-ended question answering [24, 35] and fact verification [12, 22, 29, 39]. Automatic DBMS configuration debugging is knowledge-extensive and requires experts to guarantee the accurate diagnosis of LLMs. In this paper, we introduce a framework called ANDROMEDA based on RAG for configuration debugging, and demonstrate its effectiveness.

**Information Retrieval.** Several information retrieval (IR) techniques focus on systems that respond to questions by utilizing a community or a dataset of previously answered questions. Conventional methods employ neural networks [45], knowledge graphs [13, 36], and user intent analysis [33, 41, 44]. With the development of LLMs, more approaches consider incorporating LLMs to tackle the question answering problems [22, 29]. [22] introduces Question-Answer Cross Attention Networks (QAN) to leverage pre-trained models for answer selection, and improves the performance using knowledge augmentation from LLMs. Compared to these methods, the unique challenge of ANDROMEDA lies in retrieving documents from multiple sources, such as historical questions and DBMS manuals, where the semantics of documents from different sources may be heterogeneous. For instance, historical questions typically describe “what is wrong”, while DBMS manuals focus on “how to fix it”. Moreover, it is hard for existing IR techniques to analyze telemetry data in the DBMS debugging scenarios.

### 3 THE ANDROMEDA FRAMEWORK

As shown in Figure 3, ANDROMEDA employs a retrieval-augmented generation (RAG) strategy that consists of offline and online stages.

**Offline Stage.** In the offline stage, ANDROMEDA learns a unified representation *i.e.*, a vector-based embedding for heterogeneous documents from multiple sources. ANDROMEDA currently considers the following two types of document sources.

- *Historical Questions*  $Q$  contains historical configuration debugging questions that users have encountered, where each instance  $\langle q_i^H, K_i \rangle$  contains a historical question  $q_i^H$  and the recommended configurations  $K_i$  corresponding to  $q_i^H$ . Figure 2 shows an historical question regarding “*slow execution while inserting large amounts of data*” with its recommended configurations,  $\text{foreign\_key\_checks} = 0$  (disabling foreign key checks) and  $\text{unique\_checks} = 0$  (disabling unique checks).
- *Troubleshooting Manuals*  $M$  consists of documents for DBMS manuals, *e.g.*, MySQL Manual [4] and PostgreSQL Manual [5], which describe how to troubleshoot configuration issues. Specifically, we first chunk each document in the manuals into sentences  $\{m_i\}$ , and then use regular-expression rules to match each sentence  $m_i$  with the recommended configurations  $K_i$ . Consider the example in Figure 2 again. A document from *MySQL Manual 10.5.5: Bulk Data Loading for InnoDB Tables* contains a sentence “*When importing data into InnoDB, turn off autocommit mode...*” with its corresponding configuration  $\text{autocommit} = 0$ . In such a way, we form a set of troubleshooting manuals  $M = \{\langle m_i, K_i \rangle\}$ .

Given document sets  $Q$  and  $M$ , ANDROMEDA utilizes a DOCUMENT ENCODER to map each document (either historical question  $q^H$  or manual instance  $m$ ) into a vector-based embedding representation. Then, it stores learned document embeddings in a vector database, which will be further used for online document retrieval.

The key challenge here is that the semantics of documents from different sources may be heterogeneous. Consider the examples of historical questions and troubleshooting manuals in Figure 2 again: historical questions typically describe “*what is wrong*”, while manuals focus on “*How to fix it*”. Thus, it is challenging to develop a unified document encoder that understands the semantics of different types of documents for effective document retrieval. To address this challenge, we propose an effective document encoding mechanism that addresses the semantic heterogeneity among different sources. We propose a contrastive learning-based approach that aligns documents from different sources into a unified embedding space, and enables accurate document retrieval on the space. We also design a data synthesis strategy to tackle the difficulty of limited training data. We will provide more details in Section 4.

**Online Stage.** In the online stage, ANDROMEDA utilizes a RAG strategy to recommend configurations  $K$  for an NL question  $q$  over a DBMS  $D$ , which consists of the following three steps.

**Step 1: Document Retrieval.** ANDROMEDA encodes the question  $q$  into its embedding representation and employs a DOCUMENT RETRIEVER to retrieve top- $k$  historical questions or troubleshooting manuals that are most similar to  $q$  in the embedding space.

**Step 2: Telemetry Analysis.** Inspired by the way a DBA performs configuration debugging, ANDROMEDA also needs to know what is happening in the DBMS  $D$ . To this end, we propose to extract key features from the telemetry data, *i.e.*, time-series data of performance metrics (*e.g.*, CPU utilization). Formally, we define the telemetry data as a set  $S = \{s_i, \{t_j^i\}\}$  of time-series performance metrics, where each instance consists of a performance metric  $s_i$  and a collection of timestamps  $\{t_j^i\}$ . Note that not all performance metrics or all timestamps of a particular metric may be relevant

to the user's question  $q$ . Thus, ANDROMEDA utilizes a **TELEMETRY ANALYZER** to detect "troublesome" telemetries  $\{s_i\}$  with the particular timestamps when a performance issue occurs (e.g., the duration of full CPU usages). Moreover, the **TELEMETRY ANALYZER** also suggests the related knobs that need adjustment. Consider the example in Figure 2. ANDROMEDA considers telemetries of DBMS  $D$  and detects three troublesome performance metrics, e.g., `innodb_log_write_requests`, and suggests the related knobs that need adjustment, e.g., `innodb_log_file_size`.

**Step 3: Configuration Reasoning.** With the above useful and contextual information, ANDROMEDA employs a **CONFIGURATION REASONER** that leverages an LLM (e.g., GPT-4 or Llama 3) to diagnose improper DBMS configurations and generate accurate troubleshooting recommendations. Specifically, it generates a prompt that consists of three components: (1) task instruction, (2) retrieved data, including historical questions, troubleshooting manuals and relevant telemetries, and (3) user's NL question  $q$ . Figure 2 provides an example of the prompting mechanism (the central part of the figure). After that, ANDROMEDA prompts an LLM to generate the configuration recommendations (as shown in the right part of Figure 2). In particular, we utilize a two-phase prompting strategy that respectively (1) identify a subset  $K \subseteq \mathbb{K}$  of knobs relevant to the corresponding configuration issues in  $q$ , and (2) recommend a proper value  $v \in \mathbb{D}_k$  for tuning each relevant knob  $k \in K$ , which makes sense because (i) Using RAG, LLM can obtain knowledge about the knobs. For example, LLM can set `autocommit=0` with manual, and it can set `innodb_log_file_size = 256` if the similar question sets the same value. (ii) LLMs may learn some database related knowledge in pre-training or fine-tuning process, which is helpful for knob value decisions. Note that there could be multiple ways, such as different tuning knobs, to fix the same issue. ANDROMEDA uses the LLM to reason about the various tuning knobs and select the most appropriate set. Specifically, ANDROMEDA generates a prompt that includes all the retrieved data and then calls the LLM to recommend the configurations.

To use ANDROMEDA flexibly, users can selectively utilize the sources (for instance, if telemetry is unavailable, they can only consider documents). Moreover, users can either use only Step 1 to identify the correct subset of knobs or perform an end-to-end process to predict the knob values.

The key challenge in the online stage is how to detect troublesome telemetries. Specifically, there are many types of telemetries in a DBMS, and only a small number of telemetries are relevant to the question. Moreover, even given a relevant telemetry, only a small proportion of timestamps are useful. Thus, it is non-trivial to detect such troublesome telemetries that are relevant to a user's NL question. To address the challenge, we introduce a telemetry analysis method to identify troublesome telemetries that are relevant to a user's NL question. The method first utilizes a seasonal-trend decomposition mechanism to find troublesome telemetries, and then judiciously selects the telemetries relevant to a user's question. Please refer to Section 5 for more details.

**Discussion on Generalizability.** ANDROMEDA can be easily generalized to other DBMSes, as the algorithms for training data generation, model training, and evaluation are adaptable to new DBMSes. When applying ANDROMEDA to a new DBMS, all that is required

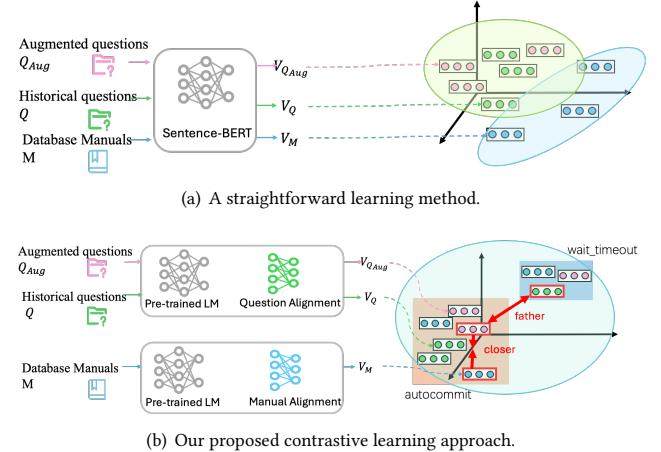


Figure 4: An overview of document representation learning.

is the specification of the knobs and the manuals of the DBMS. ANDROMEDA can then use these inputs to automatically generate augmented data and train a new model during the offline stage, while supporting evaluation on the DBMS in the online stage.

## 4 MULTI-SOURCE DOCUMENT RETRIEVAL

Given a user's NL question  $q$ , multi-source document retrieval aims to retrieve relevant historical questions from  $Q$  and related troubleshooting manuals from  $M$ . To achieve this goal, we introduce an encoder model  $f_{enc}$  that maps the question  $q$ , historical questions  $Q$  and DBMS manuals  $M$  into dense vectors, as shown in Figure 4. Then, we can efficiently retrieve top- $k$  relevant documents by using the normalized L2 distance of the question  $q$  and a document (i.e., a historical question  $q^H$  or a manual instance  $m$ ).

Intuitively, we develop techniques to address two main research challenges in designing encoder  $f_{enc}$ .

Firstly, as mentioned previously, manuals (usually including explanations of knobs and solutions) and questions (typically including configuration issues and user behaviors) are semantically heterogeneous. Thus, a straightforward encoding solution (e.g., using Sentence-Bert [30]) may result in distributional divergence among documents from different sources, as shown in Figure 4 (a). To address this challenge, we develop a contrastive learning approach that constructs *comparative relationships* for documents from different sources, so as to *align* the heterogeneous documents in a unified representation space, as shown in Figure 4 (b).

Secondly, constructing such an encoder  $f_{enc}$  that aligns documents from different sources requires a substantial amount of training data. However, in practice, there is often an insufficiency in both quantity and quality of the training data. To address the challenge, we propose a logic-based task decomposition method that uses LLMs to perform logical reasoning with the knowledge from manuals, thereby augmenting high-quality training data.

### 4.1 Document Representation Learning

We present our contrastive learning approach that constructs *comparative relationships* for documents from different sources.

**Model Structure.** We construct  $f_{enc}$  as a deep neural network. As shown in Figure 4, we design representation models separately for different document types, *i.e.*, questions and manuals. Specifically, for each document type, we firstly employ a pre-trained language model (LM) (*e.g.*, Sentence-Bert) to encode a document into a vector-based representation. Then, we add an *Alignment* component to map the original embeddings into a unified representation space. In particular, the Alignment component is implemented using a neural network with multiple fully-connected layers.

