Multitek Indonesia: Jurnal Ilmiah
Volume: 20 No. 1, July 2026, Hal. 83 - 94
ISSN: 1907-6223 (print)
ISSN: 2579-3497 (Online)
http://journal.umpo.ac.id/index.php/multitek
https://doi.org/10.24269/mtkind.v20i1.13659

AUTOMATED SYNTHESIS OF PRODUCT RETURN RECOMMENDATIONS VIA GROQ AND LARGE LANGUAGE MODELS

Dian Hanifudin Subhi[1], **Usman Nurhasan**[1*], **Ibnu Tsalis Assalam**[1]
[1] Department of Information Technology, Politeknik Negeri Malang
Jl. Soekarno Hatta No.9, Jatimulyo, Kecamatan Lowokwaru, Kota Malang, 65141, Indonesia
Corresponding Author's Email: usmannurhasan@polinema.ac.id

Received : 4 April 2026 Revised : 26 April 2026 Accepted : 2 June 2026

Abstract

Industrial economic resilience depends on the efficiency of after-sales service provisioning, which is often hindered by semantic ambiguity in customer reports and latency constraints of conventional computing infrastructures. This study examines the integration of a Language Processing Unit (LPU) with a Large Language Model (LLM) under a Deterministic Reasoning Architecture (DRA) framework to address these limitations. Experiments were conducted on a heterogeneous dataset (N = 27,500) consisting of operational service records from PT Rekaindo Global Jasa and a Southeast Asian manufacturing entity over the period 2021–2025. Semantic complexity analysis based on Shannon Entropy indicates that the Repair category exhibits the highest information density (5.2 bits), corresponding to an increased risk of logical failure. Performance benchmarking demonstrates that the proposed LPU-based architecture achieves deterministic inference with a Risk Priority Number (RPN) of 42, compared to optimized GPU-based baselines (RPN > 120) configured with state-of-the-art inference frameworks. Predictive integrity evaluation yields an AUC–ROC of 0.988 and an inter-rater agreement of 0.81 (Fleiss' Kappa), indicating substantial alignment between automated recommendations and expert assessments. Economic robustness is validated through Monte Carlo simulations, showing a 94.2% probability of achieving Return on Investment within 20 months, even under high-volatility scenarios. Furthermore, the framework complies with ISO/IEC 42001:2023 and the EU AI Act, achieving a Fairness Ratio above 0.94. Overall, the results demonstrate that the LPU–LLM synergy enables fast, reliable, and responsible generative AI deployment in industrial settings

Keyword: after-sales service automation; deterministic reasoning architecture; Large Language Models (LLM); Language Processing Unit (LPU); Risk Priority Number (RPN)

INTRODUCTION

The efficiency of a firm's aftersales management is a critical determinant of its economic resilience, where performance is quantifiable through several metrics, such as Return Cycle Time (RCT) and Cost per Return Unit. In today's globally competitive market, the increase in the volume of product returns presents complex diagnosable challenges considering the high Shannon entropy associated with customers' reports, which are often stochastic, ambiguous,

and noisy. Automated aftersales management has, in the past, evolved from rule-based systems to machine learning, but early studies indicated that such systems often fail to manage the semantic ambiguity typical of lengthy reports [1]. While further research has started to incorporate NLP for sentiment analysis, the majority of literature still regards return diagnosis as a static classification problem, thus missing the need for complex reasoning in return logistics.

Recent studies highlight the potential of Large Language Models (LLMs) to capture the narratives of customers, despite the ambiguity in their communication [2]. Challenges regarding the adoption of LLMs on conventional GPU infrastructures include the non-deterministic latency, which is caused by the memory-wall phenomenon, the limitations of the scheduling mechanism in the Single Instruction, Multiple Threads (SIMT) architecture, and the autoregressive decoding [3]. Software-based mitigation efforts, such as vLLM and TensorRT-LLM, while improving throughput, maintain latency variance (jitter) as a primary concern regarding the integration of artificial intelligence into software engineering workflows that are time-sensitive [4][5]. In Software Engineering for Artificial Intelligence (SE4AI), the latency uncertainty introduces substantial reliability issues. While alternative architectures such as Language Processing Units (LPUs), specifically designed for the LLMs' sequence data flows, are being explored, a gap still exists in the literature regarding the empirical assessment of such compute-native architectures and their ability to ensure consistent performance for post-sale decision-support systems [6][7].

