Evaluating the Quality of AI-Generated Resolutions from Conversational vs Structured Sources: Implications for Enterprise Knowledge Automation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Enterprises increasingly rely on historical data to extract resolutions for problems and to automate knowledge mining. While structured ticketing systems (such as ServiceNow, Freshservice, Zendesk etc.) are well-established sources for resolutions, conversational platforms like Slack also capture valuable knowledge in less formal contexts. This paper proposes a Resolution Extraction System to extract meaningful resolutions for IT support cases from noisy, unstructured conversational platforms like Slack, MS Teams, etc. The paper then compares these AI-generated resolutions extracted from Slack conversations (RES) to AI-generated resolutions extracted from structured ticketing systems (RET). We evaluate six key performance indicators (KPIs) - context relevance, completeness, conciseness, noise, perplexity, and readability across 1,000 samples. Our results reveal systematic differences between structured and conversational sources. The analysis shows that with high-precision filtering, conversational sources can be transformed into a meaningful source of resolutions despite the challenges of building reliable enterprise knowledge systems from noisy data. Slack-based resolutions are more relevant and concise but noisier and less readable, whereas ticket-based resolutions are more structured and easier to interpret. These findings highlight the complementary role of conversational data for enterprise knowledge mining and provide guidance on integrating multiple sources into AI-driven automation for support and resolution.

1 Introduction

Most enterprises use historical information from structured ticketing systems to help solve new user problems that take the form of service desk tickets [15, 1]. Usually, companies focus on finding permanent resolutions, which are a set of instructions that can solve the user problem [11, 9]. Therefore, providing fast and high-quality resolutions can help boost productivity in these enterprises by reducing the mean time to resolution (MTTR). Conversational platforms (e.g., Slack, Webex, MS Teams chats) are another source of resolutions for service desk tickets. The aim of this study is to understand whether Slack conversations from specific channels could be a meaningful source of resolutions by comparing it to resolution from tickets. This matters in an enterprise context because it enables discovery of untapped knowledge and provides a foundation for scalable automated resolution extraction.

Slack, despite its informal and often noisy nature, contains rich, actionable content that can be effectively mined using AI-based resolution extraction and with the right pre-processing, the quality of resolutions extracted from Slack matches that of structured tickets. This study is directly motivated by the challenges of reliable ML under imperfect data conditions [8]. Unlike structured ticketing

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

systems, Slack conversations mix relevant resolutions with tangential discussions, emojis, and informal shorthand. This heterogeneity poses a major challenge for enterprise knowledge automation. Therefore, we propose and evaluate a resolution extraction pipeline that adapts large language models to transform Slack conversational data into reliable structured resolutions. This paper makes the following novel contributions:

- A system that transforms noisy, multi-turn enterprise chat threads into validated and structured resolutions by identifying high-quality conversations.
- Resolution-focused extraction from multi-turn enterprise chat threads, not just general summarization.
- Quality Scoring of resolutions using enterprise-specific KPIs.
- Creation of RES as distinct, reusable units in enterprise resolution serving pipelines.

2 Why are Conversational Platforms like Slack considered unreliable?

Conversations are often fragmented and noisy as participants often join or leave them, and important decisions are buried among reactions. Moreover, shorthand or emojis replace explicit statements. Unlike ticketing systems, which enforce schema, conversational threads lack standardized markers of "problem," "root cause," or "resolution," making downstream extraction ambiguous.

Illustrative examples make this contrast clearer. A Slack conversation might contain:

"I think it was a config problem? → Yeah, changed env var, fixed → :thumbsup:"

Whereas the equivalent ticket explicitly records:

Problem: Service outage. Root Cause: Misconfigured environment variable. Resolution: Updated deployment configuration.

Even advanced dialogue summarization approaches require explicit modeling of discourse relations to reconstruct coherence lost in conversational data [5]. Studies of dialogue annotation further reveal how unclear discourse and fragmented turns lead to inconsistency in identifying even basic features such as addressees [18]. Without additional pre-processing, a resolution extraction pipeline will miss critical context, propagate errors, or introduce low-confidence knowledge into enterprise automation pipelines.