**Model Training.** We design a contrastive learning approach to train the above model, *i.e.*, optimizing parameters for both pre-trained LMs and our Alignment components. To this end, we need to prepare both *positive* and *negative* documents for each target question  $q^*$  in the training data, which are described as follows.

(1) *Positive document* is defined as the historical question  $q^H$  or manual instance  $m$  related to  $q^*$ , *i.e.*, the recommended knobs of  $q^H$  or  $m$  match that of  $q^*$ . To efficiently prepare positive documents, we first use Sentence-Bert to identify the similar documents from  $Q$  and  $M$  to  $q^*$ , and then filter out the documents with knobs not overlapping with the recommended knobs of  $q^*$ .

(2) *Negative document* is defined as the historical question  $q^H$  or manual instance  $m$  unrelated to  $q^*$ , *i.e.*, the recommended knobs of  $q^H$  or  $m$  are not overlapping with that of  $q^*$ . Similar to positive documents, we also use Sentence-Bert to identify the similar documents from  $Q$  and  $M$  to  $q^*$ , and then only consider the documents with knobs not intersecting with that of  $q^*$ .

In such a way, we can prepare three types of training data instances. (1) *question-manual*: given a target question, the positive documents are historical questions and negative documents are manuals. (2) *question-question*: given a target question, both positive and negative documents are historical questions. (3) *manual-manual*: given a target question, both positive and negative documents are manuals. For each type, we prepare training data with an equal number of positive and negative documents. Additionally, we prioritize training manuals-manuals to achieve overall alignment of the manuals, and then proceed for other types.

For model training, we freeze the pre-trained LM (*i.e.*, Sentence-Bert in our implementation) and optimize the alignment models through contrastive learning, with the InfoNCE loss function,

$$L_{InfoNCE} = -\log \frac{\exp(||V_{q^*}, V_{d^+}||^2/\tau)}{\sum_{d^- \in D^-} \exp(||V_{q^*}, V_{d^-}||^2/\tau)}, \quad (1)$$

where  $d^+ \in D^+$  is positive document, while  $D^-$  are negative documents. The basic idea of the above InfoNCE loss brings positive retrieval texts closer in the embedding space while pushing negative retrieval texts further apart, as shown in Figure 4 (b).

## 4.2 Training Data Augmentation

For better training  $f_{enc}$ , we need a sufficient amount of training data, *i.e.*, a large number of target queries ( $q^*$ ) with ground-truth knob configurations. However, it is time- and effort-consuming to collect a large amount of high-quality training data, which hinders the effectiveness of document representation learning. To solve this problem, we introduce a novel **data synthesis** mechanism

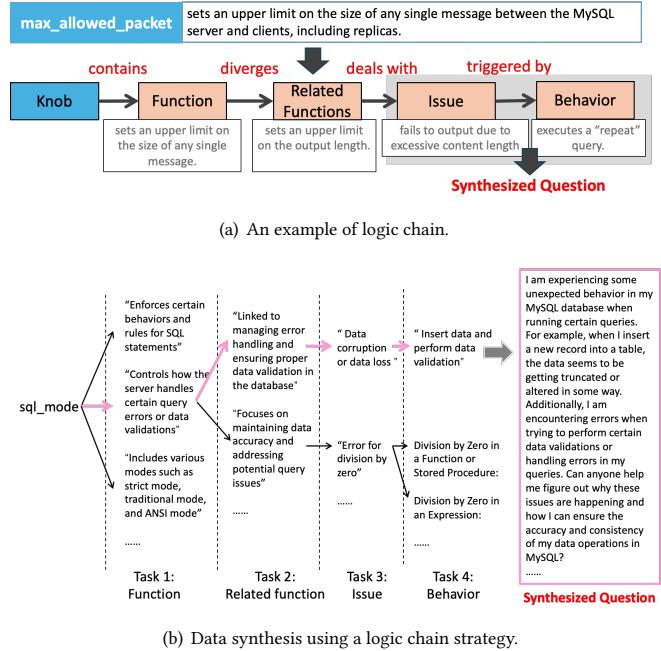


Figure 5: An overview of training data augmentation.

that utilizes an LLM to automatically generate NL questions with corresponding knob configurations from DBMS manuals.

Unfortunately, a straightforward data synthesis method, which directly prompts the LLM with manuals and asks it to generate possible user's questions based on the manuals, may not perform well. The main reason is that the generated questions are not *realistic*, *e.g.*, lacking contextual information, such as the user's behavior details. Moreover, the generated questions may also fall short of diversity. We illustrate this in the following example.

**EXAMPLE 3 (STRAIGHTFORWARD DATA SYNTHESIS).** Consider the manual of knob `max_allowed_packet` shown in Figure 5(a). Directly prompting an LLM (*e.g.*, GPT-4) returns the following question: “I'm experiencing issues with my MySQL replication setup. Occasionally, large BLOB or TEXT fields cause errors on the source server, which then leads to the replication process stopping unexpectedly. What could be causing this issue and how can I resolve it?”. Compared with the real questions shown in Figure 1, it is not difficult to find such questions are not realistic.

**Our Logic Chain Strategy.** To address the above limitation, we introduce a logic-chain based strategy for effective data synthesis. The basic idea is to generate realistic questions by simulating a real user's experience when encountering a DBMS configuration issue. Specifically, we consider the natural logical reasoning chain from issues to the solutions (*e.g.*, recommended knob configurations):

Triggering behaviors → Behaviors causing issues → Specific functions resolving issues → knobs with the corresponding functions.

Intuitively, if we traverse the logic chain reversely, we can inject domain knowledge about the logical relationships between knobs and issues into the LLM. Moreover, as the functions can be quite

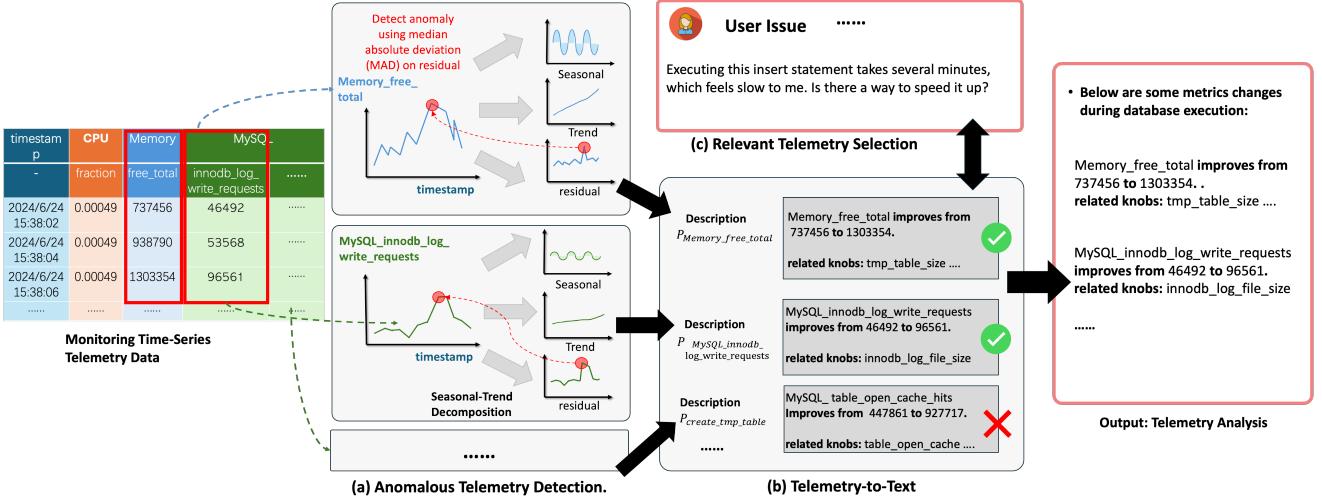


Figure 6: An overview of telemetry data analysis, which employs anomalous telemetry detection, telemetry-to-text, and relevant telemetry selection to report the “anomalous” status of database  $D$ .

diverse, given a knob, we can generate a wide range of possible questions, which is helpful to improve the diversity of training data.

**EXAMPLE 4 (OUR LOGIC CHAIN BASED SYNTHESIS STRATEGY.)** Consider the manual of `max_allowed_packet` in the previous example again. We can extract diverse knob functions from the manual, and then we can infer that function “set an upper limit on the output length” can resolve issue “fail to output due to excessive content length”, which is triggered by the behavior “executes a repeat command”. We provide such a logic chain to the LLM, and obtain a more realistic question “I want to connect to the remote SQL server to simulate traffic using the REPEAT command in MySQL. And I executed SQL: REPEAT(‘A’, 26214400). The connection is successful, but I encounter an error that returns NULL. How can I solve this for MySQL?”

Based on the above idea, we utilize an LLM with a logic-chain prompting to generate questions. Specifically, we decompose the data synthesis task into several sub-tasks. Each sub-task aims to generate details for each edge in the logic chain using the LLM, as shown in Figure 5(b). For instance, for “knob->function” in the logic chain, we can construct the sub-task with question “What are the functions of `max_allowed_packet`?”, which can be answered as `spec_func` by the LLM. Next, for “function->Anomaly”, the question can be “What anomaly can be solved with the `spec_func` of `max_allowed_packet`”. When the LLM provides answers for each edge in the complete logic chain, we can integrate all the answers to identify a specific configuration debugging question.

## 5 TELEMETRY DATA ANALYSIS

We introduce a telemetry data analysis approach consisting of anomalous telemetry detection, telemetry-to-text and relevant telemetry selection. We develop techniques to address the following two challenges in the above three steps. The first challenge is detecting anomalous telemetries from a huge amount of time-series data, as it is not easy to identify what is anomalous. To address this, we first use a seasonal-trend decomposition method to decompose time series data and detect anomalous telemetries using statistics on the residual. Secondly, there may exist hundreds of telemetries,

many of which may not be related to our NL question  $q$ . To tackle this challenge, we propose a telemetry-to-text mechanism that enables to select telemetries relevant to  $q$ .

### 5.1 Anomalous Telemetry Detection

Given a time series  $T_{s_i}$  of the telemetry metric  $s_i$ , the goal of this step is to detect a set of  $\{(j, t_j)\}$  where the value  $t_j$  on timestamp  $j$  in  $T_{s_i}$  is anomalous (higher or lower than normal values). Inspired by [14, 20], for  $T_{s_i}$ , we use a seasonal-trend decomposition mechanism based on LOESS (STL) [14], which is a method that decomposes time series into seasonal, trend, and residual components for time series analysis, i.e.,

$$T_{s_i} = Seasonal_{s_i} + Trend_{s_i} + \epsilon_{s_i} \quad (2)$$

where  $Seasonal_{s_i}$  represents the captured repeating seasonal patterns by smoothing the data within each period.  $Trend_{s_i}$  is the long-term trend captured in the time series using local regression smoothing (LOESS), and  $\epsilon$  is the unexplained random part of the time series after removing the trend and seasonal components. With STL, we can decompose seasonal and trends (normal variations), and assess whether the value on each timestamp is anomalous or not by analyzing the residual data  $\epsilon_{s_i}$ . After removing the effects of seasonal and trend components, it is easier to detect anomalous deviations for  $s_i$  using statistic-based methods.