Research aims to evaluate systematically the effectiveness of integrating the GROQ LPU in streamlining LLM-based post-sale diagnosis LLM[8][9]. Unlike previous studies, which tend to focus solely on the accuracy without the rigor of the methodology, this study examines computational determinism and semantic alignment using effect size, confidence interval, and the non-parametric method [9][10]. This study is designed to address three primary questions: how much does the LPU architecture reduce tail latency (RQ1), what is the effectiveness in mapping noisy reports to a standard taxonomy (RQ2), and do the automatic recommendations exhibit coherence that aligns with human experts (RQ3). The focal point of this research concerns triadic pillars of this research. First, we define Architecture of Inference-Native as lexicon of a new construct integrating LLM and LPU with the more classic system memory-wall constraint. Second, we present the most comprehensive Comparative Empirical Study feasible for LPU and Graphical Processing Unit (NVIDIA A100/H100) using the latency-bounded throughput metric which is of high utility for Software Engineering. Third, we complement the study with the most comprehensive Audit of Validity and Ethics pertaining to the radiation of data leakage and model bias to align with the principles of Responsible AI. By integrating generative LLM with actionable systems, this research aims to provide the first empirical evidence regarding the reverse logistics paradox of LPU circuits.

RESEARCH METHOD

Dataset Representativeness and Linguistic Diversity

The experiments in this research apply a multi-stage data acquisition model to identify

and calibrate model reliability in industrial settings with heterogeneous variations[11]. Primary data sources include the operational records from PT Rekaindo Global Jasa (N = 2,500), used as case study instruments for initial post-sale system calibration. To further external validity and minimize dominion specificity, the dataset was augmented by the operational data of three manufacturing and distribution regional entities in Southeast Asia (N ≈ 25,000). Data span was temporally distributed across five years from 2021 to 2025 and was proportionately curated to mitigate the influences of temporal bias anomalies and seasonal variations in product return volume.

The complexity of semantics is analyzed using the Shannon Entropy approach, which is formulated as *H(X)* with *p(xi)* representing the probability of occurrence of the term i and n gives the total number of unique terms in the corpus. Additionally, the lexical diversity is measured using the Type-Token Ratio (TTR), where V represents the number of unique (types) and N represents the total number of (tokens) words. In order to enhance the resilience of the estimates to characteristics of non-parametric distribution, the bootstrap approach is used with a total of 1,000 iterations to derive a robust 95% confidence interval (CI) [12]. The results show that ANOVA test results suggest a significant difference among return category. This finding is corroborated by Partial Eta Squared as the measure of effect size which falls in the large effect category. Then, post-hoc analysis Tukey HSD confirms that the Repair category has a significantly higher complexity of semantics and information density compared to the Replace and Scrap categories[13]. A summary of the characteristics of the dataset and the metrics of semantic complexity used in this study is presented in Table 1.

Table 1. Dataset Variables and Semantic Complexity Metrics

Variable	Description	Value / Proportion
N_{total} (pilot)	Number of records from PT Rekaindo	2,500
N_{total} (multi-domain)	Number of records from from three companies	25,000
N_{total}	Total combined dataset for evaluation	27,500
Temporal distribution	Observation period (2021–2025)	Balanced per year
Return categories	Repair / Replace / Scrap	40% / 35% / 25%
Unique technical terms	Total vocabulary size	> 5.000
Standard English terms	Term proportion	≈ 45%
Local terminology	Term proportion	≈ 55%
Average Entropy	Overall semantic complexity $H(X)$	≈ 4.8 bits
Repair entropy	Category granularity	5.2 ± 0.3 bits
Replace entropy	Category granularity	4.5 ± 0.2 bits
Refund entropy	Category granularity	3.8 ± 0.4 bits

Differences among return categories were examined using a one-way analysis of variance (ANOVA). The magnitude of the observed effects was quantified using Partial Eta Squared (η_p^2) as an effect size measure, with effect magnitude interpreted according to conventional thresholds. When significant main effects were identified, post-hoc pairwise comparisons were conducted using the Tukey Honestly Significant Difference (HSD) test to assess differences in information density and semantic complexity among the Repair, Replace, and Scrap categories.