Table 1: Comparison of Conversational vs Ticket platforms for resolution extraction

Aspect	**Conversational Platforms (like Slack)**	**Ticketing Platforms (like ServiceNow)**
Structure	Informal, free-form, conversational; includes abbreviations, emojis, and incomplete sentences	Structured, standardized fields for problem description, resolution, status, priority
Information Fragmentation	Resolution often spread across multiple messages and threads; contributions from multiple users	Centralized; single authoritative resolution authored or approved by responsible agent
Mandatory Fields	No enforced title, description and resolution field; updates are inconsistent or optional	Field like title, description and resolution are often mandatory. Additionally. closure criteria is enforced
Noise	High Noise due to off-topic chatter, jokes, outdated suggestions	Low Noise: chronological updates focused on resolution
Ownership	Unclear; multiple users contribute; difficult to identify responsible resolver	Clear ownership - resolution attributed to specific agent or team
Auditability	Limited; messages may be deleted or edited, less queryable	Strong auditability; immutable logs, versioning, queryable records
Status	Slack conversations do not have a status that indicates if the conversation is resolved.	Tickets tend to have Resolved status.

3 Related Work

Prior work shows that noise in dialogue (e.g., off-topic content, interruptions, unexpected phrasing) significantly hampers model performance unless carefully addressed [3]. Some aspects of knowledge mining from conversational platforms have also been explored. Dialogue summarization has been explored both in academic and industrial contexts. Feng et al. proposed a dialogue heterogeneous graph network enhanced with commonsense knowledge to improve summarization quality [4]. Wang et al. introduced Instructive Dialogue Summarization that tailors outputs via query prompts [22]. Other work has focused on extracting structured policies from dialogues in an unsupervised manner [21]. In enterprise settings, solutions such as Microsoft's Conversation Knowledge Mining leverage entity extraction, summarization, and RAG over chat transcripts for knowledge management [13, 14]. Our work extends this line of work by specifically focusing structured resolution extraction from conversational platforms. Moreover, we perform a systematic comparison of RES and RET, and evaluate across six KPIs using statistical rigor.

These characteristics motivate a method that mitigates unreliability to use Slack as a reliable contributor to enterprise resolution systems.

4 Methodology

4.1 What is a Resolution Extraction Model?

A Resolution Extraction Model is an AI system designed to process tickets and raw chats/conversations to extract three key components: *Problem*, *Root Cause* (the underlying reason), and *Resolution*. Both RES and RET will contain a Problem, Root Cause and Resolution.

4.2 Resolution Extraction Model for Conversational Data Sources like Slack

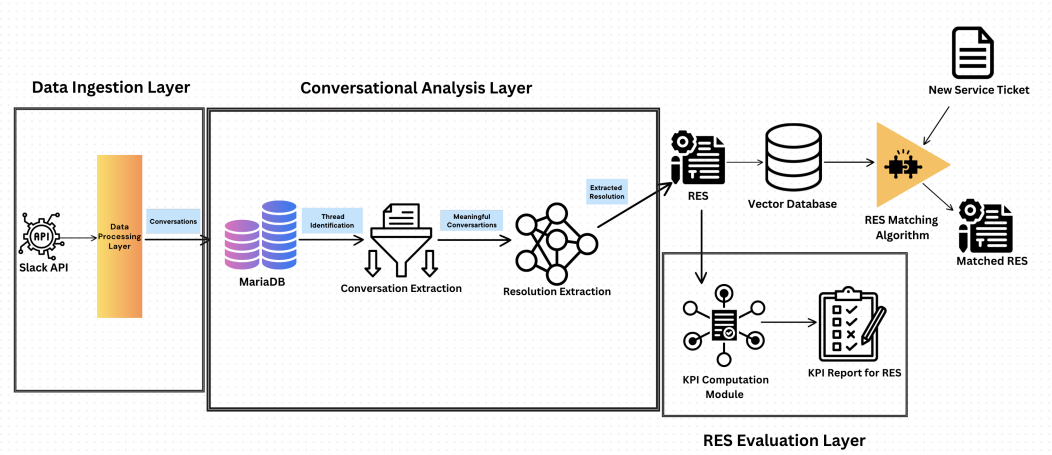

Figure 1: Resolution Extraction Pipeline from Conversational Data Source (Slack)

We propose a pipeline that adapts large language models (LLMs) to extract structured resolutions from Slack, addressing the challenge of unreliability in conversational enterprise data. The workflow is as follows:

1. **Input:** Raw Slack messages from channels via Slack Conversations API.
2. **Preprocessing:**
   - Combine related messages into coherent threads.
   - Remove Slack-specific metadata such markup tags, markdowns etc., replace IDs with names, and handle PII while retaining the textual content.
   - Preserve reactions and emojis for sentiment cues.

- Handle shorthand and informal abbreviations.