We use a statistic method called Extreme Studentized Deviate (ESD) [32] to detect anomalous deviations on  $\epsilon_{s_i}$ . The advantage of ESD is that it does not require pre-setting fixed thresholds or critical values for anomaly detection. Instead, it dynamically determines anomalous thresholds based on the actual distribution of data, which facilitates its adaptability across different temporal metrics in our scenarios. Specifically, ESD is an iterative algorithm to detect anomalous deviations. In each iteration  $k$ , we need to select an anomalous timestamp by calculating a deviation score  $C_{t_j}$  for the value  $t_j$  in each timestamp  $j$  and a threshold  $\lambda_k$ . We select the timestamp  $j_{max}$  with the highest deviation score  $C_{max}$  and compare it with the threshold  $\lambda_k$ . If  $C_{max} > \lambda_k$ ,  $t_{j_{max}}$  is anomalous.

and we drop it. We define the deviation score  $C_{t_j}$  as:

$$C_{t_j} = \frac{|\epsilon_{t_j} - \tilde{\epsilon}_{s_i}|}{MAD_j} \quad (3)$$

Different from using mean and standard deviation, we use median and the median absolute deviation (MAD) [21] to calculate  $C_{t_j}$ , as they are robust to outliers [23]. In particular,  $\tilde{\epsilon}_{s_i}$  in Equation (3) is the median of  $\epsilon_{s_i}$ , the residual  $\epsilon_{t_j}$  for timestamp  $j$  in  $\epsilon_{s_i}$  is:

$$\epsilon_{t_j} = t_j - Seasonal_{s_i j} - Trend_{s_i j} \quad (4)$$

The median absolute deviation (MAD) is the median of the absolute deviations from the median of all data points in  $\lambda_{s_i}$ , denoted as:

$$MAD_j = median_j(|\epsilon_j - \tilde{\epsilon}_{s_i}|) \quad (5)$$

After calculating the deviation  $C_{t_j}$ , we need to calculate the critical value  $\lambda_k$  of  $k^{th}$  iteration as threshold:

$$\lambda_k = \frac{(n-k)t_{p,n-k-1}}{\sqrt{(n-k-1+t_{p,n-k-1}^2)(n-k+1)}} \quad (6)$$

where  $t_{p,n-k-1}$  is the critical value of the t-distribution. If the point with the highest deviation exceeds the critical value  $\lambda_k$ , that point is considered an anomaly and be dropped. The process is then repeated until no new anomalies are detected or the predetermined limit on the number of anomalies  $N$  is reached and we can detect all the anomalous telemetries as  $A$ .

## 5.2 Telemetry-to-Text and Telemetry Selection

After detecting anomalous telemetries  $A$ , for each metric  $s_i$  in  $T_{s_i}$ , we would like to convert these anomalies into NL descriptions  $p_{s_i}$ . In particular, each description  $p_{s_i} \in p_{s_i}$  should include the metric, an explanation of the metric, a description of the deviation (including the normal value before the deviation and the abnormal value after the deviation) and the corresponding related knobs. The reason for such a telemetry-to-text process is twofold. Firstly, it is helpful to identify the relevance between  $p_{s_i}$  and the user's question  $q$ . Secondly, converting it into natural language enables the LLM to understand the information about users' DBMS status.

To this end, we first need to identify the normal values of each detected metric  $s_i$  using a percentile point of the time series to represent the normal point (e.g. 5%). In addition, we also attach the corresponding knobs that can adjust the metric  $s_i$ . Using this Telemetry-to-Text process, we can obtain many telemetry descriptions in NL, such as "*the value of an innodb\_log\_write\_requests changed from 46492 to 96561 and the related knob is innodb\_log\_file\_size*".

Due to the input-length limitations of LLM and the possibility that irrelevant metrics may mislead the LLM, we need to select useful anomalous telemetries from a large set as prompts to assist LLM to perform reasoning of DBMS diagnosis. Intuitively, we need to select metrics that are relevant to the user's question  $q$ . To achieve this goal, we rank the anomalous telemetry descriptions on the basis of the semantic similarity between the descriptions and the user's NL question (e.g., using Sentence-Bert). Finally, we select the top- $K$  anomalous descriptions as the final prompts.

**EXAMPLE 5.** As shown in Figure 6, given a user's NL question about the slow execution time of "INSERT" statement, we monitor the

temporal performance  $T_S$  as a temporal telemetry data with hundreds of metrics about CPU, Memory and MySQL system. For time-series data  $T_{s_i}$  of each metric  $s_i$ , such as Memory\_free\_total recording the remaining memory, we iteratively detect the most anomalous value in  $T_{s_i}$ . In each iteration, we use STL to decompose  $T_{s_i}$  as seasonal, trend and residual. For residual, we used the mean absolute deviation (MAD) to assess the severity of the anomaly at each timestamp  $j$  and collect them. After all anomalous deviation collected (Memory\_free\_total, MySQL\_innodb\_log\_wirte\_requests and others), we convert them into text and select relevant telemetries using semantic similarity with question  $q$ . Finally, we concatenate the generated text as the result of telemetry data analysis (the right part of the figure).

## 6 EXPERIMENTS

### 6.1 Experimental Setup

**DBMS and Configurable Knobs.** We conduct our experiments on MySQL 5.7 and PostgreSQL (PG) 12.0. MySQL contains 582 knobs and PostgreSQL contains 202 knobs.

**Datasets.** We use four real-world datasets in two types of experimental settings (**NL** and **Runnable**). Table 1 provides statistics of the datasets. Note that, the training labels can be automatically annotated. For the original training data, we use regular expressions to extract relevant knobs from accepted answers or comments on questions from StackOverflow or the MySQL Forum. For the augmented training data, we generate questions based on a knob's manual, with the knob serving as the label corresponding to the questions. To validate labeling quality, we have randomly sampled some training and testing cases for manual verification, which have confirmed that our automatic labeling is accurate.

**(1) The NL Setting:** We use DBMS configuration debugging questions posted by real users in StackOverflow [9] and MySQL community [3]. Specifically, we consider the following three datasets.

- **MySQL Stackoverflow (MySQL SO)** contains real questions regarding MySQL DBMS from StackOverflow [9].
- **MySQL Forum** contains questions about MySQL DBMS from MySQL community [3].
- **PostgreSQL Stackoverflow (PG SO)** contains questions about PostgreSQL DBMS from StackOverflow [9].

For each of the above datasets, each instance  $\langle q, K_q \rangle$  consists of an NL question  $q$  from real-world users and the ground-truth knob configuration  $K_q$ . Note that the ground-truth  $K_q$  is manually extracted from the accepted answers or comments in Stackoverflow or MySQL community. In addition, for each dataset, we split all instances  $\{\langle q, K_q \rangle\}$  with a ratio of 7:2:1 into three subsets, namely historical questions, training data and test data. We also merge historical questions of MySQL SO and MySQL Forum to form a more complete historical question set for these two datasets.

**(2) The Runnable Setting:** One limitation of the above NL setting is that it takes the accepted answers as ground-truth configurations and cannot evaluate whether the configuration issues are solved. Moreover, the NL setting does not have the telemetry data of the underlying DBMS. Thus, we also consider a *Runnable* setting. In this setting, besides providing an NL question  $q$ , users should also give the telemetry data  $S$  of their DBMS. To this end, we derive a dataset

**MySQL Run** from the above MySQL datasets. In particular, we prepare a test set in **MySQL Run** by selecting 70 test questions that can reproduce the DBMS runtime environment from the MySQL dataset in the NL setting. Moreover, we manually design 30 NL debugging questions with runtime environments.

**Evaluation Methods.** We conduct two evaluation methods for the NL and Runnable setting respectively.

(1) *Evaluation for the NL setting.* We evaluate the performance of knob diagnosis of various algorithms. For each test question  $q$ , given a set of knobs  $K_A$  predicted by an algorithm  $A$ , we compare  $K_A$  with ground-truth  $K_q$  and calculate the Recall, Precision and F1-score of question  $q$ . F1-score is computed as  $2 \cdot P \cdot R / (P + R)$ , where precision  $P$  is the proportion of predicted correct knobs to all predicted matching pairs, recall  $R$  is the proportion of predicted correct knobs to all correct knobs. Then, for the entire test dataset, we compute the *average F1-score* to evaluate performance.

(2) *Evaluation for the Runnable setting.* We evaluate the performance of DBMS configuration debugging of various approaches. For each test instance with DBMS  $D$ , after an algorithm  $A$  predicts a set of knobs  $K_A$ , the algorithm also predicts value  $v$  for each knob ( $k = v$ ),  $k \in K_A$ , which are used to tune DBMS  $D$ . Then, we employ DBMS experts to verify whether the recommended configurations successfully solve the issue. Next, we introduce *SuccessRate*, which is the ratio of successful configurations that solve the issues over all the issues,  $\text{SuccessRate} = \frac{\sum_{q \in Q_{\text{Test}}} \text{Solved}(q)}{|Q_{\text{Test}}|}$ .

**Methods Compared.** We compare the following methods.

**PLM.** For pre-trained language model (PLM), we use Sentence-BERT [31] with the published model [8] as the base model, which takes NL questions as input. Moreover, we add an output layer that outputs a vector of size  $|K|$ , where  $K$  represents all knobs. Each dimension of this vector represents the score that the corresponding knob should be recommended. **We train PLM using the training data and PLM+DA** using augmented training data in the NL setting. During the inference phase, we select the set of knobs whose output values are higher than 0.5.

**LLMs.** For large language models, we consider open-source and closed-source LLMs. For open-source LLMs, we evaluate **Llama8B** (Meta-Llama-3-8B-instruct[2]), **Llama70B** (Meta-Llama-3-70B-instruct[1]), **Qwen7B** (Qwen2-7B-Instruct[7]) and **Qwen72B** (Qwen2-72B-Instruct[6]). For closed-source LLMs, we evaluate **GPT-3.5** and **GPT-4**. For each LLM, prompt contains a task instruction and an NL question. Moreover, we ask the LLM to strictly output a knob list so that we can extract the knob set for evaluation. For value recommendation, we ask LLMs to strictly output a dictionary as `knob : value`. If the LLMs do not output the correct format, we regard it as failures.

**LLMs (all knobs).** LLMs (all knobs) improves the LLMs by modifying the prompts, i.e., adding all knobs  $\mathbb{K}$  to the original prompt and requiring the model to select among them. We also ask the LLMs to output a knob list and dictionary for value recommendation.

**LLMs (prompt engineering).** We use two prompting strategies: (1) **CoT**, which asks the LLM to think step by step, and (2) **Task Decomposition**, which asks the LLM to decompose the entire process into multiple steps and solve them in a divide-and-conquer manner.

