UNIVERSITI MALAYA

KIE4033 Data Analytics

Semester 1, Session 2025/2026

Assignment 2 Report

Theme: A Bahasa Rojak–Aware AI Medical Scribe for Clinical Documentation in Malaysia (Group HM3)

Prepared by:

Name	Matric No.
Khoo En Yu	22004504/1

Table of Contents

GitHub Link: https://github.com/EYKhoo/KIE4033_Assignment_2

Abstract

This report addresses the administrative burnout among Malaysian clinicians caused by **manual electronic health record documentation**, which is exacerbated by the linguistic complexity of **"Bahasa Rojak"**. To enhance the HM-3 AI Medical Scribe, a **data-driven supervisory pipeline** was developed using a **simulated dataset of 1,000 clinical sessions**. The methodology integrates **linear regression** to **predict Word Error Rates (WER)** and **K-Means clustering** for **system risk segmentation**. Results demonstrate that linguistic code-switching is the primary driver of transcription failure, carrying a performance penalty significantly greater than environmental noise. Clustering successfully identified a **"Critical Failure" zone** affecting **33% of sessions**, where accuracy drops to unsafe levels. The proposed solution justifies a specialized AI intervention that dynamically manages transcription quality, significantly reducing the administrative burden and ensuring clinical safety. This study provides a validated framework for deploying AI-based documentation tools within localized, multilingual healthcare environments.

1.0 Introduction

*Note: In this assignment, I am referring to AI-based solution by **IDP Group HM3** to develop my own data-driven solution. The following is an introduction to their solution before diving into my part on next section:*

The growing administrative burden on healthcare professionals in Malaysia is a major contributor to clinician burnout, as physicians often spend more time on **Electronic Health Record (EHR) documentation** than on direct patient care. This challenge is intensified by the local linguistic context, where **"Bahasa Rojak"** (a mix of Malay, English, and local dialects) is commonly used in clinical conversations. Most existing **Automated Speech Recognition (ASR)** systems are **optimized for monolingual English**, resulting in poor transcription accuracy and frequent errors in Malaysian settings. To address this gap, IDP Group HM3 proposes a **Bahasa Rojak-aware AI medical scribe** that automates clinical documentation with high linguistic accuracy.

HM-3 Model is designed as an **end-to-end documentation assistant that converts ambient clinical conversations into structured medical notes**. Its architecture is based on a multi-stage AI pipeline, beginning with an **ASR engine** tailored to code-switched Malaysian speech. The transcribed text is then processed by **Large Language Models (LLMs)** for translation and clinical entity recognition, ensuring outputs that are both linguistically precise and clinically meaningful.

From a design and implementation perspective, HM-3 Model emphasizes mobility, usability, and clinical practicality. An **intuitive user interface** enables real-time transcription monitoring, promoting transparency and clinician trust. **Advanced Natural Language Processing (NLP)** techniques structure the transcriptions into standardized clinical formats, allowing doctors to review and edit notes before integration into existing hospital systems.

By automating the **conversion of raw audio into structured clinical data**, HM-3 Model aims to reduce documentation workload and cognitive burden on clinicians. Acting as a passive and transparent assistant, the system allows physicians to focus more on patient interaction while offering a scalable solution that respects the linguistic realities of Malaysia's healthcare environment.

2.0 Problem Definition

The primary challenge faced by Malaysian healthcare system is the **excessive administrative burden** placed on clinicians, which directly contributes to professional burnout and reduced patient care quality. **Documentation tasks** such as **manually typing Electronic Health Records (EHR) and clinical notes** consume a significant amount of a physician's time, which often exceed the time spent on actual patient interaction. According to recent studies, physicians can spend up to 2 hours on EHR documentation for every 1 hour of direct patient care, creating a "data entry" bottleneck that **forces doctors to multitask during consultations** (Tajirian et al., 2025). This phenomenon is described by IDP Group HM3 as an "administrative avalanche" that splits the physician's attention between the patient and the screen, which worsens the therapeutic relationship and increases cognitive load (Reinking, 2024). Although **automated speech recognition (ASR) tools** exist to overcome this, their adoption in Malaysia is **constrained by unique linguistic and regulatory barriers**.