3. **Thread identification**:
   - Identify Slack messages that belong to one thread and order them chronologically to make it meaningful.
4. **Generate RES using an LLM :** Generate an RES using GPT-4o. Each RES contains the problem, the root cause and the resolution.
   - The LLM identifies resolution and creates an RES only if a concrete and complete resolution is present in the Slack thread.
   - This step acts as a crucial filtering mechanism. It is important to note that only a subset of raw threads contained a discernible resolution. For this study, we processed an initial corpus of 4500 threads to yield the 500 high-quality RES samples, resulting in a yield rate of approximately 11%. This yield rate is a key finding in itself, underscoring the necessity of the extraction pipeline to isolate actionable knowledge from conversational noise.
   - **RES Quality**: The Quality of RES has been provided in Appendix B and Appendix C.
5. **Postprocessing:** Clean and standardize the extracted outputs.
6. **Indexing:** Store RES in a vector database.
7. **Matching Algorithm:** For a new service desk ticket, find the suitable RES that will resolve the ticket.

4.3 Resolution Extraction Model from Tickets

The RET pipeline follows a similar process as RES pipeline with the exception of certain pre-processing steps that are specific to RES. These pipelines tend to use standard fields such as title, description and comments for resolution extraction. For this study, we processed an initial corpus of 1700 tickets to yield the 500 high-quality RET, resulting in a yield rate of approximately 30%. The yield rate is higher when compared to RES because of the structured nature of tickets. Similar to RES, we have used GPT-4o to extract resolutions from Tickets.

5 Experimental Setup

In our experiment, RET serves as a control, while RES represent extractions from noisy conversational data source. Our experiments test whether the proposed pipeline can extract reliable resolutions from Slack threads, and how closely these outputs approximate the quality of RET across multiple KPIs. We randomly sampled 500 RES and 500 RET outputs. Six KPIs were defined and computed for both groups, then compared using statistical tests. Each KPI is visualized using kernel density plots and analyzed by measures of central tendency. Three different LLMs: GPT-4o, o4-mini (a specialized Chain of Thought model), and Llama 3.3 70B were used to compute the KPIs. Since the initial resolution extraction was performed with GPT-4o, incorporating multiple models for evaluation helped minimize bias and ensured a more balanced and reliable judgment. Llama3.3 was included specifically to ensure that KPIs are also computed from a LLM outside of GPT family of LLMs.

5.1 Hardware and Software requirements

- All analysis was performed using Python libraries: `scipy`, `numpy`, `matplotlib`, `seaborn`, `pandas`, `textstat`, `transformers`, `torch`, and `openai`.
- Compute resources: The experiment runs an AWS instance - `r5.4xlarge`. It also uses `GPT-4o-2025-01-01-preview`, `o4-mini-2024-12-01-preview` hosted on Azure OpenAI, and `Llama-3.3-70B-Instruct` hosted on an AWS instance - `p5en h200`

5.2 KPIs Explained

Four KPIs (Context Relevance, Completeness, Conciseness, Noise) were computed using LLMs. Perplexity and Readability are non-LLM based KPIs and were computed using pretrained model `(EleutherAI/gpt-neo-1.3B)` [2, 6] and standard readability score `(Flesch-Kincaid)` [19, 7] respectively.

- **Context Relevance:** Relevance to the source ticket or Slack conversation. Higher is better.
- **Completeness:** Does the resolution fully address the issue? Higher is better.
- **Conciseness:** How succinct the resolution is without losing meaning. Higher is better.
- **Noise:** Amount of irrelevant or extraneous information. Lower is better.
- **Perplexity:** Measure of textual complexity. Lower is better.
- **Readability:** Ease of reading, favoring structured resolutions. Higher is better.

5.3 Statistical Tests

Slack and Ticket are completely independent data sources which means that the data distributions for RES and RET are also independent. To compare RET and RES distributions for each KPI, we applied statistical significance tests: Mann-Whitney U test [16] and Kolmogorov-Smirnov (KS) test (which does not assume normality) [20]. Furthermore, we also applied statistical tests for effect size measures: Cliff's delta, a robust nonparametric metric that is particularly suitable for nonnormal data [12, 10], and Wasserstein distance [17].

6 Results & Analysis

6.1 Significance Tests

The null hypothesis stated there would be no difference between RET and RES KPIs. For GPT-4o three out of four KPIs, p-values < 0.05 allowed rejection of the null hypothesis, confirming the differences in each KPI distribution. For Conciseness, Mann-Whitney $p = 0.0524$ failed to reject the null hypothesis, while KS test indicated distribution differences.