**Table 1: Statistics of Datasets, where  $Q$  means historical questions and  $M$  means DBMS manuals for retrieval, “# test/train” means the number of question in test/train set and “% test/| $Q$ |” means the coverage ratio of knobs in test/historical questions set. Moreover,  $|M|$  covers all knobs.**

Datasets	# test	# train	$Q$	$M$	% test	%   $Q$	Type
MySQL SO	174	620	1632	3506	30%	25%	NL
MySQL Forum	49	620	1632	3506	17%	25%	NL
PG SO	57	114	402	756	38%	19%	NL
MySQL Run	100	620	1632	3506	15%	25%	Runnable

**LLMs (SFT).** We use the training data to construct the “Instruction-Answer” pairs for fine-tuning gpt-3.5-turbo, where the “Instruction” is the prompts in the *LLM* setting, and the “Answer” consists of a list of ground-truth knobs from the training data. For inference, we apply the same prompting strategies in the *LLMs* setting.

**Implementation.** All experiments were conducted on a device with Linux Ubuntu 20.04.1, 20 vCPU cores and 1.0TB memory. For document retrieval, we use top-5 retrieved documents (questions & manuals). For GPTs, we use Python to call OpenAI’s API, where the versions are gpt-3.5-turbo for GPT-3.5 and gpt-4-turbo GPT-4 with *temperature* = 0 and *random\_seed* = 42. For telemetry data analysis, we use Prometheus to monitor 557 performance metrics with a timestamp interval of 1 second. **We use Meta Faiss as our vector database to index the embeddings of historical questions and manuals.** In the current implementation of **ANDROMEDA**, we have indexed 5,138 vectors for MySQL and 1,158 vectors for PostgreSQL, with each vector having a dimension of 768.

## 6.2 Overall Comparisons

**Exp-1: How does ANDROMEDA perform compared with LM-based methods?** We report the overall results in Table 2.

**Evaluation in the NL Setting.** We report the Recall, Precision and F1-scores of different approaches on the three datasets, MySQL SO, MySQL Forum and PG SO, in the NL setting, as shown in Table 2. The PLM-based method does not perform well due to the limited availability of training data. Even though PLM+DA incorporates more training data, the improvement still remains marginal. This is because the PLM, with its fewer parameters, struggles to understand the context, particularly in longer and expert-domain texts.

LLM-based methods (LLMs in Table 2) perform better than PLM-based methods. However, they also have unsatisfactory performance. The main reason is hallucination - they often predict knobs that do not exist in reality, even for GPT-4. **Moreover, GPT models outperform open-source LLMs, while smaller-sized models (Llama8B and Qwen7B) fail to deliver satisfactory performance.**

This phenomenon gets better with LLMs (all knobs), because this essentially turns a generation problem into a selection problem, which significantly reduces hallucinations. However, these methods still have limitations. In the database configuration debugging scenario, the database contains a large number of knobs with complex functions, and there are intricate logical relationships between the knobs and anomalies. To deal with such complex questions, LLM models still lack expert knowledge and experience. Additionally, we find that sometimes GPT-4 does not perform better than GPT-3.5, such as in the MySQL Forum, because GPT-4 may not strictly output results as our required format.

**Table 2: Overall results, where we measure Precision, Recall and F1-score in NL setting, and SuccessRate in Runnable setting.**

Methods	Settings	MySQL SO			MySQL Forum			PG SO			MySQL Run SuccessRate
		Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score	
PLM	PLM	0.529	0.004	0.008	0.547	0.008	0.015	<b>0.558</b>	0.003	0.006	-
	PLM+DA	0.518	0.004	0.008	<b>0.587</b>	0.009	0.017	0.485	0.005	0.01	-
LLMs	GPT-3.5	0.329	0.205	0.221	0.207	0.145	0.153	0.32	0.183	0.201	0.37
	GPT-4	0.291	0.201	0.222	0.143	0.052	0.06	0.272	0.14	0.161	0.43
	Llama8B	0.07	0.025	0.035	0.078	0.021	0.033	0.056	0.023	0.032	0.25
	Llama70B	0.054	0.027	0.034	0.051	0.026	0.034	0.091	0.031	0.045	0.26
	Qwen7B	0.024	0.024	0.019	0.008	0.006	0.007	0.044	0.015	0.021	0.11
	Qwen72B	0.208	0.125	0.140	0.171	0.112	0.116	0.114	0.05	0.065	0.31
LLMs (all knobs)	GPT-3.5	0.448	0.293	0.308	0.411	0.247	0.246	0.465	0.268	0.299	0.52
	GPT-4	0.507	0.300	0.332	0.371	0.227	0.235	0.504	0.38	0.390	0.49
	Llama8B	0.116	0.048	0.057	0.11	0.054	0.058	0.263	0.094	0.121	0.31
	Llama70B	0.536	0.217	0.268	0.368	0.166	0.201	0.488	0.25	0.277	0.57
	Qwen7B	0.103	0.086	0.087	0.136	0.164	0.137	0.232	0.247	0.222	0.10
	Qwen72B	0.451	0.3	0.321	0.384	0.248	0.26	0.521	0.293	0.305	0.54
LLMs (PE)	CoT	0.308	0.170	0.196	0.161	0.115	0.121	0.226	0.144	0.158	0.53
	Task Dec.	0.143	0.136	0.127	0.111	0.160	0.100	0.207	0.169	0.179	0.5
LLM (SFT)	GPT-3.5	0.345	0.391	0.347	0.333	0.6	0.390	0.343	0.383	0.347	0.46
Andromeda	GPT-3.5	0.512	0.318	0.348	0.541	0.383	0.382	0.461	0.305	0.321	<b>0.79</b>
	GPT-4	<b>0.557</b>	0.426	<b>0.441</b>	0.528	<b>0.495</b>	<b>0.449</b>	0.47	0.402	0.398	0.76
	Llama8B	0.32	0.183	0.201	0.441	0.244	0.256	0.39	0.241	0.270	0.44
	Llama70B	0.507	0.315	0.345	0.482	0.331	0.347	0.452	0.31	0.335	0.60
	Qwen7B	0.294	0.245	0.242	0.253	0.192	0.195	0.275	0.225	0.222	0.46
	Qwen72B	0.529	<b>0.435</b>	0.437	0.488	0.45	0.412	0.449	<b>0.415</b>	<b>0.410</b>	0.60

We can see that ANDROMEDA outperforms other methods significantly. This is mainly because ANDROMEDA employs a retrieval-augmented generation (RAG) strategy that enriches the NL debugging questions with domain-specific context drawn from multiple sources, including historical questions and troubleshooting manuals. The results have shown that the RAG strategy can significantly improve the quality and relevance of its recommendations. Moreover, with ANDROMEDA, the performance gap between open-source models and GPTs narrows. Larger-size open-source LLMs (Llama70B and Qwen72B) become comparable to GPTs due to the advantage of retrieved knowledge.

The prompt engineering methods (CoT and Task Decomposition in Table 2) are not very effective because the configuration debugging process doesn't always involve multiple steps for decomposition. Instead, it requires expert knowledge of DBMS. Additionally, the fine-tuning methods do not show significant improvement, primarily due to the limitations of the training data. Unlike other classification tasks, knob diagnosis requires predictions across hundreds of knobs with diverse functions, each needing substantial training data to achieve accurate results.

**Evaluation in the Runnable setting.** ANDROMEDA produces high-accuracy (over 0.7) and outperforms other approaches, which means most cases are practically solved by ANDROMEDA. Note that the result of PLM is not reported, because it predicts a very large number of knobs (more than 100) and fails to solve the issues.

In detailed analysis, the number of knobs predicted on average by ANDROMEDA is 4.85 with GPT-3.5 and 2.63 with GPT-4, which means that users can solve most problems by adjusting just a few knobs. Interestingly, ANDROMEDA achieves higher accuracy when

using GPT-3.5 compared to GPT-4. This is partly attributed to GPT-4's tendency to deviate from our specified output requirements, and it produces fewer knobs. In contrast, when the number of adjustable knobs is controlled, GPT-3.5 recommends more knobs, making it easier to find knobs that can solve the problem. Compared with other methods, ANDROMEDA shows significant improvements compared with GPTs and GPTs (all knobs).

**Finding 1: ANDROMEDA outperforms other LM-based methods in both NL and Runnable evaluation settings, demonstrating the effectiveness of our RAG strategy.**

### 6.3 Evaluation on Document Retrieval

In this section, we evaluate the impact of our document retrieval module. In addition to F1-score for end-to-end results, we also directly evaluate the retrieval quality. We measure the quality of the documents by comparing the knobs of the top- $k$  retrieved documents with the ground truth knobs, thereby eliminating the influence of the reasoner module on the evaluation. To achieve it, we utilize **Document retrieval recall** metric. For algorithm  $A$  on test dataset  $Q_{\text{Test}}$ , **Document retrieval recall** is calculated as:  $\text{Recall}_{\text{doc}}(A, k, Q_{\text{Test}}) = \frac{1}{|Q_{\text{Test}}|} \sum_{q \in Q_{\text{Test}}} \frac{|(\bigcup_{d \in D_{A,k,q}} K_d) \cap K_q|}{|K_q|}$ , where  $D_{A,k,q}$  is the top- $k$  retrieved documents from retrieval data of  $A$  on  $Q$ , and  $K_d$  is the corresponding knobs of document  $d$ .

**Exp-2: How does data augmentation work in document retrieval module?** In order to see the effect of the augmentation module, we consider an **ANDROMEDA-docs-only** version of our framework by disabling telemetry analysis. We compare **ANDROMEDA-docs-only** with the following document augmentation methods:

**(1) No Augment:** training model using the original training data (as shown in Table 1). **(2) CoT Augment:** let LLM generate the same volume questions from manuals using CoT method. The results are provided in Table 3 for **Document retrieval recall** and Table 4 for **F1-score**, which show that ANDROMEDA-docs-only outperforms these two methods. ANDROMEDA-docs-only and CoT Augment perform better than No Augment, showing that the new generated question-knob pairs are helpful. ANDROMEDA-docs-only performs better than CoT Augment due to our proposed logic chain strategy.

**Finding 2: The logic chain based data synthesis strategy for training data augmentation is very helpful and outperforms other data augmentation methods.**

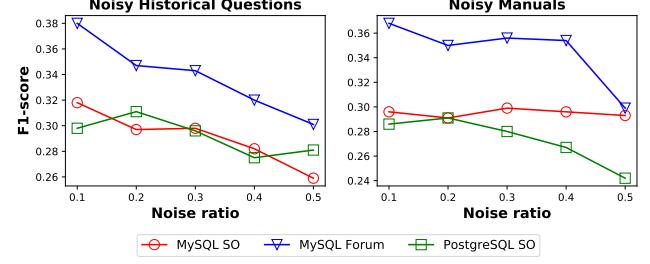
**Exp-3: How does document representation model work in document retrieval?** We compare **ANDROMEDA-docs-only** with document representation methods: (1) **Q-only**: only retrieving most similar questions using Sentence-BERT. (2) **M-only**: only retrieving most similar manuals using Sentence-BERT. (3) **No train**: retrieving similar questions and manuals using Sentence-BERT.

The results are reported in Table 5 and Table 6. For document retrieval recall shown in Table 5, ANDROMEDA-docs-only outperforms the other three methods. Compared with Q-only and M-only, No Train performs better, showing the benefit of retrieving documents from multiple sources. However, No Train has limited performance compared with ANDROMEDA-docs-only, which validates our claim that the semantics of documents from different sources may be heterogeneous. These results also demonstrate the effectiveness of our proposed contrastive learning approach that aligns documents from different sources into a unified representation space. Table 6 shows the F1-score of DBMS configuration debugging results with the top-5 retrieved documents. ANDROMEDA-docs-only also outperforms other methods in all datasets, showing the benefits of the unified representation learning model.