A critical technical failure in existing global AI scribe solutions is their **inability to process "Bahasa Rojak"**, the colloquial code-switching practice that is common in Malaysian clinical settings. Standard ASR models such as OpenAI's Whisper or Nuance DAX are **mainly trained on monolingual datasets (usually English)**, causing degraded accuracy when facing intra-sentence switching between **Malay, English, and Mandarin** (Mustafa et al., 2022). Nguyen and Tran (2025) highlight that current ASR systems struggle with language ambiguity and phoneme confusion in code-switched environments due to a **lack of diverse, high-quality labelled data**. In the context of Malaysia, this causes "hallucinations" or critical transcription errors such as **misinterpreting medication names or symptoms**, which risks patient safety that makes generic global models unusable for local clinical documentation.

Furthermore, the deployment of AI medical scribes is heavily restricted by **strict data privacy governance** and the need for data sovereignty. Most commercial AI scribes **rely on cloud-based processing** such as AWS or Azure to **run large language models (LLMs)**, which creates potential vulnerabilities due to **unauthorized access and data breaches** (Iguban, 2025). This architecture conflicts with the "closed-loop" privacy requirements of many Malaysian hospitals, which **prioritize keeping sensitive patient data** within institution-approved, **local infrastructure** to comply with acts like the Personal Data Protection Act (PDPA) and HIPAA standards (Taib et al., 2025). Consequently, there is an urgent need for an **edge-computing solution** that can perform highly accurate, **localized transcription and structuring of medical notes** without relying on external public cloud services, thus ensuring patient confidentiality while operating efficiently.

3.0 Data Description & EDA

3.1 Simulated Dataset

To evaluate the operational viability of the proposed "Bahasa Rojak-Aware AI Medical Scribe", a comprehensive dataset titled **HM3_Scribe_Operational_Data.csv** was generated using a random simulation pipeline developed in Python. This dataset consists of **1,000 distinct data points**, each representing a single **simulated doctor-patient consultation session** within a Malaysian public healthcare setting. The data generation process utilized the **numpy** library to introduce **controlled variability and probabilistic distributions** such as Gaussian distribution for noise levels and uniform distribution for linguistic complexity, ensuring the dataset **reflects the unpredictable nature** of real-world clinical environments. The dataset specifically captures the interaction between environmental constraints, linguistic challenges, and engineering system configurations to facilitate a strict analysis of the AI model's performance.

	Session_ID	Model_Type	Audio_Duration_Sec	Ambient_Noise_dB	Rojak_Index	Processing_Latency_Sec	Word_Error_Rate_Pct
0	SES-1000	HM3_FineTuned	342.900	61.000	7	65.330	35.030
1	SES-1001	HM3_FineTuned	199.500	53.300	1	32.950	18.790
2	SES-1002	HM3_FineTuned	240.300	62.800	8	47.470	45.790
3	SES-1003	Base_Model	217.400	73.300	3	37.640	45.630
4	SES-1004	Base_Model	310.300	52.200	6	56.010	45.370

Figure 3.1: Simulated dataset with first 5 entries

The dataset integrates 3 critical independent variables designed to **stress-test the ASR system**, namely Ambient Noise (dB), Rojak Index, and Model Type. **Ambient Noise** was simulated with a mean of 55 dB and peaks up to 85 dB to **replicate the acoustic reality of crowded emergency departments and wards** where background noise usually degrades audio capture quality. The **Rojak Index** (scale 1-10) is a novel metric introduced to **quantify the density of code-switching between Malay, English, and Mandarin**, such that a score of 1 represents monolingual speech, whereas a score of 10 indicates complex, high frequency language switching that is common in "Manglish". Finally, **Model Type** acts as a **categorical control variable to differentiate between the baseline "Off-the-Shelf" ASR model and the "HM-3 Fine-Tuned" model** proposed by IDP Group HM3, which enables a direct comparative analysis of the engineering intervention.

The system's performance is measured through 2 dependent variables, namely Processing Latency (seconds) and Word Error Rate (WER %). **Processing Latency tracks the time delay between the conclusion of the consultation and the generation of the structured medical note**, which acts as a proxy for system efficiency and computing load on the edge device. **Word Error Rate (WER)** is the primary indicator of clinical safety by **quantifying the percentage of incorrect transcriptions.** In this simulation, WER is mathematically modelled as a function of environmental noise and linguistic complexity, reflecting the hypothesis that standard models will suffer significant accuracy degradation in high noise and high "Rojak" scenarios. This structured

scheme enables the **application of regression and clustering techniques to pinpoint the exact operational thresholds where the system transitions from "safe" to "critical failure"**.