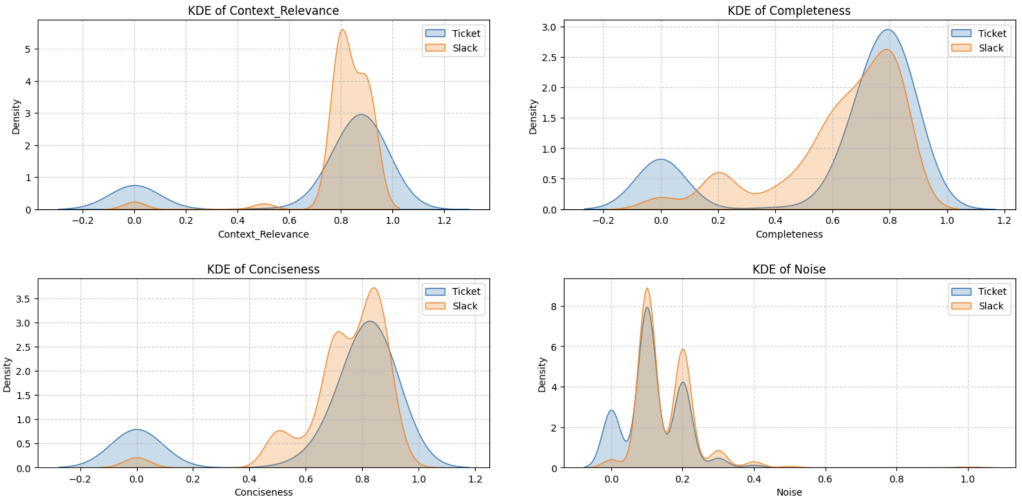

Figure 2: Kernel density plots of KPIs computed by GPT-4o

Extending the analysis to KPIs generated by o4-mini and Llama3.3, we again applied both statistical tests across all six KPIs. For o4-mini, three of the four KPIs produced statistically significant differences between RES and RET ($p < 0.05$), Unlike GPT-4o, both tests failed to reject the null hypothesis for Completeness (Mann-Whitney $p = 0.688$ while the KS test $p = 0.0534$.
For Llama3.3, all four KPIs produced statistically significant differences between RES and RET ($p < 0.05$).

This means that across GPT-4o, o4-mini and Llama3.3, the results consistently indicate that RES and RET differ for the majority of KPI distributions.