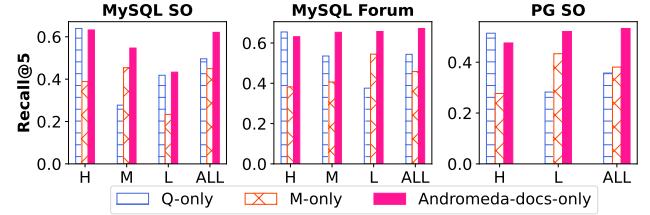
**Different knob frequencies.** Next, we provide an in-depth analysis to examine the effect of knob frequency in historical questions. We first rank knobs according to their occurrence frequency in historical questions and then categorize them into high, medium, and low-frequency buckets.

We report the recall@5 in Figure 8 for retrieval. The results show that, during the retrieval stage, as the knob frequency decreases, the document retrieval recall of Q-only continuously decreases. This indicates that, if the knobs are rare in the historical question set, it is difficult to find useful information solely by retrieving similar questions. In contrast, the document retrieval recall of M-only performs similarly across different knob frequencies, and is lower than that of similar questions in high frequency. This suggests that due to the distribution differences between the questions and the manuals, finding similar manuals is a more challenging process. However, it is not affected by the frequency of the knob and is relatively effective for rare knobs. The performance of ANDROMEDA-docs-only is generally superior demonstrating that it is beneficial to retrieve documents from multiple sources and our contrastive learning method is effective.

We report the F1-score in Figure 9 for knob diagnosis, which illustrates the impact of different retrieval data sources on the knob diagnosis of LLMs under knobs of varying frequencies. For both



**Figure 7: F1-score of knob diagnosis for retrieval document with different quality of source.**



**Figure 8: Document retrieval recall@5 in the NL evaluation setting by varying knob frequencies , where ‘H’/‘M’/‘L’ means high/median/low frequency, respectively.**

GPT-3.5 and GPT-4, the F1-score of **ANDROMEDA-docs-only** is generally higher than that of Q-only and M-only. Furthermore, with **ANDROMEDA-docs-only**, the frequency of knobs has a smaller effect on the knob diagnosis of LLMs. This is attributed to the advantages of document representation with two types: question retrieval aids in high-frequency knobs, while manual retrieval compensates for the deficiencies of question retrieval in low-frequency knobs.

**Finding 3: Combining different types of documents is more effective than using a single type of document. Also, our representation model can effectively align different types of documents, achieving better document retrieval results.**

**Exp-4: How does inaccuracies in the retrieval sources impact the performance of ANDROMEDA?** We manually inject noise into these sources by replacing the ground-truth labels in historical questions and database manuals with incorrect ones in various ratios, i.e., 0.1, 0.2, 0.3, 0.4, and 0.5. Figure 7 shows that as the noise ratio increases, the overall performance of ANDROMEDA decreases, highlighting the importance of the quality of retrieval sources. In particular, on some datasets (e.g., MySQL SO in the Noisy Manuals setting), the performance does not significantly change. This is because, despite the presence of noisy documents, useful ones can still be retrieved to guide the LLMs.

**Finding 4: The inaccuracies in the retrieval sources would impact the overall performance of ANDROMEDA.**

## 6.4 Evaluation on Telemetry Data Analysis

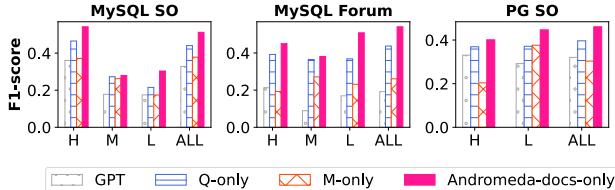
**Exp-5: How does telemetry data analysis module work?** To examine the effect of our telemetry data module, we compare GPTs and ANDROMEDA with/without telemetry data analysis. We compare the quality of (1) GPTs, (2) **ANDROMEDA-telemetry-only**:

**Table 3: Document retrieval recall of different augmentation methods in document retrieval in the NL evaluation setting.**

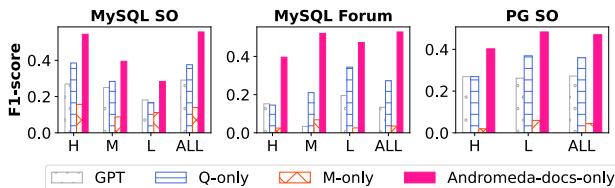
Datasets	MySQL SO				MySQL Forum				PG SO			
	top- <i>k</i>	1	3	5	10	1	3	5	10	1	3	5
No Augment	0.323	0.407	0.474	0.55	0.314	0.474	0.582	0.63	0.194	0.277	0.35	0.407
CoT Augment	0.32	0.435	0.516	0.587	0.305	0.431	0.524	0.612	0.247	0.373	0.454	<b>0.556</b>
ANDROMEDA-docs-only	<b>0.327</b>	<b>0.461</b>	<b>0.54</b>	<b>0.621</b>	<b>0.336</b>	<b>0.492</b>	<b>0.595</b>	<b>0.672</b>	<b>0.299</b>	<b>0.426</b>	<b>0.473</b>	0.533

**Table 4: Reasoning result on F1-score of augmentation methods, where No Aug. means No Augment, CoT Aug. means CoT Augment, and And-d means ANDROMEDA-docs-only.**

Datasets	Reasoners	No Aug.	CoT Aug.	And-d
MySQL SO	GPT 3.5	0.317	0.331	<b>0.348</b>
	GPT 4	0.365	0.407	<b>0.441</b>
MySQL Forum	GPT 3.5	<b>0.395</b>	0.387	0.382
	GPT 4	0.400	0.360	<b>0.449</b>
PG SO	GPT 3.5	0.256	0.290	<b>0.321</b>
	GPT 4	0.286	0.338	<b>0.398</b>



(a) Using GPT-3.5 for reasoning.



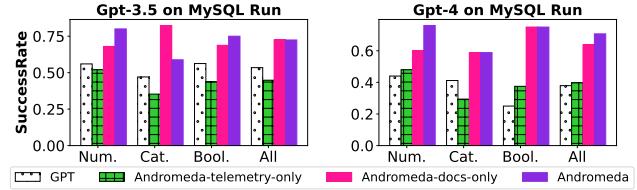
(b) Using GPT-4 for reasoning.

**Figure 9: F1-scores of knob diagnosis in the NL evaluation setting by varying knob frequencies, where ‘H’/‘M’/‘L’ means high/median/low frequency, respectively.**

ANDROMEDA without document retrieval, (3) **ANDROMEDA-docs-only**: ANDROMEDA without telemetry data analysis, and (4) **ANDROMEDA** respectively.

The results are presented in Table 7. GPTs and ANDROMEDA-telemetry-only perform similarly. This indicates that using telemetry data analysis alone has limit effects as it only diagnosis a part of configuration issues that are highly related to telemetries: the telemetry data analysis can provide description on anomalous telemetry changes, which are useful for diagnosing anomalies caused by these anomalous changes. However, some questions are not due to the state of runtime environment resources, so using only telemetry data analysis is insufficient.

Comparing the results of ANDROMEDA-docs-only and ANDROMEDA, we can observe how telemetry data analysis works after adding document retrieval. The results show that ANDROMEDA outperforms ANDROMEDA-docs-only, demonstrating the effectiveness

**Figure 10: SuccessRates on various knob types.**

of telemetry data analysis. Question and manual retrieval can add more domain knowledge, while telemetry data analysis can further determine issues through metric changes.

**Different knob type.** We provide an in-depth analysis to examine the effect of varying types of knobs by categorizing them based on their value ranges. According to the descriptions of knobs in the manual, we classify all knobs into Numeric, Categorical, and Boolean types. We calculate the accuracy to analyze the knobs based on these different categories in the Runnable setting. The results are reported in Figure 10. ANDROMEDA outperforms other methods across various types of knobs, indicating that both telemetry data analysis and text retrieval are beneficial for knob diagnosis. More specifically, the impact of document retrieval is more pronounced, as ANDROMEDA-docs-only generally performs better than ANDROMEDA-telemetry-only. This is because related questions and manuals contain richer knowledge. However, for numeric knobs, we observe that the improvement of ANDROMEDA-docs-only is not as significant for GPT-3.5 compared to other types of knobs. Moreover, ANDROMEDA significantly outperforms ANDROMEDA-docs-only on numeric knobs. Numeric knobs typically control resource allocation and usage, such as `innodb_buffer_pool_size` in MySQL, and these knobs often address issues caused by improper resource allocation. These issues are reflected in telemetry data analysis, making telemetry data analysis more helpful for diagnosing numeric knobs.

**Finding 5: The telemetry data analysis module is helpful for DBMS configuration debugging. When including retrieved documents, the improvement from adding telemetry data analysis is more significant.**

## 6.5 Efficiency Evaluation

**Exp-6: How efficient is ANDROMEDA?** We report the average runtime breakdown for diagnosis debugging in ANDROMEDA, as shown in Table 8. We have the following observations. First, the average end-to-end runtime used to recommend knobs is in seconds, which means that ANDROMEDA is efficient to meet DBMS configuration debugging requirements. Second, we find that telemetry data analysis and reasoning take most of the time. For telemetry data analysis, all metrics need seasonal-trend decomposition and statistical test, which would take time. For reasoning, the LLM has a large number of parameters and thus requires some time to generate outputs.