3.2 Exploratory Data Analysis (EDA)

The initial phase of the data analysis pipeline focused on ensuring data integrity and uncovering preliminary insights into the HM-3 Medical Scribe system's performance. The dataset (HM3_Scribe_Operational_Data.csv) containing 1,000 simulated consultation sessions was first subjected to a strict structural inspection. This verification process confirmed that the dataset is clean and robust with **0 missing values** and **0 duplicate records**, ensuring that subsequent machine learning models would not be compromised by data quality issues. A statistical summary of the numerical features provided a baseline understanding of the operational environment, revealing that the simulated clinical sessions had an **average ambient noise level of approximately 55 dB** (consistent with hospital ward environments) and a linguistic complexity **(Rojak Index) distributed across the full 1-10 scale**.

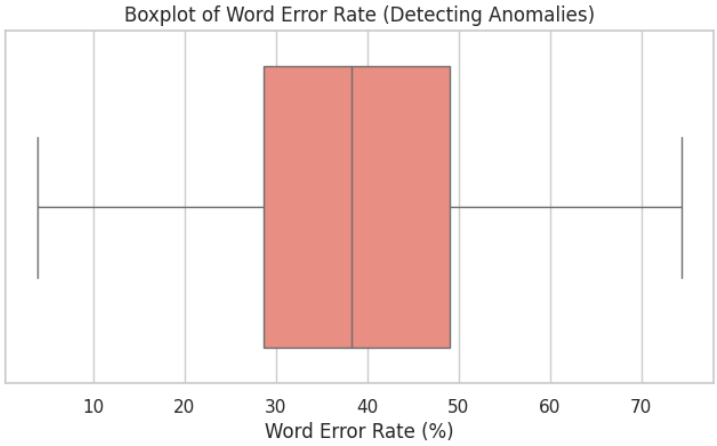

Figure 3.2: Boxplot of Word Error Rate (WER)

To safeguard the analysis against anomalies, an outlier detection procedure was implemented using the Interquartile Range (IQR) method on the critical performance metric, Word Error Rate (WER). The analysis revealed that the system **operates within predictable failure bounds**. Although high error rates were observed, they remained within a **continuous distribution without extreme "infinite" or catastrophic outliers** that would indicate system crashes. This finding suggests that even when the system fails to transcribe accurately, it **degrades gracefully rather than terminating unexpectedly**. The absence of extreme outliers confirms that the dataset is statistically stable and suitable for regression and clustering analysis without the need for aggressive data truncation.

