Adaptive In-Context Learning in Large Language Models: A Zero-Leakage Benchmark Using Fictionalized Extensions of the Khom Thai Script

Aakash Kuragayala, Dr. Rishi Jain

Asian Institute of Technology
Bangkok, Thailand
*Corresponding email: aakashkuragayala40@gmail.com

Abstract: Despite rapid progress in large language models, existing benchmarks often fail to distinguish genuine adaptive reasoning from training data memorization. To address this, we introduce KHOM THAI, a zero-leakage benchmark designed to rigorously test a model's ability to acquire, retain, and flexibly apply novel symbolic rules entirely in-context. Built upon a structured, fictionalized extension of the historical Khom Thai script, the benchmark challenges models with a complex romanized notation system governed by interlocking rule sets. These rules encompass context-sensitive subscript clusters, multi-position vowel attachments, consonant class distinctions, and intricate interaction effects such as rwrap-induced order inversion and class-sensitive consonant doubling ensuring models must master an entirely novel system without relying on pre-trained linguistic priors. The benchmark isolates distinct facets of adaptive learning through three complementary tasks: Task A evaluates rule generalization across varying complexities, Task B tests long-context retention using semantically unrelated distractors, and Task C examines belief updating through targeted feedback. The evaluation of 14 frontier and strong open models reveals substantial performance disparities; while leading models like Gemini 2.5 Pro achieve near-ceiling accuracy (96.4%), many others exhibit sharp performance drops when handling rare cluster interactions and complex vowel modifications. Ultimately, KHOM THAI provides a clean, highly discriminative testbed for evaluating robust, human-like in-context rule learning, offering a valuable signal for measuring genuine abstraction and adaptive intelligence in AI systems.

Keywords: Adaptive In-Context Learning, Zero-Leakage Benchmarking, Artificial General Intelligence (AGI), Cognitive Plasticity**,** LLM Evaluation, Data Contamination, Synthetic Dataset Generation, Procedural Content Generation

1. Introduction

The rapid advancement of large language models (LLMs) has been accompanied by a persistent challenge in reliable evaluation. While models continue to post impressive scores on popular benchmarks, there is growing evidence that these results are often inflated by data contamination the unintentional (or sometimes deliberate) inclusion of evaluation examples or closely related content in pre-training corpora[1], [2], [3]. This phenomenon undermines the validity of

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

benchmark scores and makes it increasingly difficult to distinguish genuine reasoning capabilities from memorized knowledge.

Data contamination has become a central concern in LLM evaluation, as static benchmarks such as MMLU and others have become heavily saturated[4], [5]. When test data or near-identical variants leak into training, reported performance no longer accurately reflects a model's ability to generalize or reason adaptively. This issue is especially problematic for measuring progress toward Artificial General Intelligence (AGI), where the core requirement is not retrieval of known facts but cognitive plasticity the capacity to rapidly acquire, retain, and flexibly apply novel rules and symbolic systems from limited contextual information.

In response to these limitations, researchers have called for new evaluation paradigms that emphasize few-shot or in-context rule learning on previously unseen tasks. A prominent example is the Abstraction and Reasoning Corpus (ARC-AGI) introduced by Chollet [6], which was explicitly designed to measure fluid intelligence and skill-acquisition efficiency rather than crystallized knowledge. However, even dynamic benchmarks face challenges as models improve and evaluation practices evolve.

To address these gaps, we present KHOM THAI, a zero-leakage benchmark specifically engineered to evaluate adaptive in-context learning. The benchmark is built upon a richly structured romanized notation system inspired by the historical Khom Thai script a Brahmic abugida traditionally used for Pali texts in Thailand but substantially extended with fictionalized rules that have virtually no overlap with existing training data.

The system features three consonant classes (mid, high, and low) that dictate notation behavior, context-sensitive subscript clusters with special cases (rwrap, hsub, msub, hh, and ras), multi-position vowel attachments (including diphthongs and two new vowels), and complex interaction rules such as high-class subscript modifiers, order inversion for above vowels under rwrap, automatic vowel lengthening for low-class initials, and class-sensitive consonant doubling during medial /a/ omission.

These interlocking rules create a challenging, hierarchical, and exception-rich symbolic domain. Through synthetic dataset generation and procedural content generation, we ensure that every test instance requires genuine on-the-fly rule internalization rather than pattern matching against prior exposure.

KHOM THAI consists of three complementary tasks that isolate distinct facets of adaptive intelligence:

- Task A (Rule Generalization): Models receive a complete rule specification in a single prompt and must apply it correctly to 500 unseen syllables across five tiers of increasing complexity.

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

- Task B (Long-Context Retention): Evaluates rule stability after the insertion of ten semantically unrelated distractor questions.
- Task C (Belief Updating): Assesses the model's ability to detect errors and revise outputs when given targeted rule-specific feedback and structural hints.