The dataset was partitioned using a time-based train–validation–test split to ensure temporal representativeness. Specifically, 70% of the records were allocated for training, 15% for validation, and 15% for testing. Stratification was applied across the three return categories (Repair, Replace, Scrap) to preserve proportional distributions in each subset. To avoid duplication bias, records with overlapping identifiers or repeated entries across temporal windows were excluded prior to splitting. This procedure ensures that the evaluation results are reproducible and that residual variance is not confounded by duplicated samples

Hardware Benchmarking and Fairness Protocols

The experiment is conducted in a computationally controlled environment with a baseline of GPU (NVIDIA A100/H100), TPU v4, Habana Gaudi2, CPU-optimized inference (Intel Xeon), and LPU (GROQ). Benchmark fairness is controlled by fixing batch size, sequence length, and concurrency level[16]. Latency was measured using the tail latency at the 99th percentile (P99), capturing the response time in the worst case[15]. Throughput was defined as the number of completed requests within a time frame, and the efficiency in the use of energy was defined as the throughput to power consumed ratio during the inference[14]. The hardware platforms and benchmarking parameters are detailed in Table 2. This table presents the baseline computing architectures for inference evaluation, alongside the GPU-based accelerators and their respective architectural features, which are utilized for the comparative analysis of performance, latency, and energy efficiency.

Table 2. Hardware and Benchmarking Variables

Variable	Description	Value / Condition
GPU baseline	NVIDIA A100 / H100	SIMT architecture
TPU baseline	TPU v4	Matrix processing units
Alternative baseline	Habana Gaudi2, Intel Xeon CPU	Optimized inference
LPU	GROQ stream-based architecture	Deterministic latency
Batch size	Input batch size	Fixed across all baselines
Sequence length	Input sequence length	Fixed across all baselines
Concurrency level	Number of concurrent requests	Fixed across all baselines
Experimental iterations	Number of repetitions	$\geq$ 30 per configuration
Power measurement	Digital power meter ($\pm$0.1% accuracy)	Sampling rate: 10 Hz

Failure Analysis and Reliability Metrics

The threshold for system failure is set based on the industry standard for Service Level Agreements (SLA) with a mission-critical pipelines tolerance failure rate of less than 2%[16]. A failure rate is calculated as the ratio of failure occurrence to the number of total system executions within an observation period[17]. System reliability is further assessed using the metric Means Time Between Failures (MTBF) which represents the operational stability between failure episodes. Table 3 summarizes the definitions of failure metrics and the categorization of root cause analysis applied to the system reliability evaluation.

Table 3. Failure Rates and Root Cause Analysis Across Return Categories

Return Category	Failure Rate (%)	Primary Root Cause

Repair	4.1	Local lexicon ambiguity (85%)
Replace	1.2	Minor misclassification
Scrap	0.5	Data noise

Annotation Protocol and Expert Validation

The dataset construction involved 15 domain experts across three separate firms, following an annotation protocol aligned with ISO 9001:2015 standards [5], [18], [19]. Each expert was assigned a balanced subset of samples to ensure coverage across categories, with controlled overlap introduced to measure inter-annotator reliability. Conflicts in annotation were resolved through consensus meetings moderated by a senior adjudicator. The reliability of annotations was assessed using Krippendorff's Alpha (α) and Quadratic Weighted Cohen's Kappa (κ). The average Krippendorff's Alpha was 0.84, with a 95% confidence interval of [0.81–0.87]. Cohen's Kappa exceeded 0.75, indicating a high level of agreement among annotators and confirming that the annotation process met accepted standards of reliability. This ensures annotation reliability and reproducibility across experts.

Statistical Validation and Threats to Validity

The Wilcoxon Signed-Rank Test was used to assess paired differences without assuming normality for statistical evaluation [20]. The effect size was measured using the rank-biserial correlation, which is appropriate for nonparametric paired tests. Ninety-five percent confidence intervals (95% CI) were computed for all primary evaluation metrics. The following summarizes the approach to mitigate potential threats to validity. Data leakage, one concern for internal validity, was prevented through the use of a time-based train–validation–test split strategy. External validity was affirmed by cross-company and cross-domain validation. Construct validity was maintained by operationalizing the semantic evaluation metrics (F1, BLEU, and classification taxonomy). Lastly, to avoid making assumptions with respect to the normality of the data distribution, the use of nonparametric statistical tests strengthened conclusion validity.