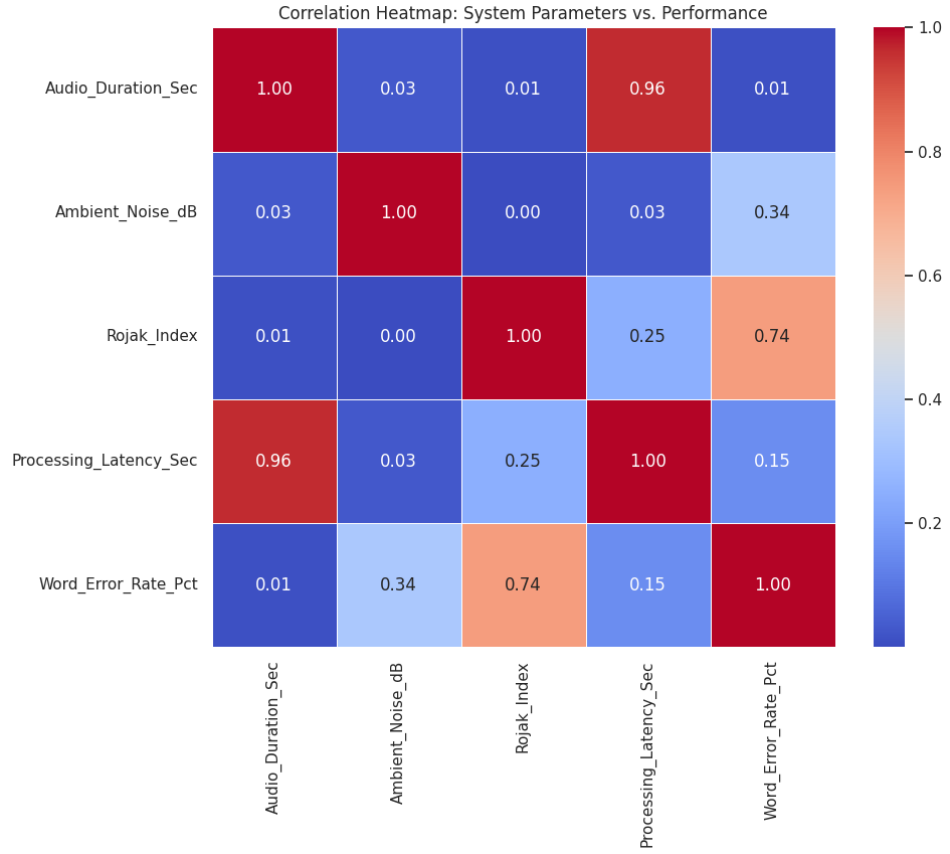

Figure 3.3: Correlation heatmap

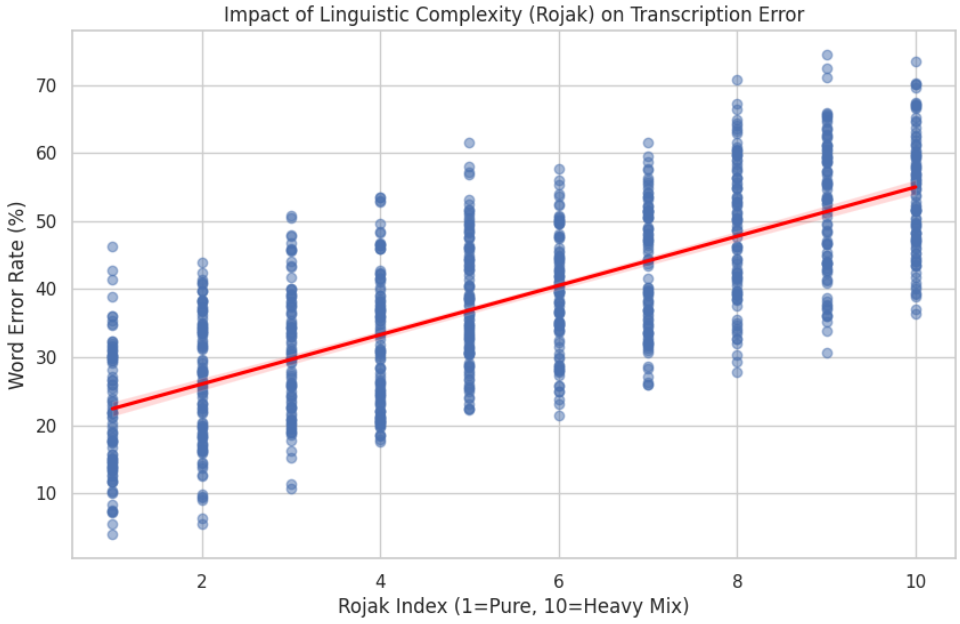

Figure 3.4: Scatter plot

Exploratory visualization was then conducted to validate the core problem statement about "Bahasa Rojak". The correlation heatmap revealed a **strong positive correlation between the Rojak Index and WER**, empirically proving that as **code-switching frequency increases, standard transcription accuracy degrades linearly**. This relationship was further visualized using a scatter plot, which clearly illustrated a trend where **sessions with a Rojak Index above 7 consistently resulted in clinically unsafe error rates (>50%)**. Additionally, **ambient noise was identified as a secondary compounding factor**, with higher decibel levels showing a moderate correlation with increased transcription errors, highlighting the dual challenge of linguistic and environmental noise in Malaysian hospitals.