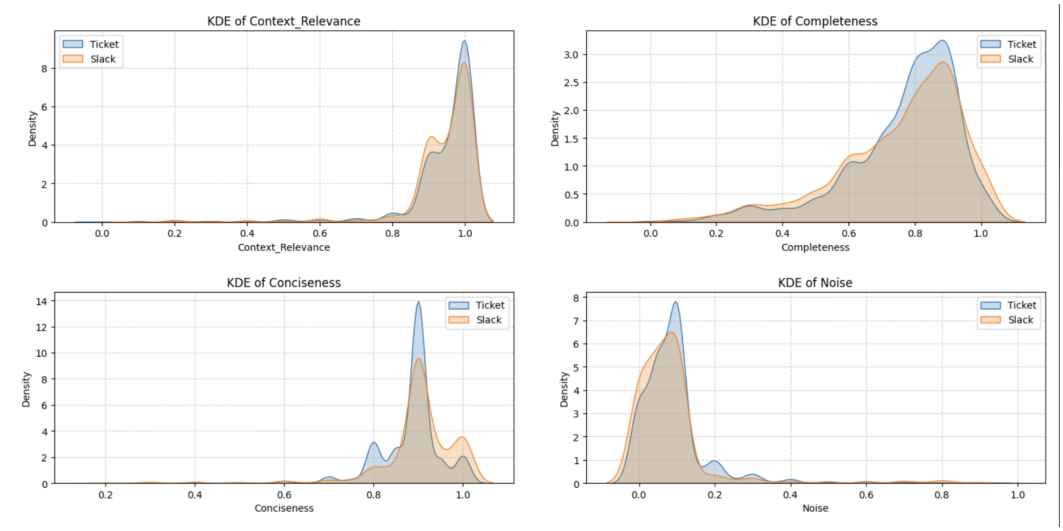

Figure 3: Kernel density plots for KPIs computed by o4-mini

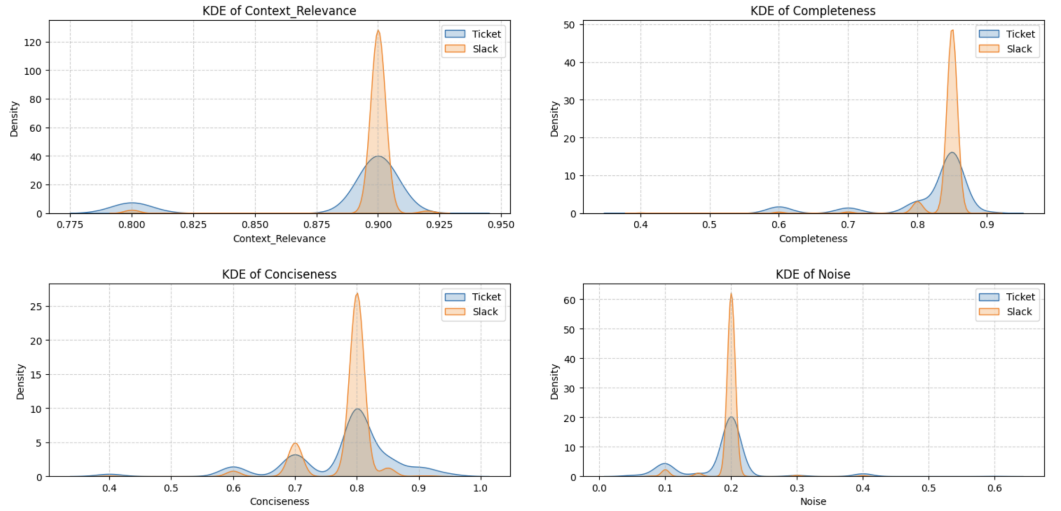

Figure 4: Kernel density plots of KPIs computed by Llama 3.3

6.2 Findings from LLM based KPIs

Table 2: RET vs RES - Qualitative comparison based on mean of the distribution across KPIs

KPI	GPT-4o	o4-mini	Llama3.3
Context Relevance	RES is better	Similar	Similar
Completeness	Similar	Similar	RES is better
Conciseness	RES is better	Similar	Similar
Noise	RET is better	Similar	Similar

As judged by all LLMs, RES has nearly identical or higher Context Relevance and Conciseness scores with GPT-4o highlighting that RES is much better than RET in terms Context Relevance and Conciseness (provided in Appendix C). Cliff's Delta scores and Wasserstein distance) suggest small

effect sizes across all LLM based KPIs, indicating that these differences are present but not large (provided in Appendix B).

6.3 Findings from Non-LLM based KPIs

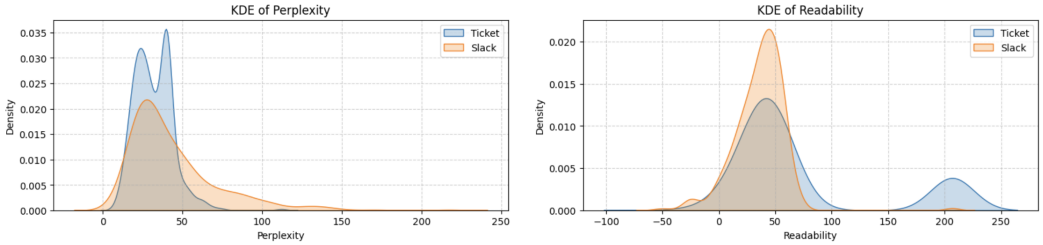

Figure 5: Kernel density plots for Perplexity and Readability for RES and RET

The two KPIs that show the most dramatic differences are Perplexity and Readability. RES has significantly higher perplexity (Wasserstein distance = 11.64), reflecting more complex and less predictable language. RET has substantially higher readability, confirmed by effect size measurements (Cliff's Delta = 0.2732 and Wasserstein distance = 34.7).

This aligns with the structured, formal nature of tickets. Based on the results, we can confirm that Readability and Perplexity are the major separator between RES and RET.

Table 3: Quantitative comparison of RET vs. RES for Perplexity and Readability.

KPI	RET (mean)	RES (mean)	RET (std)	RES (std)
Perplexity	31.98	43.62	11.813	27.86
Readability	70.41	35.71	66.47	22.92

6.4 KPI Interpretations

These findings demonstrate how Slack has unfavorable KPI values for Noise, Perplexity and Readability (probably due to off-topic chatter, informal language). Although differences are statistically significant, not all have strong practical implications. Illustrative examples make this clearer, here's a RES that has low readability and high perplexity because of unusual phrasings and less predictable langugage structure.

***Problem**: Payroll tax calculation updates for Q1 are not applied automatically.*
***Root Cause**: The system requires customers to manually enable the new payroll tax calculation settings.*
***Resolution**:*
1. Inform customers that Q1 payroll tax updates are not applied automatically.
2. Advise customers to manually enable the new calculation settings.
3. Include a note in the release documentation to highlight this requirement and the system's upcoming flagging of discrepancies in old settings.

In contrast to this, here is the RET, which has high readability and low perplexity.

***Problem**: Unable to update user role permissions in Admin Portal.*
***Root Cause**: Mismatch in the role ID mapping for the user in the database.*
***Resolution**:*
1. Identify the affected user and confirm the issue is specific to their profile.
2. Attempt to remove the user from their current role, save the changes, and re-add them with updated permissions.
3. If the issue persists, create a new user profile for the individual and assign the updated permissions

to the new profile as a temporary workaround.
4. Wait for the engineering team to deploy a patch to fix the role ID mapping issue in the database.
5. Once the patch is deployed, revert to the original profile and test the changes.