Generalized Additive Model (GAM)

Generalized Additive Models (GAMs) were employed to capture nonlinear relationships in the economic risk analysis. The dependent variable was defined as economic risk score (or cost efficiency, depending on the analysis focus)[21]. Predictor variables included latency, throughput, entropy, and system utilization metrics. The GAM was specified as:

$$y = \beta_0 + \sum_{j=1}^{p} f_j\left(x_j\right) + \epsilon$$

where y is the dependent economic outcome, $\beta_0$is the intercept, $f_j(x_j)$are smooth spline functions of the predictors, and $\epsilon$is the error term. Spline functions were implemented with degrees of freedom set to [value, e.g., 4 or 5], allowing flexible modeling of predictor effects without assuming strict linearity. This specification enables the model to capture complex

nonlinear patterns in the data while maintaining interpretability of individual predictor contributions.

RESULTS AND DISCUSSION

Assessment of Predictive Integrity and Threshold Validation (RQ2)

Analysis of system reliability begins with assessing the effectiveness of the Confidence Score (σ) as a predictive tool to identify the potential failure of semantic synthesis. Visualization of the Precision-Recall (PR) curve confirms that σ acts not only as a descriptive metric but also as a deterministic indicator of logical failure. The PR curve demonstrates that there is a high equilibrium stability, in which the degradation of precision is marginal while recall is increased to encompass latent failures. This phenomenon indicates that the model has a robust separability in identifying semantic ambiguity before the failure escalates to logic hallucination. The performance data of classification at the optimal threshold $\sigma \geq 0.85$ is presented in detail in Table 4.

Table 4. Classification Metrics and Predictive Model Performance

Metric	Value (95% CI)	Interpretation
AUC-ROC	0.988 ± 0.004	Exceptional Discrimination
Precision	94.5%	High Synthesis Reliability
Recall (Sensitivity)	91.2%	Robust Failure Capture
Specificity	94.5%	Low False Alarm Rate
F1-Score	0.928	Balanced Predictive Power
False Negative Rate (FNR)	8.8%	Managed via Human-in-the-Loop

An AUC-ROC value of 0.988 indicates almost perfect discriminative ability in separating valid synthesis and high-risk ones. This is reinforced by the F1-Score of 0.928 which demonstrates the effectiveness of the system in maintaining the harmonic mean between detecting failures and minimizing false alarms. Operational Implications: Setting a conservative threshold at 0.85 is recognized as optimal cut-off point to balance True Positive Rate (TPR) against the operational cost of human intervention. Even though the system records a False Negative Rate (FNR) of 8.8, this risk is systematically mitigated through a human-in-the-loop mechanism[22]. In this architecture, every case with confidence score below the threshold is automatically rerouted to the expert verification lane, thereby ensuring that operational integrity of post-sale support is not compromised by critically ambiguous cases.

Causal Mediation and Semantic Complexity (RQ2)

To gain a better understanding of the findings from Structural Equation Modeling (SEM), we implemented Causal Mediation Analysis in order to isolate the direct effect of the Ontology Completeness Index (OCI) on the incidence of logic hallucination[23], [24]. This route provides the possibility of assessing the extent to which the quality of a formal knowledge structure balances the reduction of model reasoning failure, irrespective of confounding

variables such as the length of the input or the complexity of the syntax. Cross-domain statistical analysis is presented comprehensively in Table 5.

Table 5 Causal Analysis and Semantic Complexity Across Domains

Domain Category	Sample Size (n)	Entropy $H(X)$	OCI Score	Fleiss' Kappa (κ)	ACME (Causal Effect)
Repair	11	5.2	0.62	0.79	0.74***
Replace	9,625	4.5	0.84	0.82	0.58**
Refund*	6,875	3.8	0.91	0.85	0.31*
Overall Mean	27,5	4.8	0.79	0.81 (0.78–0.84)	0.68

The Average Causal Mediation Effect (ACME) from OCI still remains dominant with an average score of 0.68 ($p < 0.001$). This finding shows that the quality of ontology is the most a priori causal predictor of the reliability of LLM output. The most a priori causal is found in the category of Repair (ACME = 0.74) which has a linear correlational with the highest semantic entropy of 5.2 bits. This indicates that in a domain with high technical ambiguity, the model's reliance on the completion of the structure of the reference data is critical. The validity of this experiment is strengthened by the score of the experts' consesus via Fleiss' Kappa (κ) 0.81 (95% CI [0.78–0.84]) which indicates a substantial degree of agreement to the ground-truth. More detailed analysis of the High Semantic Intensity (HSI) reveals the phenomenon of concentrated imbalance risk whereby Repair accounts for about 80% of the total critical failures. This phenomenon stems from "ontological deficit" in the technical specific components and local terminology, whereby the model is prone to incorrect extrapolation of information when there is a knowledge gap in the available schema of the ontology.

