Benchmarking Malaysia's Sovereign AI Towards a State-of-the-Art Malay Large Language Model



Abstract

Achieving **79.61% accuracy on the Malay MMLU benchmark**, Malaysia's sovereign Malay Large Language Model (LLM) has unlocked academia-grade reasoning proficiency. This analysis confirms high competency in core domains like Language (83.60%) and STEM (80.80%), particularly in factual and cultural understanding. However, the evaluation strategically highlights clear opportunities for growth. Performance gaps were identified in numerical reasoning, complex social science synthesis, and procedural tasks. This report provides the foundational insights for the next phase of development, guiding targeted data augmentation and fine-tuning to engineer a state-of-the-art model for its upcoming public launch.

1. Introduction

The development of Large Language Models (LLMs) has become a critical measure of a nation's technological progress and digital sovereignty. For Malaysia, the creation of a sovereign AI, specifically a State-of-the-Art Malay Large Language Model is a strategic imperative. Global models often lack the nuance to navigate the complexities of *Bahasa Melayu* and our multicultural social fabric, risking misinterpretation in critical contexts. A sovereign model ensures that our digital infrastructure is built upon, and respectful of, our unique linguistic heritage and national values. This endeavor is not only about technological autonomy but also about ensuring that the digital future reflects Malaysia's unique linguistic, cultural, and academic contexts.

However, the creation of such a foundational model is only the first milestone. To transition from a research asset to a reliable, publicly deployable tool, a rigorous and transparent assessment of its true capabilities is essential. A model's "black box" nature must be illuminated, its strengths must be validated, and its weaknesses must be precisely identified and understood.

This report addresses this critical next phase. It presents a comprehensive functional and technical gap analysis of the Malay LLM, based on its performance against the robust MalayMMLU benchmark.

The purpose of this document is to provide a granular, multi-dimensional diagnostic of the model's reasoning abilities. It moves beyond what the model gets right or wrong to analyze why, by:

1. Deconstructing performance across diverse academic categories, from Language and STEM to Social Sciences and vocational subjects.
2. Analyzing the performance delta between primary and secondary education levels to understand its grasp of reasoning complexity.
3. Conducting a root cause analysis to pinpoint the specific data and training gaps responsible for its limitations.

The insights from this analysis are the foundation for the strategic roadmap presented herein. This report provides the definitive, data-driven evidence needed to guide the next cycle of targeted data enrichment, model fine-tuning, and architectural optimization. It is the critical step in transforming our sovereign model from "proficient" to "state-of-the-art."

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 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	83.60
Social Science	6,918	28.6	76.30
Humanities	4,395	18.1	80.80
Others	4,169	17.2	77.10
STEM	2,443	10.1	80.80
Total	**24,213**	**100.0**	**79.60**

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**	**79.60**

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 (category) 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
Language	**83.60**	Strong understanding of Bahasa semantics, idioms, and syntax.
Humanities	**80.80**	Solid interpretive reasoning; reliable comprehension of abstract and moral texts.
STEM	**80.80**	High domain understanding, especially biology and chemistry.
Social Science	**76.28**	Struggles with multi-step causality (e.g., history, economics).
Others	**77.14**	Moderate performance in applied and vocational topics.

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 training 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.

As the subjects get more difficult, 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 with 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.

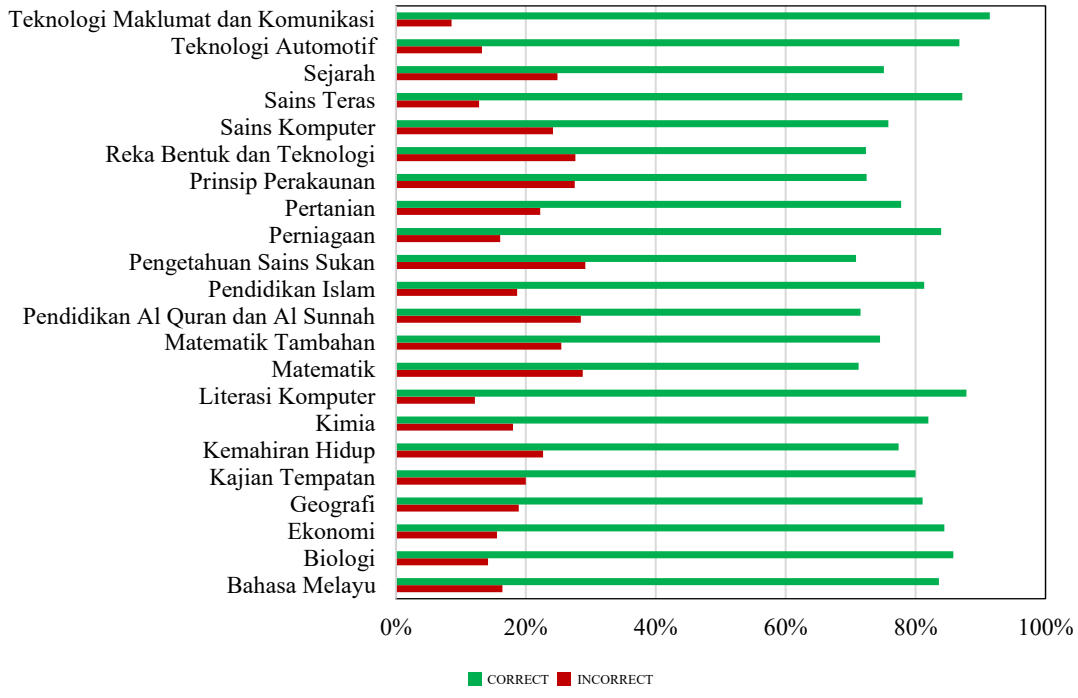

Figure 1. Subject-level breakdown

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	45.00%	Factual recall, definition, comprehension, and textual interpretation.
STEM (Descriptive)	Sains Teras, Biologi, Kimia	25.00%	Conceptual understanding, terminology, and cause-and-effect explanations.
Social Science & Business	Ekonomi, Perniagaan, Geografi	15.00%	Definitions, scenario-based descriptions, and simple causal relationships.
STEM (Numerical & Symbolic)	Matematik, Matematik Tambahan	8.00%	Primarily focused on definitions and single-step problems. Lacks multi-step procedural examples.

Applied & Vocational	Prinsip Perakaunan, Reka Bentuk & Teknologi, Kemahiran Hidup, Sains Sukan	7.00%	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. 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 has unlocked academia-grade reasoning capabilities 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 integration of various training methodologies 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.

Endorsed By:



Shageenderan Sapai

Lead AI Engineer

Doctor of Philosophy (PhD) in AI,

Enabling AI Models for Robotics,



Ts. Muhammad Syafiq Asyraff

Senior AI Business Analyst,

Member of Malaysian Board of Technology

