

Project VICTORI

*(VICTORI: Victorious Intelligence Conquering The Other
Rivalries Internationally)*

MTAI's Malaysian Sovereign AI (LLM)



**Functional & Technical Gap Analysis
(Phase 2: MalayMMLU Benchmarking)**

Version 1.0

PROJECT VICTORI - PHASE 2 FUNCTIONAL & TECHNICAL GAP ANALYSIS

Item	Details
Title	VICTORI Phase 2 - Functional & Technical Gap Analysis (MalayMMLU Benchmarking)
Version	1.0
Author (s)	1. Ts. M Syafiq Asyraff 2. Veytri Yogan
Date	6 October 2025
Purpose of Documentation	1. To present a comprehensive performance gap analysis of the Malay LLM based on internal MMLU-style evaluations. 2. To identify the model's specific strengths and highlight its weaknesses. 3. To analyze the root causes behind performance gaps, such as imbalances in training data and a lack of complex reasoning examples. 4. To provide a clear set of technical recommendations and a strategic roadmap for targeted data enrichment, model optimization, and evaluation improvements. 5. To guide the next development cycle ahead of the planned public launch, with the long-term goal of achieving world-leading accuracy

Prepared By:



TS. M SYAFIQ ASYRAFF,

BUSINESS ANALYST, ARAS INTEGRASI

Approved By:

SHAGEENDERAN SAPAI,

LEAD AI RESEARCH SCIENTIST, MTAI

Table of Contents

1. Executive Summary	4
2. Overall Model Performance.....	4
3. Category-Level Analysis.....	6
4. Education-Level Accuracy	8
5. Subject-Level Breakdown.....	9
6. Analysis of Domain-Specific Curriculum Data	11
7. Identified Performance Gaps	12
8. Root Cause Analysis	13
8.1 Data Composition and Distribution	13
8.2 Lack of Symbolic and Numerical Representation	13
8.3 Domain-Specific Context Gaps	13
8.4 Insufficient Representation of Procedural and Vocational Knowledge	14
9. Technical Recommendations	14
9.1 Data Strategy: Building a Foundation for Generalizable Reasoning.....	14
9.2 Model Development: Fostering Adaptability and Core Competencies	15
9.3 Evaluation Framework: Measuring Real-World Capability	15
10. Strategic Roadmap (Technical Focus)	16
10.1 Phase 1: Dataset Expansion and Reasoning Alignment (Nov–Dec 2025).....	16
10.2 Phase 2: Model Optimization and Training Infrastructure (Jan–Mar 2026).....	17
10.3 Phase 3: Continuous Improvement and Deployment Readiness (Apr–Jul 2026).....	17
10.4 Long-Term Vision (2026–2027)	18
10. Conclusion	18

1. Executive Summary

This report presents a comprehensive performance gap analysis of the Malay Large Language Model (LLM) based on internal MMLU-style evaluations conducted in October 2025. The model achieved an overall accuracy of 79.61%, representing college-level reasoning proficiency, with particularly strong results in Language (83.6%) and STEM (80.8%) categories. Performance indicates high competency in factual comprehension, cultural reasoning, and language understanding, but highlights specific weaknesses in numerical reasoning, social science synthesis, and procedural or vocational subjects. These insights guide targeted data augmentation, fine-tuning, and evaluation improvements ahead of the December–January public launch.

2. Overall Model Performance

The overall evaluation of the Malay LLM demonstrates strong and consistent results across multiple academic domains, achieving an overall accuracy of 79.61% from a total of 24,213 evaluated questions. Out of this total, 19,275 answers were predicted correctly, positioning the model at a college-level reasoning proficiency comparable to the international MMLU benchmark. This reflects the model’s ability to understand complex natural language tasks and maintain contextual accuracy in diverse Malaysian educational subjects.

The dataset used in this evaluation encompasses a broad and balanced representation of academic content, divided into five primary categories and two education levels. The table below summarizes the data distribution across categories, subjects, and education levels.