By requiring models to master an entirely novel symbolic transcription system entirely within context, KHOM THAI provides a clean and highly discriminative testbed for studying adaptive in-context learning. We believe benchmarks of this nature are essential for meaningfully tracking progress toward robust, human-like intelligence.

2. Literature Review

2.1 Data Contamination in LLM Evaluation

The reliability of LLM benchmarks has come under increasing scrutiny due to data contamination the leakage of evaluation examples or closely related content into pre-training corpora. Early awareness of this issue appeared in the GPT-3 paper [7], which used n-gram overlap detection to identify potential contamination. Subsequent work has shown that contamination is widespread and can significantly inflate reported performance.

Deng et al. [2024] investigated modern benchmarks and found that models like ChatGPT and GPT-4 could predict missing options in MMLU questions with surprisingly high accuracy (52% and 57% exact match rates, respectively)[3], suggesting substantial memorization. Xu et al. [2024] provided a comprehensive survey on benchmark data contamination, highlighting detection methods, impacts on evaluation validity, and the growing difficulty of maintaining contamination-free test sets as models scale.

More recent surveys, including Cheng et al. [2025] ("A Survey on Data Contamination for Large Language Models"), emphasize the need for rigorous, contamination-free evaluation protocols. They categorize contamination into input-only and input-and-label types, and stress that larger models tend to benefit more from even minor leakage due to their superior memorization capacity. Other studies [3] further demonstrate that contamination effects are not negligible and can lead to misleading conclusions about model capabilities.

These findings underscore a fundamental limitation of static benchmarks: as training datasets grow to internet scale, it becomes nearly impossible to guarantee that test instances have not been indirectly encountered during pre-training.

2.2 In-Context Learning and Adaptive Reasoning

In-context learning (ICL) the ability of LLMs to adapt to new tasks using only a few examples provided in the prompt, without parameter updates has emerged as a core capability of modern

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

LLMs [7]. While ICL enables impressive few-shot performance on many tasks, its limits in true rule acquisition and generalization remain an active area of research.

Recent work has explored more demanding forms of ICL, such as many-shot inductive reasoning [MIR-Bench, 2025] and general-purpose in-context learning benchmarks that test acquisition of novel patterns or even entirely new "languages" [8]. These studies show that while frontier models can handle simple pattern matching well, they often struggle with complex, hierarchical, or exception-rich rule systems when the rules must be fully internalized from context.

2.3 Abstraction and Reasoning Benchmarks

A seminal contribution to measuring fluid intelligence and skill-acquisition efficiency is the Abstraction and Reasoning Corpus (ARC-AGI) introduced by Chollet [2019]. Chollet argued that true intelligence should be measured by the ability to efficiently acquire new skills on novel problems rather than by performance on crystallized knowledge tasks. The ARC benchmark and its subsequent iterations (ARC-AGI-2 and ARC-AGI-3) have become important reference points for evaluating core knowledge priors and abstraction abilities [9].

Despite its influence, even ARC-style benchmarks face challenges as models improve through targeted engineering and refinement loops. The ARC Prize series has documented progressive improvements in model scores while emphasizing the need for continually evolving, low-leakage tasks that remain easy for humans but difficult for current AI systems.

2.4 Synthetic and Procedural Benchmarks for Zero-Leakage Evaluation

To mitigate contamination risks, researchers have increasingly turned **to** synthetic dataset generation and procedural content generation. These approaches create fresh evaluation instances on-the-fly or through algorithmic rules, making prior exposure highly unlikely.

Examples include procedurally generated mazes for navigation and exploration [benchmarking GPICL], meta-languages for testing language acquisition from scratch, and dynamic task suites that emphasize rule induction over memorization. Such benchmarks align closely with the goals of measuring cognitive plasticity the flexible adaptation to novel symbolic or logical systems.

Our work builds upon and extends this line of research by introducing a linguistically inspired yet fully fictionalized symbolic notation system. Unlike purely abstract grid-based tasks (e.g., ARC), KHOM THAI leverages the structural richness of a Brahmic-style script (consonant classes, subscripts, vowel positioning, and complex interactions) while ensuring zero realistic leakage through deliberate fictional extensions and procedural gold generation.

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

2.5 Belief Updating and Long-Context Capabilities

Finally, related work has examined models' ability to update beliefs based on corrective feedback and to maintain learned knowledge over long contexts filled with distractors. While some studies show promising correction abilities in short interactions, retention over extended contexts with interfering content remains a notable weakness in many models.

KHOM THAI uniquely integrates these threads contamination resistance, rule generalization, long-context retention, and belief updating into a single, coherent, and highly structured benchmark domain.