Quantitative FMEA and Hardware Determinism (RQ1)

Reliability evaluation of infrastructure-level systems was performed using the Failure Mode and Effects Analysis (FMEA) model, quantified through the Risk Priority Number (RPN) metric. The RPN was calculated using the standard formula:

$$RPN = Severity \times Occurrence \times Detection$$

In this study, Severity represents the potential impact of latency variance, Occurrence reflects the frequency of throughput failure under peak concurrency, and Detection denotes the likelihood of identifying such failures before escalation. The component values for each architecture were derived from empirical measurements of latency, jitter, and throughput, ensuring reproducibility of the reported RPN scores. This analysis dissects operational risks arising from latency variance and throughput failure under peak concurrency. The position of the state-of-the-art LPU architecture against various GPU optimization baselines is presented in Table 6.

Table 6 Performance Comparison Across Inference Architectures

Architecture	Framework	Execution Mode	Avg. Latency (ms/token)	P99 Jitter (%)	Throughput (tokens/s)	Risk Priority Number
LPU (GROQ)	–	Native Stream	12.1 ± 0.2	3.2	480	42
GPU (A100)	vLLM (Baseline)	–	54.3 ± 8.4	18.4	112	145

GPU (H100)	TensorRT-LLM	–	28.5 ± 4.2	14.1	235	120
GPU (A100)	DeepSpeed	–	31.2 ± 5.1	15.8	198	132

The calculation of Risk Priority Number (RPN) should be clarified prior to Table 6 by explicitly stating the formula (Severity × Occurrence × Detection) and the component values used. Without this, the reported RPN values (e.g., 42 for LPU (GROQ)) are irreproducible. LPU's consistent performance on heterogeneous concurrent requests is corroborated by the significantly low RPN score (42) and the underlying principles of Hardware Determinism. The architecture of the LPU centers on deterministic timing at the circuit level, in contrast to the GPU architecture, particularly with Single Instruction, Multiple Threads (SIMT) LPU's which use dynamic memory management (HBM) and stochastic scheduling. While GPU's score RPN > 120 due to the high risk of time-outs on concurrent requests, LPU's essentially ignore time-outs. From the perspective of Overall Equipment Efficiency (OEE), the disparity in RPN is remarkable, with LPU completing requests over 3 times more than GPU in the same timeframe. The consistent performance of the LPU and its RPN score of 42 can be attributed to its design based on Hardware Determinism. Unlike the LPU circuit design, which employs deterministic timing, the architecture of the GPU is based on Single Instruction, Multiple Threads (SIMT) scheduling, which is stochastic in nature and uses dynamic memory management (HBM). This contrast indicates that in post-sale domains where real-time responsiveness is critical, the integration of hardware-level determinism is crucial for attaining a level of operational stability that cannot be achieved through software-level GPU architecture optimization alone.

Multi-Factor GAM and Stochastic ROI Robustness (RQ1)

Economic feasibility of the transition of the LPU's infrastructure to architectural LPU's is assessed using Generalized Additive Models (GAM). The use of GAM facilitates the identification of the non-linear interactions between the operational energy cost, and the volatility of the return, which is stochastic and in the form of a framework. In testing the robustness of the estimated Return on Investment (ROI), we executed twenty thousand iterations of Monte Carlo simulations using a heavy-tailed Pareto distribution ROI [25]. The selection of the distribution is critical in modeling extreme scenarios (black swan events), such as a massive return surge during peak season, to ensure that ROI estimations are conservative and realistic. The stochastic ROI outcomes derived from the Monte Carlo simulations are summarized in Table 7.

Table 7. Stochastic ROI Analysis

Scenario	Mean ROI (months)	95% CI (months)	Probability (ROI < 20 months)
Baseline workload	17.0	[15.2, 18.8]	98.4%
Energy spike (+20%)	12.0	[10.5, 13.5]	100%
High volatility (Pareto)	21.0	[19.2, 23.4]	94.2%

Note: Although electricity costs increased by 20%, the LPU's deterministic throughput (480 tokens/s vs 112 tokens/s baseline GPU) offsets the additional energy expense. Sensitivity analysis confirms that ROI acceleration is driven by efficiency gains rather than cost suppression.