This shows that Slack's lower Readability and higher Perplexity doesn't make it unusable, it means that:

- Slack is context-rich and, therefore, requires more sophisticated pre-processing steps to remove noise.
- Better techniques for consolidating multiple threads about the same context to extract complete resolutions.

Table 4: Summary of KPI interpretations comparing RES vs RET

KPI	**Interpretation**
Context Relevance	Slack shows better relevance.
Completeness	Nearly identical overall; small statistical difference but not practically significant.
Conciseness	Slack is more concise on average.
Noise	Slack has slightly more noise overall, consistent with conversational style.
Perplexity	Slack exhibits much more varied and complex text.
Readability	Tickets are significantly more readable, confirmed with effect size measures.

7 Limitations

The following are the limitations of this study:

1. The prompt for LLM-based KPI computation are same across the three LLMs (GPT-4o, o4-mini, Llama3.3)
2. What constitutes a "complete resolution" may vary between users, teams, or domains, affecting consistency.
3. Model performance may vary with update to LLMs, making replication challenging if a different version is used.
4. Limited sample size of 500 for both RET and RES.

8 Conclusion

Our results show that a carefully designed resolution-extraction pipeline can yield knowledge artifacts from successfully filtered conversational enterprise data, making it a reliable source for automation. Moreover, the comparative analysis demonstrates that the source of resolutions - Slack or Tickets, has a statistically significant and practically relevant impact on the linguistic and structural characteristics of the extracted resolution.

- Resolutions extracted from Slack (RES) are more relevant and concise but could be less readable. It also suggests that Slack can be a really good source of information for problem resolution. Extracting Resolution from Slack requires a "High precision" extraction system to filter out noise.
- Resolutions extracted from Tickets (RET) are more structured, readable.

Given the high volume and speed of discussions that take place on collaborative platforms like Slack, this opens up promising opportunities for knowledge mining, incident resolution, and building self-service AI assistants. For enterprises, this means that valuable operational knowledge is not limited to structured ticketing systems. It is also embedded in peer-to-peer discussions on Slack (or similar platforms such as MS Teams, Cisco Webex). In doing so, enterprises can reduce MTTR. This study makes the following contributions:

1. It demonstrates how noisy, multi-turn chat threads can be transformed into validated and structured resolutions.
2. It focuses on resolution-extraction rather than general summarization.
3. It introduces quality scoring driven by enterprise-specific KPIs.
4. it defines RES as distinct, reusable units enterprise resolution pipelines.

By systematically quantifying reliability dimensions and validating with statistical tests, this work contributes practical insights into automated resolution extraction systems for noisy, real-world enterprise applications.

References

[1] Bhargav Balakrishnan. Efficient management of it infrastructure implementation and support at enterprise level. *arXiv preprint arXiv:1109.2293*, 2011. Highlights documentation and ticketing systems as key support workflows in enterprise IT infrastructure.

[2] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-TensorFlow, 2021. URL `https://doi.org/10.5281/zenodo.5297715`.

[3] Derek Chen and Zhou Yu. A taxonomy of noise in dialogue and how to address it. *arXiv preprint arXiv:2212.02745*, 2022.

[4] Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. *arXiv preprint arXiv:2010.10044*, 2020.

[5] Prakhar Ganesh, Saket Dingliwal, Rahul Joshi, Manish Shrivastava, Shachi Jat, Karthik Sankaranarayanan, and Dinesh Dey. Abstractive dialogue summarization with discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1835–1845, 2019.

[6] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[7] Y. Guo, W. Qiu, Y. Wang, and T. Cohen. Leveraging large language models for lay summarization. *Journal of Biomedical Informatics*, 32(2):102, 2024. doi: 10.1016/j.jbi.2024.102.

[8] Bo Han. Trustworthy machine learning under imperfect data. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI '24)*, pages 8535–8540, 2024. doi: 10.24963/ijcai.2024/978. Discusses challenges in trustworthy ML under imperfect data (noisy labels, adversarial examples, OOD shifts).

[9] S. Harish, Chetana K. Nayak, and Joy Bose. A scalable and high availability solution for recommending resolutions to problem tickets. *arXiv preprint arXiv:2507.19846*, 2025. Identifies permanent resolutions via machine learning from structured ticket histories.