**Table 5: Document retrieval recall of different documents representation methods in the NL evaluation setting.**

Datasets	MySQL SO				MySQL Forum				PG SO			
	top- <i>k</i>	1	3	5	10	1	3	5	10	1	3	5
Q-only	0.265	0.36	0.412	0.496	0.257	0.461	0.494	0.544	0.203	0.277	0.315	0.357
M-only	0.218	0.345	0.391	0.45	0.253	0.368	0.392	0.458	0.158	0.175	0.228	0.381
No Train	0.298	0.453	0.533	0.615	0.233	0.445	0.526	0.581	0.203	0.277	0.312	0.438
ANDROMEDA-docs-only	<b>0.327</b>	<b>0.461</b>	<b>0.54</b>	<b>0.621</b>	<b>0.336</b>	<b>0.492</b>	<b>0.595</b>	<b>0.672</b>	<b>0.299</b>	<b>0.426</b>	<b>0.473</b>	<b>0.533</b>

**Table 6: F1-scores of knob diagnosis of different document representation methods (And-d:ANDROMEDA-docs-only).**

Datasets	Reasoners	Q-only	M-only	No Train	And-d
MySQL SO	GPT 3.5	0.334	0.304	0.338	<b>0.348</b>
	GPT 4	0.304	0.102	0.400	<b>0.441</b>
MySQL Forum	GPT 3.5	0.361	0.233	0.367	<b>0.382</b>
	GPT 4	0.220	0.04	0.381	<b>0.449</b>
PG SO	GPT 3.5	0.298	0.244	0.251	<b>0.321</b>
	GPT 4	0.308	0.03	0.261	<b>0.398</b>

**Table 7: SuccessRates w/o telemetry analysis (And-d: Andromeda-docs-only; And-t: Andromeda-telemetry-only).**

Datasets	Reasoners	GPTs	And-t	And-d	ANDROMEDA
MySQL Run	GPT 3.5	0.37	0.34	0.58	<b>0.79</b>
	GPT 4	0.43	0.64	0.63	<b>0.76</b>

**Table 8: Time of ANDROMEDA (in seconds).**

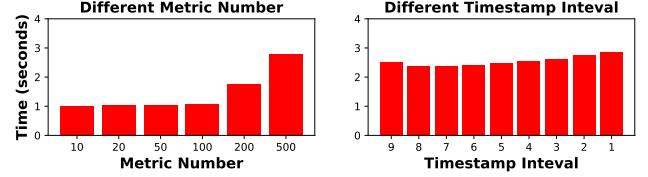
Time	Document Retrieval	Telemetry Analysis	Reasoning	Total
GPT 3.5	0.07	2.85	2.45	5.37
GPT 4	0.07	2.85	3.87	6.79

**Table 9: Comparison of on SuccessRate and end-to-end time with using manuals and search engines in Runnable setting.**

Metrics	Successful Ratio		Time (minutes)	
	# 1	# 2	# 1	# 2
DBA Experts	0.42	0.38	14.79	24.51
using manuals	0.54	0.46	14.55	9.75
using search engine	<b>0.76</b>	<b>0.76</b>	<b>7.55</b>	<b>3.79</b>

**Finding 6: ANDROMEDA can efficiently support automatic DBMS configuration debugging by answering a wide range of NL questions regarding configuration issues.**

**Exp-7: Whether ANDROMEDA effectively reduces the end-to-end time?** We evaluate whether ANDROMEDA effectively reduce the end-to-end time (*i.e.*, operational burdens) for DBAs. To this end, we employ two experienced DBAs to solve questions in the Runnable setting and provide them with historical questions, database manuals, and telemetry data. We then consider three strategies: (1) *using manuals*: DBAs manually search DBMS manuals to provide diagnostic suggestions for tuning knobs; (2) *using search engines*: DBAs use search engine to provide diagnostic suggestions for tuning knobs; (3) *using Andromeda*: DBAs follow ANDROMEDA’s recommendations to address the issues. We compare the average time taken by the DBAs to troubleshoot configuration issues, as well as the SuccessRate scores. As shown in Table 9, ANDROMEDA achieves a higher SuccessRate than the alternative strategies and significantly reduces end-to-end time. These results demonstrate that ANDROMEDA effectively reduces end-to-end time for DBAs while providing accurate troubleshooting suggestions.

**Figure 11: Evaluation on scalability of telemetry analysis.**

**Finding 7: ANDROMEDA significantly reduces the end-to-end time for DBAs during database configuration debugging.**

**Exp-8: How does ANDROMEDA perform in terms of the scalability for telemetry data analysis.** We vary the timestamp interval (*i.e.*, the sampling interval for each telemetry metric  $s_i$ ) and the number of metrics (*i.e.*, the total count of telemetry metrics  $s$ ). As shown in Figure 11, with the increase of metric number and the decrease of timestamp interval, the time required for telemetry analysis increases in a sub-linear manner. Moreover, the overall time remains within a few seconds, which is totally acceptable in our configuration debugging settings.

**Finding 8: ANDROMEDA performs well in terms of scalability for telemetry data analysis.**

**Exp-9: Whether ANDROMEDA has good performance on cost efficiency?** We respectively calculate the average cost regarding calling APIs for the approaches, namely LLMs, LLMs (all knobs) and ANDROMEDA for GPT 3.5 and GPT 4. For each NL debugging question, we have the results: 0.0017\$ for GPT-3.5, 0.053\$ for GPT-4, 0.012\$ for GPT-3.5 (all knobs), 0.039\$ for GPT-4 (all knobs), 0.006\$ for ANDROMEDA-3.5 and 0.022\$ for ANDROMEDA-4.

**Finding 9: ANDROMEDA performs well on cost efficiency.**

## 7 CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced a novel framework ANDROMEDA which serves as a natural surrogate of DBAs to answer a wide range of natural language (NL) questions regarding DBMS configuration issues, and generates recommendations to fix these issues. ANDROMEDA utilizes a retrieval-augmented generation (RAG) strategy that effectively provides domain-specific context for questions from multiple sources, which significantly improves the performance of DBMS configuration debugging. Experiments on real-world DBMS configuration debugging datasets show that ANDROMEDA significantly outperforms existing solutions.

Several questions still remain to be answered. First, besides DBMS configuration debugging, we can consider addressing more types of issues, such as SQL rewriting, index tuning, etc. Second, a more effective method should be proposed to assist the LLM in value recommendation for knobs. This is a more challenging problem because the range of values can be continuous.

## REFERENCES

- [1] [n. d.]. Llama-3.1-70B-Instruct. <https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct>.
- [2] [n. d.]. Llama-3.1-8B-Instruct. <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>.
- [3] [n. d.]. MySQL Forum Dataset. <https://forums.mysql.com>.
- [4] [n. d.]. MySQL official manual. <https://dev.mysql.com/doc/refman/8.0/en/>.
- [5] [n. d.]. PostgreSQL official manual. <https://www.postgresql.org/docs/>.
- [6] [n. d.]. Qwen2-72B-Instruct. <https://huggingface.co/Qwen/Qwen2-72B-Instruct>.
- [7] [n. d.]. Qwen2-7B-Instruct. <https://huggingface.co/Qwen/Qwen2-7B-Instruct>.
- [8] [n. d.]. Sentence-Bert Model. <https://huggingface.co/sentence-transformers/all-mmpnet-base-v2>.
- [9] [n. d.]. StackOverflow Dataset. <https://www.kaggle.com/datasets/stackoverflow/stackoverflow>.
- [10] [n. d.]. Technique report of Andromeda. <https://anonymous.4open.science/r/Andromeda-D150>.
- [11] Dana Van Aken, Andrew Pavlo, Geoffrey J. Gordon, and Bohan Zhang. 2017. Automatic Database Management System Tuning Through Large-scale Machine Learning. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14–19, 2017*, Semih Salihoglu, Wenchao Zhou, Rada Chirkova, Jun Yang, and Dan Suciu (Eds.). ACM, 1009–1024. <https://doi.org/10.1145/3035918.3064029>
- [12] Akari Asai, Ziqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *CoRR* abs/2310.11511 (2023). [https://doi.org/10.48550/ARXIV.2310.11511 arXiv:2310.11511](https://doi.org/10.48550/ARXIV.2310.11511)
- [13] Junwei Bao, Nan Duan, Ming Zhou, and Tiejun Zhao. 2014. Knowledge-Based Question Answering as Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22–27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*. The Association for Computer Linguistics, 967–976. <https://doi.org/10.3115/V1/P14-1091>
- [14] R. B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma J. Terpenning. 1990. STL: A seasonal-trend decomposition procedure based on loess (with discussion). <https://api.semanticscholar.org/CorpusID:268093671>
- [15] Karl Dias, Mark Ramacher, Uri Shaft, Venkateshwaran Venkataramani, and Graham Wood. 2005. Automatic Performance Diagnosis and Tuning in Oracle. In *Second Biennial Conference on Innovative Data Systems Research, CIDR 2005, Asilomar, CA, USA, January 4–7, 2005, Online Proceedings*. www.cidrdb.org, 84–94. <http://cidrdb.org/cidr2005/papers/P07.pdf>
- [16] Karl Dias, Mark Ramacher, Uri Shaft, Venkateshwaran Venkataramani, and Graham Wood. 2005. Automatic Performance Diagnosis and Tuning in Oracle. In *Second Biennial Conference on Innovative Data Systems Research, CIDR 2005, Asilomar, CA, USA, January 4–7, 2005, Online Proceedings*. www.cidrdb.org, 84–94. <http://cidrdb.org/cidr2005/papers/P07.pdf>
- [17] Songyun Duan, Vamsidhar Thummalapala, and Shivnath Babu. 2009. Tuning Database Configuration Parameters with iTuned. *Proc. VLDB Endow.* 2, 1 (2009), 1246–1257. <https://doi.org/10.14778/1687627.1687767>
- [18] Ayat Fekry, Lucian Carata, Thomas F. J.-M. Pasquier, Andrew Rice, and Andy Hopper. 2020. To Tune or Not to Tune? In Search of Optimal Configurations for Data Analytics. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23–27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 2494–2504. <https://doi.org/10.1145/3394486.3403299>
- [19] Matthias Feurer. 2018. Scalable Meta-Learning for Bayesian Optimization using Ranking-Weighted Gaussian Process Ensembles. <https://api.semanticscholar.org/CorpusID:51795721>
- [20] Xiao He, Ya Li, Jian Tan, Bin Wu, and Feifei Li. 2023. OneShotSTL: One-Shot Seasonal-Trend Decomposition For Online Time Series Anomaly Detection And Forecasting. *Proc. VLDB Endow.* 16, 6 (2023), 1399–1412. <https://doi.org/10.14778/3583140.3583155>
- [21] Jordan Hochenbaum, Owen S. Vallis, and Arun Kejariwal. 2017. Automatic Anomaly Detection in the Cloud Via Statistical Learning. *CoRR* abs/1704.07706 (2017). [arXiv:1704.07706 http://arxiv.org/abs/1704.07706](https://arxiv.org/abs/1704.07706)
- [22] Xinghang Hu. 2023. Enhancing Answer Selection in Community Question Answering with Pre-trained and Large Language Models. *CoRR* abs/2311.17502 (2023). [https://doi.org/10.48550/ARXIV.2311.17502 arXiv:2311.17502](https://doi.org/10.48550/ARXIV.2311.17502)
- [23] Peter J Huber and Elvezio M Ronchetti. 2011. *Robust statistics*. John Wiley & Sons.
- [24] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. *CoRR* abs/2403.14403 (2024). [https://doi.org/10.48550/ARXIV.2403.14403 arXiv:2403.14403](https://doi.org/10.48550/ARXIV.2403.14403)
- [25] Jiale Lao, Yibo Wang, Yufei Li, Jianping Wang, Yunjia Zhang, Zhiyuan Cheng, Wanghu Chen, Mingjie Tang, and Jianguo Wang. 2023. GPTuner: A Manual-Reading Database Tuning System via GPT-Guided Bayesian Optimization. *CoRR* abs/2311.03157 (2023). [https://doi.org/10.48550/ARXIV.2311.03157 arXiv:2311.03157](https://doi.org/10.48550/ARXIV.2311.03157)
- [26] Guoliang Li, Xuanhe Zhou, Shifu Li, and Bo Gao. 2019. QTune: A Query-Aware Database Tuning System with Deep Reinforcement Learning. *Proc. VLDB Endow.* 12, 12 (2019), 2118–2130. <https://doi.org/10.14778/3352063.3352129>
- [27] Yuhe Liu, Changhua Pei, Longlong Xu, Bohan Chen, Mingze Sun, Zhirui Zhang, Yongqian Sun, Shenglin Zhang, Kun Wang, Haiping Zhang, Jianhui Li, Gaogang Xie, Xidao Wen, Xiaohui Nie, and Dan Pei. 2023. OpsEval: A Comprehensive Task-Oriented AIOps Benchmark for Large Language Models. *CoRR* abs/2310.07637 (2023). <https://doi.org/10.48550/ARXIV.2310.07637 arXiv:2310.07637>
- [28] Minghua Ma, Zheng Yin, Shenglin Zhang, Sheng Wang, Christopher Zheng, Xin-hao Jiang, Hanwen Hu, Cheng Luo, Yilin Li, Nengjun Qiu, Feifei Li, Changcheng Chen, and Dan Pei. 2020. Diagnosing Root Causes of Intermittent Slow Queries in Large-Scale Cloud Databases. *Proc. VLDB Endow.* 13, 8 (2020), 1176–1189. <https://doi.org/10.14778/3389133.3389136>
- [29] Reieichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *CoRR* abs/2112.09332 (2021). [arXiv:2112.09332 https://arxiv.org/abs/2112.09332](https://arxiv.org/abs/2112.09332)
- [30] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. <https://doi.org/10.18653/V1/D19-1410>
- [31] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. <https://doi.org/10.18653/V1/D19-1410>
- [32] Bernard Rosner. 1975. On the detection of many outliers. *Technometrics* 17, 2 (1975), 221–227.
- [33] Chirag Shah and Jeffrey Pomerantz. 2010. Evaluating and predicting answer quality in community QA. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19–23, 2010*, Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Eftimis N. Eftimidis, and Jacques Savoy (Eds.). ACM, 411–418. <https://doi.org/10.1145/1835449.1835518>
- [34] Vikramank Y. Singh, Kapil Vaidya, Vinay Shekhar Bannhatti Kumar, Sopan Khosla, Balakrishnan Narayanasamy, Rashmi Gangadharai, and Tim Kraska. 2024. Panda: Performance Debugging for Databases using LLM Agents. In *14th Conference on Innovative Data Systems Research, CIDR 2024, Chaminade, HI, USA, January 14–17, 2024*. www.cidrdb.org. <https://www.cidrdb.org/cidr2024/papers/p6-singh.pdf>
- [35] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics* 11 (2023), 1–17.
- [36] Ulli Waltinger, Alexa Breuing, and Ipke Wachsmuth. 2011. Interfacing Virtual Agents with Collaborative Knowledge: Open Domain Question Answering Using Wikipedia-Based Topic Models. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16–22, 2011*, Tony Walsh (Ed.). IJCAI/AAAI, 1896–1902. <https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-317>
- [37] Dong Young Yoon, Ning Niu, and Barzan Mozafari. 2016. DBSherlock: A Performance Diagnostic Tool for Transactional Databases. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 – July 01, 2016*, Fatma Özcan, Georgia Koutrika, and Sam Madden (Eds.). ACM, 1599–1614. <https://doi.org/10.1145/2882903.2915218>
- [38] Dong Young Yoon, Ning Niu, and Barzan Mozafari. 2016. DBSherlock: A Performance Diagnostic Tool for Transactional Databases. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 – July 01, 2016*, Fatma Özcan, Georgia Koutrika, and Sam Madden (Eds.). ACM, 1599–1614. <https://doi.org/10.1145/2882903.2915218>
- [39] Hamed Zamani and Michael Bendersky. 2024. Stochastic RAG: End-to-End Retrieval-Augmented Generation through Expected Utility Maximization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2641–2646.
- [40] Ji Zhang, Yu Liu, Ke Zhou, Guoliang Li, Zhili Xiao, Bin Cheng, Jiashu Xing, Yangtao Wang, Tianheng Cheng, Li Liu, Minwei Ran, and Zekang Li. 2019. An End-to-End Automatic Cloud Database Tuning System Using Deep Reinforcement Learning. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 – July 5, 2019*, Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Michael J. Franklin (Eds.). ACM, 137–148. <https://doi.org/10.1145/3291380.3305830>