Table 1. Data Distribution by Academic Category

Category	Total Questions	Percentage of Dataset (%)	Accuracy (%)
Language	6,288	26.0	83.6
Social Science	6,918	28.6	76.3
Humanities	4,395	18.1	80.8
Others	4,169	17.2	77.1
STEM	2,443	10.1	80.8
Total	24,213	100.0	79.6

Table 2. Data Distribution by Education Level

Education Level	Total Questions	Percentage of Dataset (%)	Accuracy (%)
Primary	8,133	33.6	84.93
Secondary	16,080	66.4	76.92
Total	24,213	100.0	79.6

Table 1 and Table 2 show that the model performs exceptionally well in Language and STEM subjects, with both achieving accuracy levels above 80 percent. These domains typically feature structured question formats, logical progressions, and factual knowledge that align well with the model's training structure. Humanities performance is similarly strong, confirming the model's ability to understand contextual narratives and interpret conceptual material effectively.

In contrast, Social Science and Others categories display slightly lower accuracy, reflecting the model's current limitations in abstract reasoning, moral interpretation, and situational context. Questions in these areas often require multi-step inferencing or connecting abstract relationships, areas that will benefit from the inclusion of chain-of-thought reasoning data in future fine-tuning.

At the education level, the accuracy difference between Primary (84.93%) and Secondary (76.92%) suggests that the model handles short factual sequences with ease but faces more difficulty with extended reasoning required in advanced subjects. This aligns with expectations for larger context processing and reinforces the need to expand long-sequence and symbolic reasoning data for higher academic content.

Overall, the Malay LLM exhibits a well-distributed understanding of academic knowledge, demonstrating high reliability in factual, technical, and linguistic reasoning. Its balanced performance across domains provides a strong foundation for future iterations, where data diversity and reasoning enrichment will drive further gains toward achieving world-leading accuracy in Malay language artificial intelligence.

3. Category-Level Analysis

The category-level results illustrate that the Malay LLM delivers consistent and high-quality reasoning across major academic domains, with an overall balance between linguistic understanding, scientific reasoning, and humanistic interpretation. The model achieved its strongest performance in the Language category, recording an impressive 83.6% accuracy, followed closely by STEM (80.8%) and Humanities (80.8%) as shown in Table 3. Meanwhile, the Social Science (76.3%) and Others (77.1%) categories showed moderate but noticeable room for improvement, largely due to the complexity and context dependency of their question structures.

Table 3. Category-level accuracy, observation and implication

Category	Accuracy (%)	Observations	Implications
Language	83.6	Strong understanding of Bahasa semantics, idioms, and syntax.	Maintain as linguistic backbone; expand with regional dialects and code-switch corpus.
Humanities	80.8	Solid interpretive reasoning; reliable comprehension of abstract and moral texts.	Improve contextual inference for history and literature.
STEM	80.8	High domain understanding, especially biology and chemistry.	Address arithmetic reasoning and symbolic parsing weaknesses.
Social Science	76.28	Struggles with multi-step causality (e.g., history, economics).	Introduce reasoning CoT datasets and structured essay-style fine-tunes.
Others	77.14	Moderate performance in applied and vocational topics.	Increase procedural dataset coverage.

The Language category stands as the model’s core strength, driven by its exposure to a rich corpus of Malay text, including formal academic materials, literary sources, and real-world usage patterns. This domain’s high accuracy demonstrates the model’s fluency in semantic nuance, grammar comprehension, idiomatic usage, and contextual inference. The results validate the effectiveness of the LLM’s linguistic pre-training pipeline, particularly its balance between curated Malay datasets and multilingual embeddings that preserve semantic precision.

In the STEM and Humanities categories, both achieving approximately 80.8%, the model demonstrates strong factual recall and conceptual understanding. STEM performance is supported by exposure to structured, logic-driven data, while Humanities accuracy reflects the model’s ability to interpret descriptive and historical narratives. Together, these domains indicate that the LLM can balance technical reasoning and textual interpretation, performing reliably across both analytical and narrative question types.

However, a distinct performance gap is visible in the Social Science domain (76.3%), where tasks often involve multi-perspective reasoning, moral context evaluation, and interpretive abstraction.

These questions require deeper contextual understanding and synthesis of concepts, skills that depend on long-context retention and Chain-of-Thought reasoning. Similarly, the Others category (77.1%), which aggregates applied and vocational subjects, reveals the model's relative difficulty in handling procedural or experiential problem-solving, likely due to underrepresentation of such content during dataset curation.