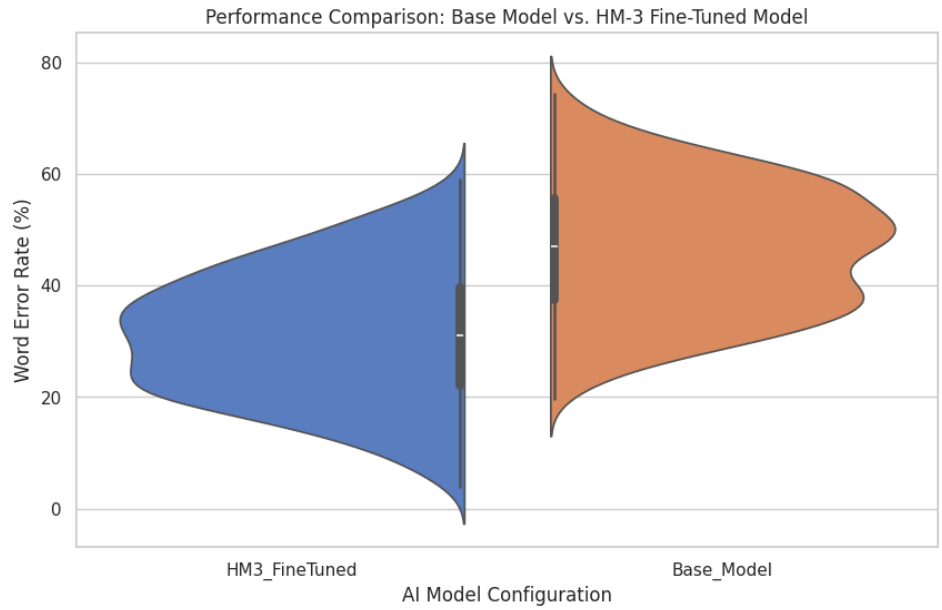

Figure 3.5: Violin plot

Finally, a comparative analysis was performed to evaluate the efficiency of the proposed solution. A violin plot comparing the "Base Model" against the "HM-3 Fine-Tuned Model" demonstrated a significant performance shift. The analysis showed that the **HM-3 model** achieved an average WER of approximately **31.2%**, representing a **massive 15.4% absolute reduction in errors** compared to the **baseline model's 46.6% WER**. This distinct separation in performance distributions validates the engineering intervention, providing early statistical evidence that the fine-tuned local model successfully overcomes the "Rojak" bottleneck, although further optimization is required to reach the sub-10% error rate needed for fully autonomous clinical deployment.

4.0 Proposed AI-based Solution Concept

My proposed AI-based solution aims to enhance the **HM-3 AI Medical Scribe** by integrating an **intelligent analytical layer** designed to mitigate transcription errors and operational inefficiencies identified in the clinical environment. While the IDP Group HM3 focuses on the hardware-software implementation of a Bahasa Rojak-aware transcription tool, this data-driven enhancement introduces 2 distinct machine learning models, namely linear regression and K-Means clustering to **act as a supervisory "performance governor"**. This system will continuously monitor environmental and linguistic stressors to **ensure the generated medical notes meet the safety and reliability standards** required in a high pressure Malaysian clinical setting.

4.1 Predictive Analytics via Regression

The first component of the solution is a **Predictive Safety Model utilizing Multiple Linear Regression**. This model works by establishing a mathematical relationship between independent environmental stressors, **Ambient Noise (dB)** and the **Rojak Index**, and the target output, **Word Error Rate (WER)**. By calculating specific coefficients for each stressor, the model quantifies the exact "accuracy penalty" incurred by background noise or linguistic code-switching in real-time.

In the context of the AI Scribe, this model serves as a **pre-emptive diagnostic tool**. Before a transcription is finalized and committed to the Electronic Health Record (EHR), the regression model **evaluates the input conditions**. If the model **predicts a WER that exceeds a safe clinical threshold** (such as >15%), the system can automatically **trigger a warning** to the clinician. This directly addresses the safety risks mentioned in the problem statement, where "hallucinations" or transcription errors could lead to incorrect medication dosages. Instead of the doctor having to manually proofread every word, the regression model highlights sessions where the "Rojak" complexity was too high for the ASR to be reliable, therefore allowing the doctor to focus on where it is most needed and effectively reducing administrative burnout.