[10] G. Macbeth, E. Ejzyk, and R. D. Ledesma. Cliff's delta calculator: A non-parametric effect size program for two groups of observations. *Revista Latinoamericana de Psicología*, 42 (2):229–235, 2010. URL `https://www.researchgate.net/publication/262763337_Cliff%27s_Delta_Calculator_A_non-parametric_effect_size_program_for_two_groups_of_observations`.

[11] Atri Mandal, Shivali Agarwal, Nikhil Malhotra, Giriprasad Sridhara, Anupama Ray, and Daivik Swarup. Improving it support by enhancing incident management process with multi-modal analysis. *arXiv preprint arXiv:1908.01351*, 2019. Historical tickets used to recommend resolutions—including permanent ones—via structured knowledge extraction and recommendation systems.

[12] Kane Meissel and Esther S. Yao. Using cliff's delta as a non-parametric effect size measure: An accessible web app and r tutorial. *Practical Assessment, Research, and Evaluation*, 27:Article 10, 2022.

[13] Microsoft. Conversation knowledge mining solution accelerator. `https://github.com/microsoft/Conversation-Knowledge-Mining-Solution-Accelerator`, 2021.

[14] Microsoft. Unlock insights from conversational data. `https://learn.microsoft.com/en-us/azure/architecture/ai-ml/idea/unlock-insights-from-conversational-data`, 2023.

[15] Robert Montgomery and Daniela Damian. What do support analysts know about their customers? on the study and prediction of support ticket escalations in large software organizations. *arXiv preprint arXiv:1901.01092*, 2019. Field study at IBM using over 2.5 million structured support tickets for escalation prediction.

[16] N. Nachar. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4(1):13–20, 2008. URL `https://www.researchgate.net/publication/49619432_The_Mann-Whitney_U_A_Test_for_Assessing_Whether_Two_Independent_Samples_Come_from_the_Same_Distribution`.

[17] A. Ponti, I. Giordani, M. Mistri, A. Candelieri, and F. Archetti. The "unreasonable" effectiveness of the wasserstein distance in analyzing key performance indicators of a network of stores. *Big Data and Cognitive Computing*, 6(4):138, 2022. doi: 10.3390/bdcc6040138. URL `https://www.mdpi.com/2504-2289/6/4/138`.

[18] Dennis Reidsma and Jean Carletta. Contextual analysis of agreement in multi-modal annotations. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, 2008.

[19] L.M. Ripoll Y Schmitz and P. Sonnleitner. Evaluating ai-generated reading comprehension passages: An expert swot analysis and comparative study for an educational large-scale assessment. *Large-scale Assessments in Education*, 13(20), 2025. doi: 10.1186/s40536-025-00255-w.

[20] SpringerLink. Kolmogorov–smirnov test. *Encyclopedia of Educational Research, Measurement, and Evaluation*, 2008. URL `https://link.springer.com/rwe/10.1007/978-0-387-32833-1_214`.

[21] Makesh Narsimhan Sreedhar, Traian Rebedea, and Christopher Parisien. Unsupervised extraction of dialogue policies from conversational data. *arXiv preprint arXiv:2406.15214*, 2024.

[22] Bin Wang, Wayne Xin Zhao, Yuan Wang, et al. Instructive dialogue summarization with query aggregations. *arXiv preprint arXiv:2310.10981*, 2023.

A Glossary of Terms

Table 5: Glossary of key terms.

Term	Definition
RET	Resolution Extracted from Tickets.
RES	Resolution Extracted from Slack.
Problem	The user-reported challenge described.
Root Cause	The underlying technical or operational reason for the issue.
Resolution	The fix, workaround, or action taken to address the problem.
Resolution Extraction Model	AI system that generates structured resolutions from text.

B Mean and Standard Deviations of LLM based KPIs

Table 6: Mean and Standard Deviations of KPIs computed by GPT-4o

KPI	RET Mean	RES Mean	RET Std Dev	RES Std Dev
Context_Relevance	0.7098	0.8152	0.3408	0.1439
Completeness	0.6352	0.6296	0.3081	0.2713
Conciseness	0.6663	0.7400	0.3209	0.1603
Noise	0.1191	0.1527	0.0877	0.0873

Table 7: Mean and Standard Deviations of KPIs computed by o4-mini

KPI	RET Mean	RES Mean	RET Std Dev	RES Std Dev
Context_Relevance	0.9460	0.9402	0.1061	0.1104
Completeness	0.7640	0.7552	0.1713	0.1882
Conciseness	0.8844	0.8985	0.0701	0.0964
Noise	0.0910	0.0841	0.0982	0.1219