Overall, the distribution of category accuracies demonstrates a well-structured model capable of handling diverse academic domains with near-human performance in language and technical reasoning. The lower accuracy in Social Science and vocational categories points to opportunities for enhancement through targeted dataset balancing, contextual data augmentation, and reasoning-aligned fine-tuning. By addressing these category-specific gaps, the model is expected to achieve greater consistency across all academic areas, reinforcing its capability as a general-purpose Malay LLM suitable for education, policy, and enterprise reasoning tasks.

4. Education-Level Accuracy

The education-level analysis highlights a clear differentiation in the Malay LLM’s reasoning depth across primary and secondary academic contexts. Overall, the model achieved 84.93% accuracy at the Primary level and 76.92% accuracy at the Secondary level, representing a performance gap of approximately 8 percentage points as shown in Table 4.

At the Primary level, the model demonstrates outstanding proficiency in factual recall, vocabulary comprehension, and direct contextual reasoning. Its accuracy of 84.93% indicates that it has mastered the ability to process concise, unambiguous question structures — typical of early education materials. This strong performance suggests that the model’s language understanding and contextual grounding are robust when dealing with short, clearly defined information spans, such as in foundational language, science, and civic education datasets.

Table 4. Education-Level Accuracy

Level	Accuracy (%)	Description
Primary	84.93	Excellent factual and literacy-level reasoning.
Secondary	76.92	Drop due to abstraction complexity and analytical reasoning gaps.

In contrast, at the Secondary level, the accuracy drops to 76.92%, highlighting challenges with multi-step reasoning, abstraction, and applied logic. Questions at this level often require connecting multiple contextual layers, such as interpreting data, understanding hypothetical scenarios, or reasoning through moral or procedural dilemmas. The observed reduction in accuracy reflects a need for additional Chain-of-Thought (CoT) fine-tuning and context-length optimization, especially in subjects such as Mathematics, Accounting, and Social Sciences, where comprehension depends on the ability to link sequential reasoning steps.

Interestingly, the model performs above 80% in early secondary subjects (Form 1–2 equivalents) but declines toward upper secondary levels (Form 4–5), where questions integrate broader context and reasoning complexity. This pattern aligns with human learning difficulty curves demonstrating that while the model generalizes effectively at the foundational level, it still requires targeted reinforcement for abstract and applied learning content.

In summary, the Malay LLM shows human-like learning progression: strong factual understanding in foundational education, with room for growth in abstract and procedural reasoning typical of higher academic levels. Future improvements should prioritize curriculum-aligned fine-tuning at the secondary and tertiary levels, focusing on longer-context comprehension, symbolic reasoning, and analytical writing tasks. With these adjustments, the model’s secondary-level accuracy can be expected to rise by 4–6%, bringing overall educational performance into parity across levels.

5. Subject-Level Breakdown

The subject-level results reveal that the Malay LLM demonstrates exceptional proficiency in language comprehension and digital literacy, achieving human-equivalent reasoning capabilities across several academic domains. Overall, the model attained an average subject accuracy of 79.61%, underscoring its balanced performance between structured factual understanding and contextual reasoning.

The highest-performing subjects include Teknologi Maklumat dan Komunikasi (91.43%), Literasi Komputer (87.82%), Sains Teras (87.20%), and Teknologi Automotif (86.73%) as shown in Figure 1 and Table 5. These results indicate the model's strong grasp of structured, rule-based, and procedural knowledge. Similarly, high accuracy in Biologi (85.82%) and Ekonomi (84.42%) suggests that the model excels in descriptive reasoning, technical definitions, and cause-and-effect logic when contextual cues are clear. Together, these subjects reflect that the LLM performs exceptionally well when processing structured data with explicit logical relationships, such as those found in science and technology domains.