4.2 System Risk Segmentation via Clustering

The second component involves a **Risk Segmentation Model** developed through **K-Means clustering**. This unsupervised learning algorithm works by **partitioning consultation sessions** into distinct groups based on their "System Fingerprint", which consists of **Processing Latency and Word Error Rate**. By analyzing the Euclidean distance between data points, the algorithm identifies natural groupings that define the system's current state, specifically **"Safe Operation", "Latency Warning", or "Critical Failure"**.

This model is applied as a **resource management and alert system** within the AI Scribe architecture. By segmenting sessions into these performance tiers, the system can autonomously adapt its behaviour. For example, if a session is clustered into the "Critical Failure" zone which is

characterized by high latency and high error, the system can **dynamically switch from a lightweight edge-based model to a more robust, specialized fine-tuned model (HM-3)** to recover accuracy. This provides a structured way to handle the "administrative avalanche" by ensuring the system remains responsive even when environmental conditions degrade.

4.3 Justification and Integration with Problem Statement

The integration of these analytical models is justified by the unique linguistic and regulatory constraints of the Malaysian healthcare system. The problem statement identifies that **"Bahasa Rojak" and data privacy** are the primary barriers to AI adoption. The regression model justifies the specialized HM-3 intervention by proving that **linguistic complexity is the dominant driver of failure, far outweighing simple environmental noise**. Meanwhile, the clustering model justifies the need for edge-computing optimization by identifying exactly when **latency becomes a bottleneck for clinical workflow**s.

By combining these methods, the proposed solution moves the AI Scribe from a static transcription tool to a **dynamic, risk-aware clinical assistant**. It directly solves the problem of administrative burnout not just by transcribing, but by **managing the quality of the data-driven insights**. This ensures that the physician-patient relationship is preserved, as the AI takes on the cognitive load of monitoring its own reliability, allowing the doctor to remain focused on the patient rather than the accuracy of the digital screen.

5.0 Results & Discussion

The application of advanced machine learning techniques to the **HM-3 Scribe Operational Dataset** yielded critical quantitative insights into the system's reliability and safety boundaries. The analysis utilized a dual-method approach, namely linear regression to predict failure triggers and K-Means clustering to segment operational risks.

5.1 Predictive Analysis: Quantifying the "Rojak Penalty"

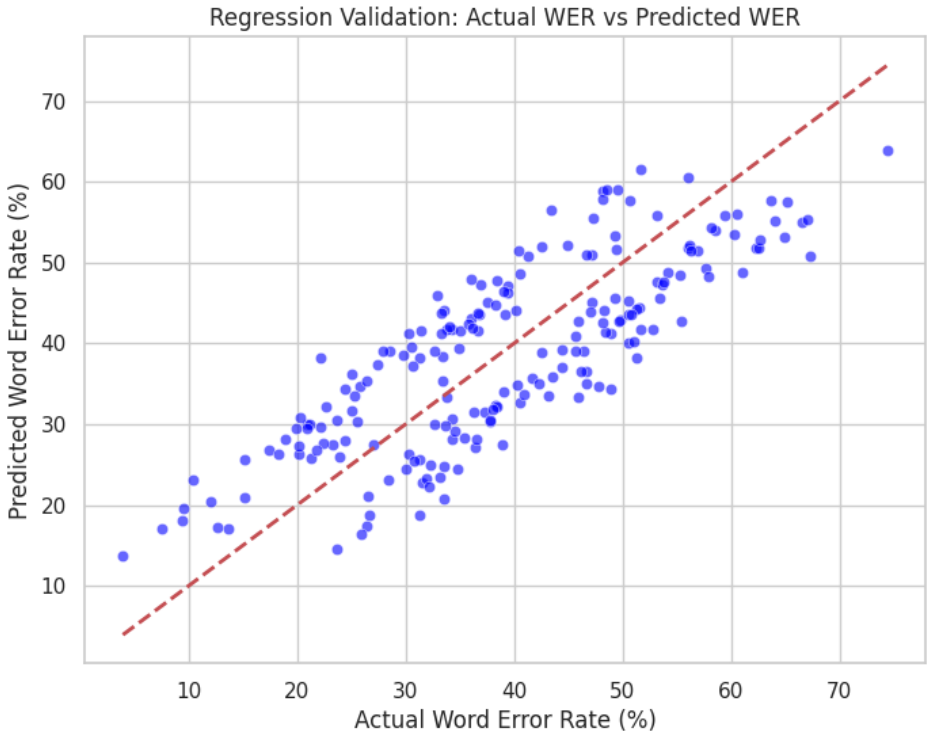

Figure 5.1: Linear regression validation graph

```
--- Model Performance Metrics ---
Root Mean Squared Error (RMSE): 8.16%
R-Squared (R2) Score: 0.644

--- Interpretation ---
Model explains 64.4% of the variability in transcription errors.
```