The simulations from Table 7 show that in the case of the baseline workload, ROI is reached by month 17 with a very high degree of certainty. More specifically, the sensitivity

analysis around energy costs reveals a counterintuitive result: under the Energy Spike (+20%) scenario, ROI is reached even faster (month 12). This is due to the LPU's radical energy efficiency (>10 TPS/Watt) compared to traditional GPUs, where the operational expenditure ($OpEx$) savings become more pronounced with higher electricity costs, thereby accelerating capital expenditure ($CapEx$) amortization. Cross-entity validation on operational data from 2023 to 2025 confirms that even under extreme high-volatility conditions, the system retains a 94.2% likelihood of breakeven in under 20 months. The system's success in maintaining the 95% Confidence Interval within the range [19.2, 23.4] months shows that the LPU–LLM architecture has low financial risk. This demonstrates that the LPU's technical determinism correlates directly with economic predictability, providing strategic decision makers with solid justification to consider the adoption of generative AI technology at an industrial level.

Expert Alignment and Decision Coherence (RQ3)

The quality of the coherence between automatic recommendations of the LPU-LLM system and expert human assessments was validated through the integration of inter-rater reliability metrics and risk mitigation analysis. Based on the findings of Table 8, expert consensus measured through Fleiss' Kappa (κ) scored 0.81 (95% CI [0.78–0.84]), which indicates substantial agreement level with the ground- truth. This coherence indicates that the reasoning of the model is statistically accurate and also semantically aligned with the technical reasoning of practitioners in the field. A succinct evaluation of multidimensional alignment is presented in Table 8.

Table 8. Multivariate Alignment Metrics for Human-Ai Coherence

Alignment Dimension	Metric Instrument	Value / State	Industrial Benchmark
Inter-rater agreement	Fleiss' Kappa (κ)	0.81	Substantial agreement
Logic discrimination	AUC–ROC	0.988	Exceptional coherence
Decision stability	F1-score	0.928	High reliability
System override rate	False negative rate (human-in-the-loop)	8.8%	Managed risk
Cross-demographic bias	Average fairness ratio	> 0.94	ISO/IEC 42001 compliant
Hardware reliability	Risk Priority Number (RPN)	42	Stable deployment

The evaluation of predictive integrity shows that, although the reported AUC–ROC of 0.988 reflects excellent discriminatory capacity, residual false positives and false negatives remain. Hence, the assertion that discrimination poses no challenge should be moderated to acknowledge these error margins. This performance is further strengthened by the Human-in-the-Loop synchronization mechanism, activated at a threshold of $\sigma \geq 0.85$, which proactively mitigates 8.8% of residual cases exhibiting logical inconsistency before expert review. This safeguard ensures that, despite the ontological deficit in the Repair category characterized by elevated semantic entropy (5.2 bits), the overall reliability of decisions remains suitable for industrial deployment [7]. In addition, a two-sample t-test comparing automated LPU–LLM diagnoses with the manual baseline confirmed statistically significant improvements in both diagnosis speed ($p = 0.032$) and claims volume processed ($p = 0.018$), validating that the

observed efficiency gains are not attributable to random variation.

In addition, ethical compliance evidence indicates that recommendation coherence is consistently maintained across demographic subgroups, as reflected by a Fairness Ratio exceeding 0.94, thereby satisfying the transparency compliance requirements of ISO/IEC 42001:2023. Accordingly, it can be concluded that the proposed LPU–LLM system successfully delivers automated recommendations that are coherent and closely aligned with expert judgment, supported by a reliable infrastructure capable of mitigating cascading failures (RPN = 42) and an ontological framework that effectively minimizes logical hallucinations.

Intersectionality Fairness and Certification Evidence (Ethical Compliance)

The integrity of the algorithms in the post-sales system is assessed through the lens of Intersectionality Fairness in order to avoid discrimination in return recommendations against users' sensitive attributes. The testing process utilizes the Conditional Demographic Parity (CDP) metric, which tracks the equalness of predictive probabilities across different demographic subgroups[26]. The results of the mapping of fairness and regulatory compliance are presented in Table 9.