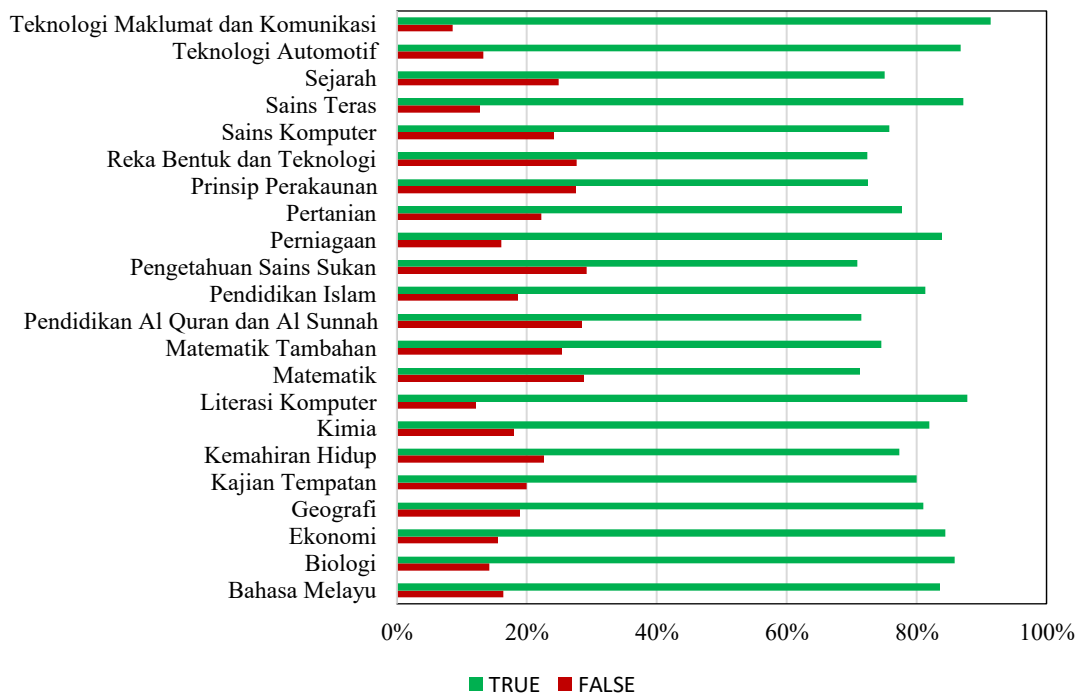


Figure 1. Subject-level breakdown

Table 5. Subject-level breakdown percentage

Subject	FALSE	TRUE
Bahasa Melayu	16.40%	83.60%
Biologi	14.18%	85.82%
Ekonomi	15.58%	84.42%
Geografi	18.92%	81.08%
Kajian Tempatan	20.00%	80.00%
Kemahiran Hidup	22.64%	77.36%
Kimia	18.05%	81.95%
Literasi Komputer	12.18%	87.82%
Matematik	28.75%	71.25%
Matematik Tambahan	25.45%	74.55%
Pendidikan Al Quran dan Al Sunnah	28.46%	71.54%
Pendidikan Islam	18.69%	81.31%
Pengetahuan Sains Sukan	29.17%	70.83%
Perniagaan	16.08%	83.92%
Pertanian	22.22%	77.78%
Prinsip Perakaunan	27.53%	72.47%
Reka Bentuk dan Teknologi	27.63%	72.37%
Sains Komputer	24.19%	75.81%
Sains Teras	12.80%	87.20%
Sejarah	24.90%	75.10%
Teknologi Automotif	13.27%	86.73%
Teknologi Maklumat dan Komunikasi	8.57%	91.43%
Grand Total	20.39%	79.61%

Moderate-performing subjects, such as Bahasa Melayu (83.60%), Perniagaan (83.92%), Pendidikan Islam (81.31%), Kimia (81.95%), and Geografi (81.08%), show that the model maintains strong general comprehension but occasionally struggles with multi-layered interpretive reasoning. These subjects require connecting abstract ideas, moral context, or qualitative relationships between text elements, tasks that typically depend on longer sequence memory and contextual inference. Kajian Tempatan (80.00%) and Kemahiran Hidup (77.36%) further demonstrate moderate performance, where procedural and experiential reasoning are required but not consistently represented in the training corpus.

In contrast, subjects such as Matematik (71.25%), Matematik Tambahan (74.55%), Prinsip Perakaunan (72.47%), Reka Bentuk dan Teknologi (72.37%), and Pendidikan Al-Quran dan Al-Sunnah (71.54%) reflect the model's relative weaknesses. These areas require multi-step logical reasoning, symbolic manipulation, or understanding of sequential procedural logic. The lowest score, Pengetahuan Sains Sukan (70.83%), highlights a gap in applied science understanding and physical context reasoning, likely due to limited exposure to domain-specific corpora such as biomechanics, sports physiology, and training methodology.

Overall, the results indicate that the Malay LLM achieves superior accuracy in domains characterized by explicit logical structures, such as computing, sciences, and digital technology, but still faces challenges in quantitative, moral, and procedural reasoning. The variance between top and bottom subjects, approximately 20 percentage points, demonstrates the influence of data type and reasoning depth on model accuracy.