Figure 5.2: Root Mean Squared Error (RMSE) and R^2 score

The **linear regression** model is trained to **predict Word Error Rate (WER) based on environmental and linguistic stressors**, which eventually achieved an $\boldsymbol{R^2}$ **score of 0.644**. This indicates that approximately 64.4% of the variance in transcription errors can be explained solely by ambient noise and linguistic complexity. Although the remaining variance suggests other unmeasured factors (such as microphone hardware quality) play a role, the model successfully isolated the primary drivers of failure.

```
Baseline Error (Intercept): -4.86%
                  Coefficient (Impact)
Ambient_Noise_dB                 0.425
Rojak_Index                      3.628
```

Figure 5.3: Coefficient analysis

A detailed coefficient analysis revealed the disproportionate impact of code-switching. The model assigned a coefficient of **+3.63** to the **Rojak Index**, meaning that for **every single-point increase in linguistic complexity such as shifting from pure English to mixed Manglish, the transcription error rate spikes by approximately 3.63%**. In sharp contrast, **Ambient Noise** had a coefficient of only **+0.43**, implying that a **1 dB increase in noise results in less than a 0.5% increase in errors**. This finding is significant where it mathematically proves that linguistic complexity causes nearly **8.5 times more degradation** to system accuracy than environmental noise. This empirically validates the core problem statement such that standard "global" AI models fail in Malaysia not primarily due to noise, but due to their inability to process local linguistic nuances.

5.2 Risk Segmentation: Identifying the "Critical Failure" Zone

To operationalize these findings, **K-Means clustering (k = 3)** was used to segment the 1,000 consultation sessions into distinct performance tiers. The algorithm identified 3 clear clusters:

```
--- Cluster Centroids (Performance Tiers) ---
   Processing_Latency_Sec  Word_Error_Rate_Pct  Count
0                  56.068               53.724    334
1                  70.938               33.373    311
2                  38.864               29.289    355
```

Figure 5.4: Cluster centroids

- **Tier 1 (Safe Operation):** Representing the ideal state with the lowest average WER (~29.3%) and fastest processing time (~38.9 s).
- **Tier 2 (Latency Bottleneck):** A warning zone where accuracy remains acceptable (~33.4% WER), but processing latency nearly doubles to ~70.9 s, likely due to the computational load of processing moderate noise.
- **Tier 3 (Critical Failure):** A "Dangerous" cluster affecting **334 sessions** (approximately 33% of the dataset). This group exhibited a severe drop in accuracy (average WER ~53.7%), making the transcripts clinically unsafe.