- and Tim Kraska (Eds.). ACM, 415–432. <https://doi.org/10.1145/3299869.3300085>
- [41] Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. 2014. Question Retrieval with High Quality Answers in Community Question Answering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, Jianzhong Li, Xiaoyang Sean Wang, Minos N. Garofalakis, Ian Soboroff, Torsten Suel, and Min Wang (Eds.). ACM, 371–380. <https://doi.org/10.1145/2661829.2661908>
- [42] Xinyi Zhang, Zhuo Chang, Yang Li, Hong Wu, Jian Tan, Feifei Li, and Bin Cui. 2022. Facilitating Database Tuning with Hyper-Parameter Optimization: A Comprehensive Experimental Evaluation. *Proc. VLDB Endow.* 15, 9 (2022), 1808–1821. <https://doi.org/10.14778/3538598.3538604>
- [43] Xinyi Zhang, Hong Wu, Zhuo Chang, Shuowei Jin, Jian Tan, Feifei Li, Tieying Zhang, and Bin Cui. 2021. ResTune: Resource Oriented Tuning Boosted by Meta-Learning for Cloud Databases. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava (Eds.). ACM, 2102–2114. <https://doi.org/10.1145/3448016.3457291>
- [44] Zhou Zhao, Lijun Zhang, Xiaofei He, and Wilfred Ng. 2015. Expert Finding for Question Answering via Graph Regularized Matrix Completion. *IEEE Trans. Knowl. Data Eng.* 27, 4 (2015), 993–1004. <https://doi.org/10.1109/TKDE.2014.2356461>
- [45] Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural Question Generation from Text: A Preliminary Study. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings (Lecture Notes in Computer Science, Vol. 10619)*, Xuanjing Huang, Jing Jiang, Dongyan Zhao, Yansong Feng, and Yu Hong (Eds.). Springer, 662–671. [https://doi.org/10.1007/978-3-319-73618-1\\_56](https://doi.org/10.1007/978-3-319-73618-1_56)
- [46] Xuanhe Zhou, Guoliang Li, Zhaoyan Sun, Zhiyuan Liu, Weize Chen, Jianming Wu, Jiesi Liu, Ruohang Feng, and Guoyang Zeng. 2023. D-bot: Database diagnosis system using large language models. *arXiv preprint arXiv:2312.01454* (2023).

## A CASE STUDY

We present two cases to demonstrate how ANDROMEDA works and identify areas for potential improvement.

### A.1 Good Cases

The first case, shown in Figure 12(a), is similar to the running example in Figure 2.

**Prompt.** In the instruction phase, the first block (gray) represents the task, assigning the role and specific responsibilities to the LLM. As indicated in the user’s question, represented by the fifth block (white), the user is working with MySQL using the InnoDB engine and needs to join two large tables for data insertion. However, the process takes several minutes, and thus the user seeks ways to optimize performance. Now, we discuss how ANDROMEDA works.

(1) For telemetry analysis, the second block (purple) provides details about changes, explaining two metrics: `created_tmp_tables` and `select_scan`. It also recommends tuning related knobs, such as `tmp_table_size`, `delay_key_write`, etc.

(2) The retrieved manual in the third block (blue) explains that using the `INSERT` command with multiple `VALUES` lists and adjusting the `bulk_insert_buffer_size` knob can accelerate the insertion process, specifically recommending tuning the `bulk_insert_buffer_size`.

(3) The retrieved historical questions in the fourth block (green) highlight two related questions about slow insertion. The first involves slow insertion of approximately 40,000 rows into a database via PHP code, with recommendations to adjust `foreign_key_check` and `unique_checks`. The second question concerns adding 300 million rows to a table at a slow pace, recommending the tuning of `innodb_buffer_pool_size`.

**Response.** For the response, the LLM predicts 10 configuration knobs along with their corresponding values. The `bulk_insert_buffer_size` is recommended based on the manuals, while `foreign_key_check`, `unique_checks`, and `innodb_buffer_pool_size` are suggested by historical questions. The remaining knobs are derived from the telemetry analysis. We tested tuning each knob to the recommended values and found the following adjustments to be particularly effective:

- `tmp_table_size` from telemetry analysis.
- `bulk_insert_buffer_size` from manual retrieval.
- `innodb_buffer_pool_size` from historical questions.
- `foreign_key_checks` from historical questions.
- `unique_checks` from historical questions.
- `sort_buffer_size` from telemetry analysis.

Some knob values were predicted as a range in the response. We tested multiple values and found that more than one setting can be effective. For example, setting `tmp_table_size` to both `70M` and `80M` yielded positive results, indicating that various values within the range can work well.

**Discussion.** With rich knowledge incorporated, the LLM accurately recommends the appropriate tuning knobs. For these multiple effective knobs, tuning a combination of them further accelerates insertion performance compared to adjusting a single knob. These knobs are sourced from various references, including manuals, historical questions, and telemetry data, highlighting the advantage of retrieval-augmented generation (RAG) from multiple sources.

### A.2 Bad Cases

The second case illustrates a failure of ANDROMEDA.

**Prompt.** As shown in Figure 13, the question (in the white block) shows that the user executed `DROP TABLE IF EXISTS "abcd"` and wanted to suppress only the warning message it triggered. Although he were aware of the `max_error_count` variable, he was concerned that setting it to zero would also suppress errors and note messages, which he wanted to avoid.

(1) For the telemetry analysis (in the purple block), ANDROMEDA reports changes and descriptions of five metrics: `connections`, `created_tmp_tables`, `table_open_cache_hits`, `select_scan`, and `bytes_sent`, recommending adjustments to `max_connections` and other related knobs.

(2) In the manual retrieval (in the blue block), ANDROMEDA retrieves three entries related to warning-related knobs. The first is a basic description of `sql_warning`, the second covers the usage of `max_error_count`, and the last explains the usage of `log_warnings` and `log_error_verbosity`.

(3) For the historical questions (in the green block), ANDROMEDA retrieves a question about how to suppress warnings in logs and recommended adjusting `log_warnings`.

**Response.** The LLM predicts `log_warnings` based on both manuals and historical questions, while `sql_warnings` and `log_error_verbosity` are suggested from the manuals.

**Discussion.** The correct solution is to set `sql_notes=0`, as detailed in the manual: the `sql_notes` system variable controls whether note messages increment the `warning_count` and whether the server stores them. By default, `sql_notes` is set to 1, but when set to 0, notes do not increment the `warning_count`, and the server does not store them.

In this case, historical questions are not helpful, as the most relevant question deals with warnings in the log, not warning messages. Without related historical questions, this source proves unhelpful. Telemetry is also not useful since adjusting warning notes does not trigger any anomalies in telemetry. For the manuals, ANDROMEDA retrieves several related entries for knobs with similar functions to `sql_notes`, but none are directly relevant. The first is `sql_warning`, which is irrelevant because it controls whether single-row `INSERT` statements produce an informational string when warnings occur (and this manual was not retrieved). The second is `max_error_count`, which the user has already tried. The third is `log_warnings` and `log_error_verbosity`, which control warnings in logs but do not affect warning messages.

**Improvement.** One of the limitations of the current version of ANDROMEDA is its difficulty in distinguishing between similar knobs when retrieving documents. In this case, `sql_notes`, `sql_warning`, `max_error_count`, `log_warnings`, and `log_error_verbosity` can all suppress warnings or errors, but in different contexts. For example, `log_warnings` suppresses warnings in logs, whereas `sql_notes` suppresses warning messages. To address this, we can improve the system by prompting LLMs to analyze the differences between the retrieved knobs and select the most relevant one based on the user’s specific question and the retrieved documentation.