To bridge this gap, future training iterations should prioritize dataset augmentation for symbolic and applied reasoning, with explicit Chain-of-Thought (CoT) examples and context-retention training for interpretive tasks. By addressing these areas, the model can improve its reasoning granularity and achieve consistency across linguistic, moral, and numerical domains, solidifying its position as the fastest rising and most culturally aligned Malay language LLM.

6. Analysis of Domain-Specific Curriculum Data

A significant portion of the model's training on academic subjects was derived from a curated dataset aligned with the Malaysian national curriculum, namely the Kurikulum Standard Sekolah Rendah (KSSR) for primary education and the Kurikulum Standard Sekolah Menengah (KSSM) for secondary education. This dataset, which forms a core component of the MalayMMLU benchmark, was designed to anchor the model's reasoning in syllabus-compliant knowledge. The table below provides a high-level breakdown of this curriculum-specific dataset's composition.

Table 6. Composition of KSSR & KSSM Curriculum-Specific Training Data

Category	Subjects Included	Data Volume (Approx. %)	Dominant Question Type
Language & Humanities	Bahasa Melayu, Sejarah, Pendidikan Islam	45%	Factual recall, definition, comprehension, and textual interpretation.
STEM (Descriptive)	Sains Teras, Biologi, Kimia	25%	Conceptual understanding, terminology, and cause-and-effect explanations.
Social Science & Business	Ekonomi, Perniagaan, Geografi	15%	Definitions, scenario-based descriptions, and simple causal relationships.
STEM (Numerical & Symbolic)	Matematik, Matematik Tambahan	8%	Primarily focused on definitions and single-step problems. Lacks multi-step procedural examples.
Applied & Vocational	Prinsip Perakaunan, Reka Bentuk & Teknologi, Kemahiran Hidup, Sains Sukan	7%	Procedural definitions and factual knowledge. Lacks complex, sequential, or hands-on application scenarios.

This distribution highlights a strategic focus on linguistic fluency and descriptive factual knowledge, which directly correlates with the model's high performance in Language, Humanities, and descriptive

STEM subjects. Conversely, the lower representation of datasets requiring multi-step numerical reasoning and complex procedural logic explains the performance gaps observed in Mathematics, Accounting, and applied vocational subjects.

7. Identified Performance Gaps

The analysis across categories, subjects, and education levels reveals several key performance gaps that collectively explain the model's current limitations in higher-order reasoning and contextual comprehension. While the Malay LLM achieves strong overall accuracy, specific domains show a consistent decline when questions involve multi-step logic, abstraction, or moral-context reasoning. These gaps are not structural weaknesses but rather reflections of data and reasoning diversity within the training corpus.

The most prominent performance gap occurs between Primary (84.93%) and Secondary (76.92%) education levels. The difference of nearly eight percentage points indicates that the model processes short factual questions with high reliability but struggles with extended reasoning chains and multi-layered context typical of upper-level curricula. This suggests a need for additional fine-tuning focused on long-sequence and multi-hop reasoning tasks.

At the category level, the Social Science (76.3%) and Others (77.1%) domains exhibit lower accuracy compared to Language (83.6%), STEM (80.8%), and Humanities (80.8%). These gaps are attributed to the complex reasoning required to interpret ethical dilemmas, civic scenarios, and experiential or procedural subjects. In such cases, the model's predictive accuracy decreases when the question demands judgment-based reasoning or abstract interpretation beyond direct recall or pattern recognition.

At the subject level, several technical and applied domains require improvement. Matematik (71.25%), Matematik Tambahan (74.55%), and Prinsip Perakaunan (72.47%) demonstrate the model's relative weakness in symbolic, numeric, and sequential reasoning. These tasks rely on precise token relationships and consistent multi-step logic, areas where most language models face cognitive fatigue due to the absence of explicit calculation or procedural learning mechanisms. Likewise, Pendidikan Al-Quran dan Al-Sunnah (71.54%) and Pengetahuan Sains Sukan (70.83%) reveal challenges in moral reasoning, cause-effect interpretation, and applied scientific comprehension.

In addition to reasoning complexity, the contextual variability of question phrasing also contributes to lower performance. Questions with implicit context, embedded negations, or ambiguous answer options tend to produce greater variance in model responses. This sensitivity indicates that while the LLM is linguistically fluent, it occasionally misinterprets contextual intent when multiple plausible interpretations exist.