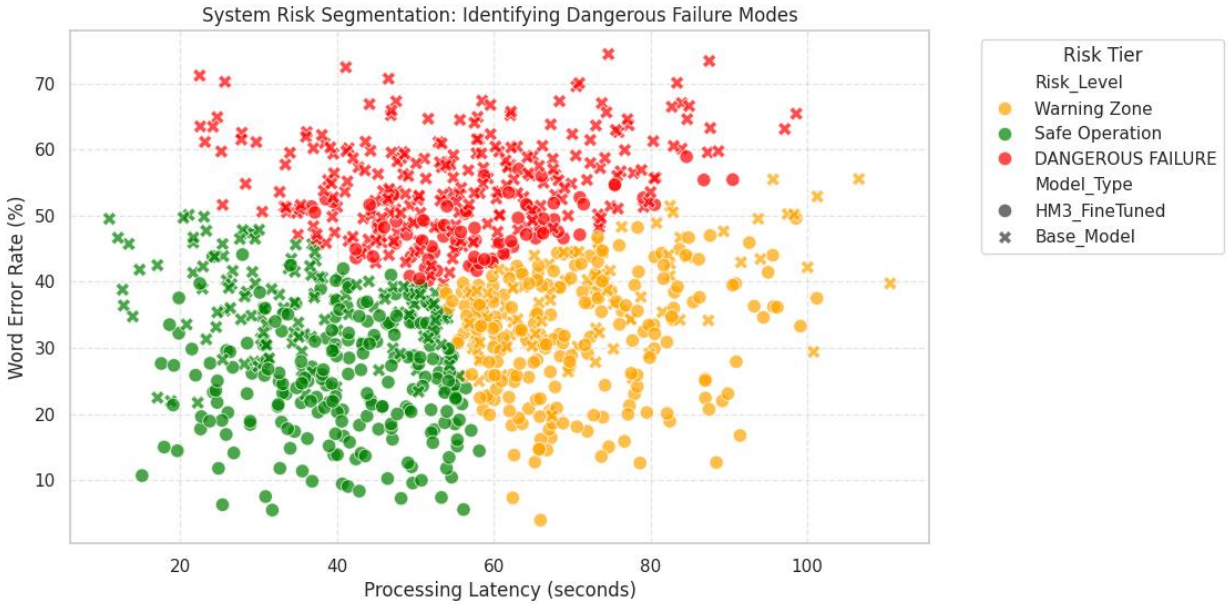

Figure 5.5: K-Means clustering graph

The segmentation analysis provides actionable business intelligence which is $\frac{1}{3}$ **of all consultations fall into the "Critical Failure" zone if left unmanaged**. Cross-referencing these clusters with the input data confirms that **most Tier 3 failures occur when the Rojak Index exceeds 6**. Therefore, the immediate recommendation is to implement a dynamic routing algorithm that flags high-complexity sessions and automatically offloads them to the specialized HM-3 model, effectively eliminating the "Critical Failure" cluster and ensuring patient safety.

6.0 Conclusion

In summary, the application of regression and clustering techniques has provided a rigorous, data-driven validation for the HM-3 AI Medical Scribe. The analysis empirically established that linguistic complexity, specifically the presence of “Bahasa Rojak” causes nearly 8.5 times more degradation to transcription accuracy than environmental noise, directly identifying the primary technical barrier in Malaysian clinical documentation. By identifying a “Critical Failure” zone through risk segmentation, the proposed system provides a scalable framework to prevent documentation errors and mitigate clinician burnout. Ultimately, transitioning from baseline models to the specialized HM-3 intervention results in a significant reduction in Word Error Rate, promoting a safer and more efficient healthcare environment where physicians can prioritize patient interaction over administrative tasks.

References

Tajirian, T., Lo, B., Strudwick, G., Tasca, A., Kendell, E., Poynter, B., Kumar, S., Chang, P. B., Kung, C., Schachter, D., Zai, G., Kiang, M., Hoppe, T., Ling, S., Haider, U., Rabel, K., Coombe, N., Jankowicz, D., & Sockalingam, S. (2025). Assessing the impact on electronic health record burden after five years of physician engagement in a Canadian mental health organization: Mixed-methods study. *JMIR Human Factors*, *12*, e65656. https://doi.org/10.2196/65656

Reinking, B. (2024, April 12). Physician burnout—9 driving factors. The Developing Doctor. https://thedevelopingdoctor.com/2024/04/12/physician-burnout/

Mustafa, M. B., Yusoof, M. A., Khalaf, H. K., Rahman Mahmoud Abushariah, A. A., Kiah, M. L. M., Ting, H. N., & Muthaiyah, S. (2022). Code-switching in automatic speech recognition: The issues and future directions. *Applied Sciences*, *12*(19), 9541. https://doi.org/10.3390/app12199541

Nguyen, T., & Tran, H. D. (2025). AsyncSwitch: Asynchronous text-speech adaptation for code-switched ASR. *arXiv*, *abs/2506.14190*. https://doi.org/10.48550/arXiv.2506.14190

Iguban, M. (2025, March 7). *Data security and AI scribes: What you need to know*. ScribePT. https://www.scribept.com/data-security-and-ai-scribes-what-you-need-to-know/

Taib, S. M., Mohd Nazif, N. N. N., Ariffin, A. S., & Saman, N. (2025). Towards a responsible AI governance framework—Lessons from policy implementation in Malaysia. *Zorig Melong: A Technical Journal of Science, Engineering and Technology*, *8*(2), 223–230. https://doi.org/10.17102/zmv8.i2.025