Assume you are a DBA and you are particularly skilled at recommending database parameters.

Below are some metrics changes during workload execution:

The value of `created_tmp_tables` increased from `141,000` to `375,000`. `created_tmp_tables` is The number of internal temporary tables created by the server while executing statements. You can compare the number of internal on-disk temporary tables created to the total number of internal temporary tables created by comparing `created_tmp_disk_tables` and `created_tmp_tables` values. See also Section 8.4.4, Internal Temporary Table Use in MySQL....Each invocation of the SHOW STATUS statement uses an internal temporary table and increments the global `created_tmp_tables` value. `created_tmp_tables` is related to the following parameters: `[tmp_table_size, delay_key_write, innodb_status_output_locks]`.

The value of `select_scan` increased from `208,000` to `558,000`. `select_scan` is The number of joins that did a full scan of the first table. `select_scan` is related to the following parameters: `[tmp_table_size, delay_key_write, innodb_status_output_locks, group_concat_max_len, innodb_status_output_locks, innodb_ft_cache_size, sort_buffer_size, myisam_sort_buffer_size, sql_select_limit]`.

Here are some documentations that you can refer:

1. **Manual:** If you are inserting many rows from the same client at the same time, use INSERT statements with multiple VALUES lists to insert several rows at a time. This is considerably faster (many times faster in some cases) than using separate single-row INSERT statements. If you are adding data to a nonempty table, you can tune the `bulk_insert_buffer_size` variable to make data insertion even faster. See Section 5.1.7, Server System Variables. **Parameters:** `[bulk_insert_buffer_size]`.

Below are some user questions and the corresponding recommended parameters:

1. **Question:** I have mysql database and I want to insert about 40'000 rows into it from PHP code , but my code takes more than 15 minutes to insert the rows, is there any chances to optimize it? where is my problem(PHP code / database design) ? Here are the details:

some code snips

First I thought it is an index problem that slows down the insertion , so I removed all the indexes from "imports" table , but it didn't go faster!!  
is the problem from the database design or from my php code?

Also note that the browser is notifying "waiting for response from the server" for the first 5 minutes then most of the remaining time is notifying "transferring data from server", is this because the response html has more than 40'000 line for the row counter 1:1 &lt;/br&ampgt 2:1 &lt;/br&ampgt .....(declared in the php code)?

please consider I'm very newbie, thanks.

**Parameters:** `[foreign_key_checks, unique_checks]`

2. **Question:** Hello there! I'm new to MySQL and PHP and for the last week or so I'm trying to add 300 Millions rows to a table with 10 columns, the system is very slow and it stops at 200k rows after 4 hours. What I did was to generate a PHP script that will automatically insert the same data to the database 300M times..... Why is it so slow? Why is it that I can not get pass 200K rows? What can I do in order to improve performance? Note: Already tried to change some parameters in my.ini like.....

Example:

```
# Set buffer pool size to 50-80% of your computer's memory
innodb_buffer_pool_size=70M
innodb_additional_mem_pool_size=10M
etc...
```

P3 933MHz, 640MB SDRAM, 30Gb hd

Thanks!

**Parameters:** `[innodb_buffer_pool_size]`

Please answer the following query and recommend parameters:

I use mysql with InnoDB containing a database with three tables:`customer, order, order_detail`. The schemas for the three tables are as follows:

```
CREATE TABLE customer(customer_id INT PRIMARY KEY, first_name VARCHAR(50), ...);
CREATE TABLE order(order_id INT PRIMARY KEY, customer_id INT, ... FOREIGN KEY (customer_id) REFERENCES customer (customer_id));
CREATE TABLE order_detail(order_id INT, customer_id INT, first_name VARCHAR(50), ... PRIMARY KEY (order_id), FOREIGN KEY (customer_id) REFERENCES customer (customer_id));
```

I have two newly generated tables for today: `today_order`, where customers are from the customer table. I need to join it and customer to obtain `today_order_detail`. Then, insert `today_order_detail` into `order_detail`. The specific SQL statement is:

```
INSERT INTO order_detail (order_id, customer_id, first_name, ...)
SELECT o.order_id, o.customer_id, c.first_name, ...
FROM today_order o JOIN customer c ON o.customer_id = c.customer_id;
```

My customer table has over 100,000 rows, and `today_order` has 2,000,000 rows. Executing this insert statement takes several minutes, which feels slow to me. Is there a way to speed it up?

Please recommend parameters occurred in reference questions and manuals. Please output list as `['parameter1', 'parameter2', ...]`

#### (a) Prompt

##### Summary of Suggested Values:

- `tmp_table_size`: 64M – 128M
- `bulk_insert_buffer_size`: 64M – 256M
- `innodb_buffer_pool_size`: 50%–75% of system RAM (e.g., 12G on a 16GB system)
- `foreign_key_checks` : 0 (during bulk inserts)
- `unique_checks` : 0 (during bulk inserts)
- `delay_key_write` : ON
- `innodb_ft_cache_size`: 32M – 128M
- `sort_buffer_size`: 4M – 16M
- `myisam_sort_buffer_size`: 16M – 64M
- `innodb_status_output_locks`: OFF (unless needed)

These values are a starting point and should be adjusted based on your system's performance, available resources, and specific workload needs. Regular monitoring is crucial to ensure the chosen settings provide the desired performance improvements.

#### (b) Response

Figure 12: A good case that user encounters a slow insertion problem, and RAG with manuals, historical questions, and telemetry can effectively resolve the issue.

**Assume you are a DBA and you are particularly skilled at recommending database parameters.**

**Below are some metrics changes during workload execution:**

1. The value of `connections` increased from 33.000 to 92.000. `connections` is The number of connection attempts (successful or not) to the MySQL server. `connections` is related to the following parameters: `[max_connections, max_user_connections, mysqlx_max_connections]`.
2. The value of `created_tmp_tables` increased from 123.000 to 354.000. `created_tmp_tables` is The number of internal temporary tables created by the server while executing statements. You can compare the number of internal on-disk temporary tables created to the total number of internal temporary tables created by comparing `Created_tmp_disk_tables` and `Created_tmp_tables` values. See also Section 08.4.4, Internal Temporary Table Use in MySQL. Each invocation of the SHOW STATUS statement uses an internal temporary table and increments the global `Created_tmp_tables` value. `created_tmp_tables` is related to the following parameters: `[max_connections, max_user_connections, mysqlx_max_connections, tmp_table_size, delay_key_write, innodb_status_output_locks]`. The value of `table_locks_immediate` increased from 128.000 to 244.000. `table_locks_immediate` is The number of times that a request for a table lock could be granted immediately. `table_locks_immediate` is related to the following parameters: `[max_connections, max_user_connections, mysqlx_max_connections, tmp_table_size, delay_key_write, innodb_status_output_locks, innodb_lock_wait_timeout, innodb_file_per_table, div_precision_increment]`.
3. The value of `table_open_cache_hits` increased from 94.000 to 210.000. `table_open_cache_hits` is The number of hits for open tables cache lookups. `table_open_cache_hits` is related to the following parameters: `[max_connections, max_user_connections, mysqlx_max_connections, tmp_table_size, delay_key_write, innodb_status_output_locks, innodb_lock_wait_timeout, innodb_file_per_table, div_precision_increment, max_prepared_stmt_count, table_open_cache, innodb_status_output]`.
4. The value of `select_scan` increased from 183.000 to 530.000. `select_scan` is The number of joins that did a full scan of the first table. `select_scan` is related to the following parameters: `[max_connections, max_user_connections, mysqlx_max_connections, tmp_table_size, delay_key_write, innodb_status_output_locks, innodb_lock_wait_timeout, innodb_file_per_table, div_precision_increment, max_prepared_stmt_count, table_open_cache, innodb_status_output, sort_buffer_size, myisam_sort_buffer_size, sql_select_limit]`.
5. The value of `bytes_sent` increased from 979335.000 to 2854371.000. `bytes_sent` is The number of bytes sent to all clients. `bytes_sent` is related to the following parameters: `[max_connections, max_user_connections, mysqlx_max_connections, tmp_table_size, delay_key_write, innodb_status_output_locks, innodb_lock_wait_timeout, innodb_file_per_table, div_precision_increment, max_prepared_stmt_count, table_open_cache, innodb_status_output, sort_buffer_size, myisam_sort_buffer_size, sql_select_limit, div_precision_increment, sql_select_limit, authentication_ldap_sasl_log_status]`.

**Here are some documentations that you can refer:**

1. **Manual:** System Variable `sql_warnings` Scope Global, Session Dynamic Yes Type Boolean Default Value OFF. Parameters: `['sql_warnings']`;
2. **Manual:** To disable message storage, set `max_error_count` to 0. In this case, `warning_count` still indicates how many warnings occurred, but messages are not stored and cannot be displayed.. Parameters: `['max_error_count']`;
3. **Manual:** Suppression of all `log_warnings` items, achieved with `log_warnings=0`, is achieved with `log_error_verbosity=1` (errors only).. Parameters: `['log_warnings', 'log_error_verbosity']`;

**Below are some user questions and the corresponding recommended parameters:**

1. **Question:** What is happening!! I have just started a MySQL 5.1.45 server running, which is replicating from a MySQL 5.0 server (and will itself replicate onwards to 3 more 5.1 servers), and my error logs are filling up with the following statement”

[Warning] Statement may not be safe to log in statement format

How can I turn this off? I have seen various bugs/posts etc. that talk about changing the `log_warnings` variable, but I have set it to 0, 1, 2, 3 and it makes no difference. Or saying that the query itself needs to be changed (even though it was never a problem before)? I can't change the query because it is to ingrained in a production system. How can I tell MySQL to ignore these warnings before I run out of disk space!!.

Parameters: `['log_warnings']`;

**Below are some corresponding recommended candidate parameters from above information: `['log_error_verbosity', 'sql_warnings', 'log_warnings', 'max_error_count']`.**

**Please answer the following query and recommend parameters:**

Lets say, I have executed a query that triggers some warning messages

`DROP TABLE IF EXISTS '\abcd';`

Is there a way to suppress only warning message that been triggering? I see there is a system variable “`max_error_count`”, changing it to zero may ignore warnings but it would also do all errors/note messages.

**Please recommend parameters occurred in reference questions and manuals. Please output list as `['parameter1', 'parameter2', ...]`,**

(a) **Prompt**

To suppress only warning messages while keeping errors and note messages intact, you can set the parameters as follows:

1. `log_warnings`: Set this to 0 to suppress warnings in the log.

```
sql 
SET GLOBAL log_warnings = 0;
```

2. `log_error_verbosity`: Set this to 1 to log only errors (no warnings or notes).

```
sql 
SET GLOBAL log_error_verbosity = 1;
```

3. `sql_warnings`: Set this to OFF to suppress warning messages for the current session.

```
sql 
SET sql_warnings = OFF;
```

These settings should ensure that warning messages are suppressed while retaining error logging.

(b) **Response**

**Figure 13: A bad case that user wants to suppress only the warning message triggered. ANDROMEDA retrieves manuals of knobs with similar functions but can not directly solve the user's question.**