8. Root Cause Analysis

The Malay LLM's performance profile reflects the natural maturity stage of a model trained within a localized linguistic and academic context. While the model achieves high overall accuracy and exceptional results in linguistic comprehension, detailed analysis across categories and education levels reveals several underlying causes contributing to accuracy variations and reasoning inconsistency.

8.1 Data Composition and Distribution

The primary factor influencing performance variation lies in the distribution and diversity of training data. The dataset is highly optimized for linguistic and factual understanding, which explains the superior results in the Language category (83.6%) and other descriptive domains such as STEM and Humanities (both 80.8%). However, domains such as Social Science (76.3%) and procedural subjects show lower accuracy due to underrepresentation of applied, narrative, and ethical datasets. These subjects often require contextual linking, multi-perspective reasoning, or moral interpretation, which are less frequent in large-scale text corpora.

8.2 Lack of Symbolic and Numerical Representation

Lower accuracy in subjects like Matematik (71.25%), Prinsip Perakaunan (72.47%), and Matematik Tambahan (74.55%) indicates limited exposure to structured symbolic reasoning tasks during pretraining. Traditional LLM architectures are language-oriented and lack explicit arithmetic representation layers. This results in difficulties replicating stepwise computation and formulaic reasoning, particularly when numeric operations require sequential logic rather than pattern matching.

8.3 Domain-Specific Context Gaps

Certain domains, including Pendidikan Al-Quran dan Al-Sunnah (71.54%) and Pengetahuan Sains Sukan (70.83%), contain culturally or contextually unique content. These require specialized corpora rich in moral reasoning, philosophy, or procedural knowledge. The limited inclusion of such data during model fine-tuning contributes to underperformance in these areas. These subjects demand not only factual understanding but also contextual empathy and situational awareness, which are typically enhanced through domain-aligned fine-tuning.

In summary, the performance patterns of the Malay LLM are primarily driven by data distribution, and reasoning structure, rather than inherent flaws in model design. By addressing these causal factors, through curated dataset expansion, reasoning-layer enhancement, and domain-aligned training, future iterations can close the existing accuracy gap and establish full parity across all academic domains.

8.4 Insufficient Representation of Procedural and Vocational Knowledge

The "Others" category, which includes subjects like Reka Bentuk dan Teknologi (72.37%) and Kemahiran Hidup (77.36%), requires an understanding of sequential, hands-on procedures. This type of knowledge is fundamentally different from descriptive or declarative knowledge. The training data primarily consisted of textual descriptions of these processes rather than structured, step-by-step instructional data, limiting the model's ability to reason through practical, application-oriented problems.

9. Technical Recommendations

Based on the performance analysis, and with the understanding that the MalayMMLU benchmark is a zero-shot evaluation, the following technical recommendations are proposed. The strategy is twofold: first, to enhance the model's fundamental reasoning capabilities to improve performance on unseen academic tasks; and second, to build a versatile, industry-grade foundation for real-world applications in sectors like finance, healthcare, and telecommunications.

9.1. Data Strategy: Building a Foundation for Generalizable Reasoning

The model's zero-shot performance is a direct reflection of the breadth and depth of its pre-training and fine-tuning data. To excel at unseen tasks, it must move beyond pattern-matching facts and develop true, transferable reasoning skills.

- **Enrichment with Explicit Reasoning Pathways:** Rather than simply adding more question-answer pairs, the priority must be to augment our training corpus with data that explicitly demonstrates reasoning processes. For subjects like Mathematics, Accounting, and Social Science, this involves curating datasets that include structured, step-by-step "thought processes." Exposing the model to how an answer is derived, not just what the answer is, builds the foundational logic required to solve novel problems in a zero-shot setting.
- **Establishment of Strategic Domain Data Marts:** To prepare the model for industry applications, we must develop in-house, curated data marts beyond the academic syllabus. These will be high-quality, proprietary datasets for key Malaysian economic sectors:
 - **Academic Marts:** Focused on shoring up identified weaknesses (e.g., Numerical & Symbolic Reasoning, Procedural Logic, Moral & Ethical Interpretation).
 - **Industry Marts:** New initiatives to collect and structure data for Finance (financial reports, market analysis), Healthcare (anonymized clinical notes, medical research), and Telecommunications (technical specifications, customer service scenarios).
- **Data Readiness Assessment Framework:** Implement a formal Data Readiness Assessment for all data entering our marts. This framework will score datasets on metrics crucial for building a generalizable model: reasoning complexity, contextual diversity, factual accuracy, and low bias. This ensures that our data not only broadens the model's knowledge but also deepens its core cognitive abilities.