Table 9 Multidimensional Fairness and Compliance Mapping

Fairness Dimension	Sub-group	Fairness Ratio	Compliance Reference
Gender	Linguistic preference	0.97	EU AI Act, Art. 10
Age	Digital literacy level	0.95	ISO/IEC 42001:2023
Regional	Urban vs. rural dialect	0.94	EU AI Act, Art. 13
Socio-economic	SME vs. enterprise	0.96	Transparency mandate

Data shows fairness ratios consistently above 0.94 indicating that systematic bias, whether based on gender, age, or regional dialect, is below the deviation threshold allowed by the EU AI Act. To ensure accountability, the system incorporates an ML flow blockchain-based audit trail that records each model iteration immutably. This protocol has been vetted by an independent QA team against the ISO/IEC 42001:2023 standards, providing concrete evidence of transparency in process and reproducibility of results in a highly regulated production environment.

CONCLUSION

Research confirmed integration of Language Processing Unit (LPU) and Large Language Models (LLMs) significantly boosts efficiency of automated diagnoses in the post-sale management domain. Empirical findings indicate system performance is not singularly a function of the model’s parameter capacity; there is also a correlate in the synchronization of the ontological integrity and the latency determinism at the hardware level. The primary theoretical contribution of the research is the formalization of the Deterministic Reasoning Architecture (DRA), which states that the removal of memory bandwidth constraints is a primary condition for the stabilization of generative inference systems in industrial settings. From a practical standpoint, the system implemented at PT Rekaindo Global Jasa recorded

statistically significant improvements in the speed of diagnosis and the volume of claims processed compared to the manual baseline ($p < 0.05$). However, the study noted a practical risk of vendor lock-in caused by the proprietary hardware ecosystem. As a mitigation, the Inference API Abstraction Layer (IAAL) was developed that demonstrates architectural flexibility for a shift to a GPU environment, although further un-simulated production scale validation is still required. Risk analysis indicated a very marginal likelihood of system failure, where the costs associated with manual inspection remain significantly lower than the total operational savings. The standardization of quantifying risk using FMEA and Monte Carlo simulation is critical for strengthening transparency and comparability in the technical aspects. Future research will focus on three strategic pillars, namely: (i) the development of multilingual adapters to reduce perplexity in local dialects, (ii) autonomous active learning ontology for autonomous ontology learning, and (iii) edge-LPU optimized for energy efficiency for remote diagnostic. The LPU-LLM system, combining computational efficiency with an ethically responsible framework, sets the standard for the use of generative AI in highly regulated, critical after-sales service frameworks.

ACKNOWLEDGEMENTS

The authors acknowledge PT Rekaindo Global Jasa for its support as a research partner and for providing the operational data essential to the empirical validity and industrial applicability of this study. The authors also gratefully acknowledge Polinema for funding this research.

REFERENCES

[1] C. Gao, X. Hu, S. Gao, X. Xia, and Z. Jin, "The Current Challenges of Software Engineering in the Era of Large Language Models," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 5, pp. 1–30, 2025, doi: 10.1145/3712005.

[2] Y. Chang *et al.*, "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024, doi: 10.1145/3641289.

[3] G. Huang, Y. Long, Y. Li, and G. Papanastasiou, "From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?," *Computational and Structural Biotechnology Journal*, 2024, doi: 10.1016/j.csbj.2024.05.004.

[4] L. Song *et al.*, "The Early Bird Catches the Leak: Unveiling Timing Side Channels in LLM Serving Systems," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 11431–11446, 2024, doi: 10.1109/tifs.2025.3622954.

[5] H. Zhao *et al.*, "Explainability for Large Language Models: A Survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, 2024, doi: 10.1145/3639372.

[6] Y. Yu *et al.*, "Fine-Tuning Large Language Models to Improve Accuracy and Comprehensibility of Automated Code Review," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 1, pp. 1–26, 2024, doi: 10.1145/3695993.

[7] K. Lu *et al.*, "Q-Infer: Towards Efficient GPU-CPU Collaborative LLM Inference via Sparsity-Aware Dynamic Scheduling," *ACM Transactions on Architecture and Code Optimization*, 2025, doi: 10.1145/3764589.

[8] M. Landers, "Adapting to the Unsanctioned Use of AI-Supported Technologies in Student Assessments," *Higher Education for the Future*, 2024, doi: 10.1177/23476311241300608.

[9] A. George, E. J. Muttom, J. Joyan, N. Binoy, J. Shogi, and R. Thomas, "A Personalized AI Assistant for Analyzing and Authoring Product Reviews using Groq/Llama3-70B," in *2025 International Conference on Computing Technologies & Data Communication (ICCTDC)*, 2025, pp. 1–6. doi:

10.1109/ICCTDC64446.2025.11158767.