Table 8: Mean and Standard Deviations of KPIs computed by Llama3.3

KPI	RET Mean	RES Mean	RET Std Dev	RES Std Dev
Context_Relevance	0.8845	0.8985	0.0361	0.0130
Completeness	0.8146	0.8437	0.0760	0.0322
Conciseness	0.7762	0.7823	0.0920	0.0499
Noise	0.1877	0.1971	.0618	0.0256

C Statistical Tests and Effect Sized for all LLM based KPIs

Table 9: Statistical Tests and Effect Sized for all KPIs computed by GPT-4o

KPI	Mann-Whitney p-value	KS Test	Cliff's Delta	Wessertein Distance
Context_Relevance	0.0091	0.0000	0.0864	0.1434
Completeness	0.0000	0.0000	0.1918	0.0944
Conciseness	0.0523	0.0000	0.0691	0.1129
Noise	0.0000	0.000007	-0.2168	0.0336

Table 10: Statistical Tests and Effect Sized for all KPIs computed by o4-mini

KPI	Mann-Whitney p-value	KS Test	Cliff's Delta	Wessertein Distance
Context_Relevance	0.0027	0.0204	0.0577	0.0078
Completeness	0.6884	0.0534	0.0084	0.0198
Conciseness	0.0000	0.0000	-0.1783	0.0250
Noise	0.0000	0.0000	0.1179	0.0180

Table 11: Statistical Tests and Effect Sized for all KPIs computed by Llama3.3

KPI	Mann-Whitney p-value	KS Test	Cliff's Delta	Wessertein Distance
Context_Relevance	0.0000	0.0000	-0.1461	0.0140
Completeness	0.0000	0.0000	-0.1937	0.0303
Conciseness	0.0107	0.0000	0.0456	0.0332
Noise	0.0000	0.0000	-0.1370	0.0239

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that we compare resolutions extracted from structured tickets (RET) and Slack conversations (RES), evaluate using six KPIs and perform statistical tests. The body presents those experiments and findings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: We have a section on limitations, explicitly noting dataset size, older model for perplexity and use of only one LLM for evaluating KPIs.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA] .

Justification: The paper is empirical; we do not introduce novel theorems or proofs, therefore formal assumptions/proofs are not applicable.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided a smaller synthetic dataset for experiment reproducibility. Additionally, we have shared details of the Resolution Extraction System and have provided KPI definitions, and statistical procedures to enable replication on similar enterprise datasets. But the complete datasets (Tickets and Slack) are proprietary and cannot be released.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Even though the complete enterprise Ticket and Slack datasets are proprietary and cannot be released, we are providing a synthetic dataset. Additionally, we are providing the code for the KPI based evaluation which takes the above mentioned dataset as inputs and provides the KPI analysis. The instructions to run KPI based evaluation is also provided. We cannot share the code for the Resolution Extraction Model since its the IP of the company.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: For this empirical comparison we report the sampling design (500 RET / 500 RES), KPI computation methods, and statistical-analysis pipeline. We do not train new large models, so hyperparameter tables for model training are not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The analysis includes nonparametric significance testing (Mann-Whitney U, KS), effect sizes (Cliff's delta), and Wasserstein distances. These quantify uncertainty and practical significance across KPI distributions.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the software stack (Python, scipy, numpy, pandas, matplotlib, seaborn, azure openai) as well as hardware details (such as AWS instance, GPU type) Since, the work primarily involved LLMs for resolution extraction and statistical analysis of those resolutions, there is no heavy model training involved.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study anonymizes enterprise logs, avoids publishing PII, documents limitations and potential harms, and uses internal compliance processes for data access — consistent with the conference ethics guidance.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses benefits of improved enterprise knowledge automation, faster resolution, broader information capture) and potential harms (data reliability issues, privacy risks), and recommends mitigation (high-precision filtering).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper mentions in experimental reproducibility checklist that data anonymization and internal compliance gating and indicates that proprietary data is not released. While detailed controlled-release protocols are not provided, the paper documents reasonable safeguards taken for this study.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA] .

Justification: License names are not listed and proprietary enterprise data is not publicly redistributed, therefore this question is inapplicable.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA] .

Justification: The paper does not release newly created public assets (e.g., datasets, trained models, or packaged code); the resolution outputs were produced for analysis but are not published as a public asset in the submission.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: The study analyzes existing enterprise artifacts such as Tickets and Slack threads rather than running crowdsourced or controlled human-subject experiments, therefore is inapplicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing, research with human subjects or participant recruitment. Therefore, IRB approvals are not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper explicitly states that a Resolution Extraction Model is powered by GPT-4o. GPT-4o was used to extract Issue/Root Cause/Resolution. Additionally, GPT-4o, o4-mini and Llama3.3 were also used to compute four of the six KPIs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.