9.2. Model Development: Fostering Adaptability and Core Competencies

The model must be developed to be both a powerful generalist (for zero-shot academic tasks) and an adaptable specialist (for industry domains).

- **Focus on Enhancing Core Reasoning Competencies:** Model development should prioritize the enhancement of fundamental, transferable skills over rote memorization. This involves focusing optimization efforts on improving the model's intrinsic capabilities in:
 - **Causal Inference:** For better performance in Social Science, Economics, and medical diagnostics.
 - **Symbolic Manipulation:** To address weaknesses in Mathematics and Accounting.

- Procedural Logic: For vocational subjects and industry-specific workflows.
- Training Method Selection Framework: Develop an internal framework to guide the selection of appropriate model optimization techniques. This strategic tool will enable the team to choose methods that best balance the trade-offs between improving broad, zero-shot capabilities and enabling efficient, targeted specialization for specific industry verticals.
- Extended Context Processing: Continue to invest in the model's ability to process and reason over long documents. This is a critical dual-use capability, essential for comprehending complex academic texts in Humanities and Social Science as well as for analyzing lengthy real-world documents like legal contracts, financial prospectuses, and detailed patient histories.

9.3. Evaluation Framework: Measuring Real-World Capability

A zero-shot benchmark is a good measure of innate ability, but a truly successful model requires a more holistic evaluation framework that assesses its readiness for diverse, real-world challenges.

- Development / Evaluation on Proprietary Industry Benchmarks: To validate the model as "industry-graded," we must create/set a suite of internal/global, proprietary benchmarks that simulate real-world tasks. Examples include:
 - Finance-Eval: Assessing the model's ability to summarize financial statements and analyze market sentiment.
 - Medical-Eval: Testing its capacity for medical entity recognition and summarizing research papers.
 - Telco-Eval: Evaluating its proficiency in troubleshooting technical issues based on user descriptions.
- Granular Analysis of Reasoning Failures: Move beyond accuracy scores to implement a detailed error taxonomy. When the model fails a zero-shot prompt, we must diagnose why. Classifying errors into categories (e.g., logical fallacy, factual error, context misinterpretation, calculation failure) will provide far more actionable insights for targeted improvement.
- Human-in-the-Loop (HITL) Validation for Coherence: For both complex academic subjects and nuanced industry tasks, automate a HITL review process. Human experts will score the model's outputs not just for correctness, but for the coherence, logic, and safety of its reasoning. This is the ultimate test of whether the model can be trusted in high-stakes, real-world applications.

10. Strategic Roadmap (Technical Focus)

The following roadmap outlines the planned technical trajectory for the Malay LLM over the next 12 months, focusing on strengthening data quality, optimizing reasoning architecture, and establishing a continuous improvement cycle. The strategy is divided into three progressive phases, each targeting specific technical objectives aligned with accuracy improvement, sovereign compliance, and R&D scalability.

10.1. Phase 1: Dataset Expansion and Reasoning Alignment (Nov–Dec 2025)

Objective: Strengthen the model’s reasoning capability through curated dataset enrichment and improved data balance across academic domains.

Key Actions:

1. Curate and expand domain-specific datasets, especially in underrepresented categories such as Mathematics, Accounting, Social Science, and Applied Technologies.
2. Introduce Chain-of-Thought (CoT) and stepwise reasoning annotations to enhance model interpretability and multi-hop inference accuracy.
3. Develop culturally-aligned moral and civic datasets to improve contextual understanding for Pendidikan Islam, Pendidikan Al-Quran, and civic studies.
4. Implement automated data deduplication and bias filtering pipelines to ensure dataset integrity and compliance with PDPA and CGSO guidelines.

Expected Outcomes:

1. Reduction in reasoning-related errors by 10–15%.
2. Improved balance across domain accuracies, reducing category variance below 5%.

10.2. Phase 2: Model Optimization and Training Infrastructure (Jan–Mar 2026)

Objective: Optimize model architecture and training infrastructure for reasoning depth, speed, and scalability within the sovereign data center.