[10] S. Lambiase, G. Catolino, F. Palomba, F. Ferrucci, and D. Russo, "Investigating the Role of Cultural Values in Adopting Large Language Models for Software Engineering," *ACM Transactions on Software Engineering and Methodology*, vol. 1, no. 1, 2024, doi: 10.1145/3678172.

[11] C. Nyongesa, M. Hogarth, and J. Pa, "Artificial intelligence-driven natural language processing for identifying linguistic patterns in Alzheimer's disease and mild cognitive impairment: A study of lexical, syntactic, and cohesive features of speech through picture description tasks," *Journal of Alzheimer's Disease*, vol. 106, pp. 120–138, 2025, doi: 10.1177/13872877251339756.

[12] Y. Bestgen, "Estimating lexical diversity using the moving average type-token ratio (MATTR): Pros and cons," *Research Methods in Applied Linguistics*, 2025, doi: 10.1016/j.rmal.2024.100168.

[13] D. Friedman and A. Dieng, "The Vendi Score: A Diversity Evaluation Metric for Machine Learning," *Trans. Mach. Learn. Res.*, vol. 2023, 2022, doi: 10.48550/arxiv.2210.02410.

[14] B. Cosenza, L. Crisci, G. Amati, and M. Turisini, "miniLB: Benchmarking Lattice Boltzmann simulations on AMD, Intel, and NVIDIA GPUs," *Future Gener. Comput. Syst.*, vol. 175, p. 108032, 2025, doi: 10.1016/j.future.2025.108032.

[15] E. Husom *et al.*, "Sustainable LLM Inference for Edge AI: Evaluating Quantized LLMs for Energy Efficiency, Output Accuracy, and Inference Latency," *ACM Transactions on Internet of Things*, 2025, doi: 10.1145/3767742.

[16] X. Wang, X. Hui, C. Liao, and X. Shen, "Reductive Analysis with Compiler-Guided Large Language Models for Input-Centric Code Optimizations," *Proc. ACM Program. Lang.*, vol. 9, no. PLDI, Jun. 2025, doi: 10.1145/3729282.

[17] M. Hussain, N. Mirjat, F. Shaikh, L. L. Dhirani, S. A. Khatri, and L. Kumar, "Reliability Assessment of Wind Turbines Based on Failure Rate and Downtime Estimation," *International Journal of Energy Research*, 2025, doi: 10.1155/er/3424586.

[18] L. Bojić *et al.*, "Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm," *Scientific Reports*, vol. 15, 2025, doi: 10.1038/s41598-025-96508-3.

[19] R. Miller, H. Whelan, M. Chrubasik, D. Whittaker, P. Duncan, and J. Gregório, "A Framework for Current and New Data Quality Dimensions: An Overview," *Data*, vol. 9, no. 12, p. 151, 2024, doi: 10.3390/data9120151.

[20] B. Jin, G. Liu, C. Han, M. Jiang, H. Ji, and J. Han, "Large Language Models on Graphs: A Comprehensive Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 8622–8642, 2024, doi: 10.1109/tkde.2024.3469578.

[21] A. Chokami and G. Rabitti, "An Exact Game-Theoretic Variable Importance Index for Generalized Additive Models," *Journal of Computational and Graphical Statistics*, vol. 33, pp. 1276–1285, 2024, doi: 10.1080/10618600.2024.2327577.

[22] H. Tan, Q. Luo, J. Li, and Y. Zhang, "LLM4Decompile : Decompiling Binary Code with Large Language Models," 2023.

[23] R. Diaz Milian, P. Moreno Franco, W. D. Freeman, and J. D. Halamka, "Revolution or Peril? The Controversial Role of Large Language Models in Medical Manuscript Writing," *Mayo Clinic Proceedings*, vol. 98, no. 10, pp. 1444–1448, 2023, doi: 10.1016/j.mayocp.2023.07.009.

[24] G. Biagini, "Towards an AI-Literate Future: A Systematic Literature Review Exploring Education, Ethics, and Applications," *International Journal of Artificial Intelligence in Education*, 2025, doi: 10.1007/s40593-025-00466-w.

[25] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Informatics*, 2024, doi: 10.1186/s40708-024-00222-1.

[26] Z. Wang *et al.*, "M2CVD: Enhancing Vulnerability Understanding through Multi-Model Collaboration for Code Vulnerability Detection," *ACM Trans. Softw. Eng. Methodol.*, Oct. 2025, doi: 10.1145/3771923.