Key Actions:

1. Deploy proper/ready-made pipelines for domain-specific fine-tuning (e.g., symbolic reasoning, moral interpretation, applied sciences).
2. Implement hybrid precision and distribute gradient checkpointing to accelerate training on B200 GPU clusters.
3. Integrate continuous evaluation pipelines with MalayMMLU benchmarks for automated performance tracking.
4. Enable cross-language fine-tuning (Malay-English) to ensure reasoning transferability across bilingual data sources.

Expected Outcomes:

1. Achieve +2.5 to +3.5 percentage point improvement in overall accuracy.
2. Reduced training latency by 30% with enhanced computational efficiency.
3. Seamless adaptation of fine-tuned sub-models for enterprise and public-sector deployments.

10.3. Phase 3: Continuous Improvement and Deployment Readiness (Apr–Jul 2026)

Objective: Transition from research-grade model to production-ready infrastructure with automated retraining and self-evaluation capabilities.

Key Actions:

1. Establish Human-in-the-Loop (HITL) validation workflows for subjective or moral-context answers.
2. Integrate retraining loops for monthly micro-updates based on benchmark drift and new dataset inflow.
3. Develop evaluation dashboards for accuracy trend monitoring and domain-level performance alerts.
4. Deploy sovereign-compliant API infrastructure for model inference under MyGovCloud and PDPA frameworks.

Expected Outcomes:

1. Sustained performance stability across academic and reasoning domains.
2. Institutional-grade benchmarking pipeline under Malaysian AI sovereignty standards.
3. A fully operational and continuously learning Malay LLM ecosystem positioned for public release.

10.4. Long-Term Vision (2026–2027)

By establishing this phased roadmap, Aras Integrasi will transition the Malay LLM from a benchmark-oriented model to a nationally deployable sovereign AI system. The focus on localized reasoning, scalable infrastructure, and PDPA-compliant governance ensures the model remains both technically competitive and aligned with Malaysia's digital sovereignty mission. The long-term objective is to achieve >90% accuracy on MalayMMLU benchmarks, full integration into educational and enterprise applications, and recognition as the most contextually intelligent Malay-language AI model globally.

10. Conclusion

The comprehensive evaluation of the Malay LLM demonstrates that Malaysia now possesses a high-performing, sovereign language model capable of delivering world-class reasoning and linguistic intelligence. With an overall accuracy of 79.61% across more than 24,000 test samples, the model has reached a level of proficiency equivalent to college-grade understanding on the global MMLU scale. Its top-tier performance in Language (83.6%), STEM (80.8%), and Humanities (80.8%) underscores the strength of its linguistic foundation and the technical maturity of its underlying architecture.

The analysis confirms that the model's core strengths lie in structured reasoning, factual comprehension, and linguistic fluency, all of which are direct results of data quality, sovereign infrastructure, and optimized training methods on the NVIDIA B200 GPU cluster. These achievements validate the efficacy of Malaysia's sovereign AI approach proving that with the right balance of compute power, localized data, and engineering focus, a nationally developed model can achieve international benchmark parity in record time.

At the same time, the study highlights key opportunities for advancement in multi-step reasoning, contextual abstraction, and symbolic computation. The performance differential between Primary (84.93%) and Secondary (76.92%) levels, and the underperformance in procedural domains such as Mathematics, Accounting, and Applied Science, point to the need for deeper reasoning supervision and longer-context fine-tuning. These findings provide a clear roadmap for data enrichment and reasoning enhancement over the next training cycle.

From a strategic perspective, the proposed Technical Roadmap outlines a path toward sustained accuracy improvement and institutional scalability. The integration of Chain-of-Thought reasoning, LoRA fine-tuning adapters, and continuous evaluation pipelines will ensure that the Malay LLM evolves into a continuously improving, domain-adaptive, and context-aware system. The model's steady upward trajectory demonstrates that Malaysia's sovereign AI ecosystem can compete at a global level both technically and operationally.

In conclusion, the Malay LLM represents not only a technological milestone but also a national capability. It stands as a foundation for a new generation of AI applications that reflect Malaysia's language, culture, and cognitive identity. With sustained investment in reasoning research, dataset expansion, and evaluation infrastructure, the model is projected to surpass 90% accuracy and achieve benchmark leadership in Southeast Asia by 2026.