3. Dataset and Methodology

This section describes the design and construction of the **KHOM THAI** benchmark, the three evaluation tasks, the procedural generation process, and the experimental setup used to evaluate frontier large language models.

3.1 Benchmark Overview

KHOM THAI v3 is built upon a romanized notation system inspired by the historical **Khom Thai** script — a Brahmic abugida traditionally used in Thailand for writing Pali texts. To ground the benchmark in linguistic authenticity while maintaining zero leakage, we extended the original script with new fictional rules and notations.[10]

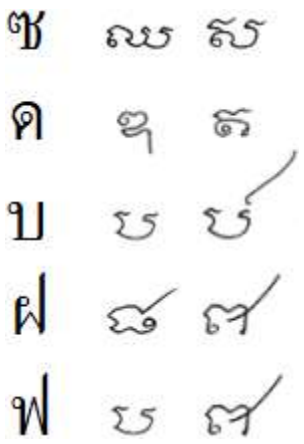


Figure 3.1: Examples of Real Khom Thai Vowels

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

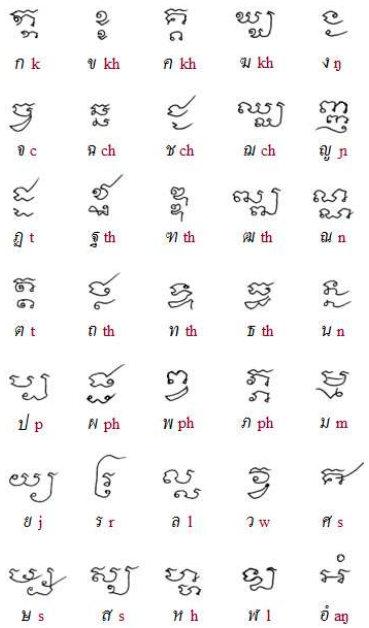


Figure 3.2: Real Khom Thai Consonant Forms

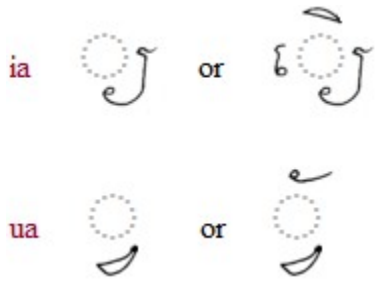


Figure 3.3: Vowel Positioning Examples (ia and ua)

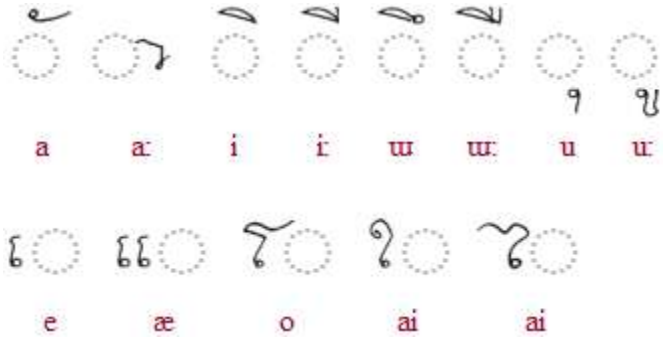


Figure 3.4: Additional Vowel and Consonant Marks

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

These authentic examples informed the structural design of our fictionalized notation system (consonant classes, subscript clusters, and vowel positioning), while all specific rules, interactions, and gold labels were procedurally generated to ensure zero training data leakage.

3.2 Rule System Design

The core of KHOM THAI is a carefully engineered symbolic notation system with four interlocking rule categories:

1. **Consonant Classes**: Consonants are divided into three classes (mid, high, low) that influence subscript and interaction behavior.
2. **Subscript Cluster Rules**: Different notations are used depending on consonant classes (hsub, msub, sub, hh, rwrap, ras).
3. **Vowel Notation Rules**: Vowels attach in different positions (above, below, left, right, surrounding) with special handling for diphthongs and two newly introduced vowels (eu, aw).
4. **Interaction Rules**: Complex interactions include high-class subscript modifiers (~), rwrap-induced order inversion for above vowels, automatic lengthening of short /a/ after low-class initials, and class-sensitive doubling (ldbl, mdbl, hdbl, rdbl) for medial /a/ omission.

These rules were deliberately fictionalized and extended beyond the original Khom Thai script to minimize any possibility of training data leakage while preserving linguistic plausibility.

3.3 Dataset Generation

We implemented a deterministic gold-standard notation engine in Python using fixed random seeds. The dataset generation process includes:

- **Tiered Difficulty Structure**: Five progressively harder tiers in Task A, ranging from simple bare consonants + vowels (Tier 1) to complex medial-a omission with class-sensitive doubling (Tier 5).
- **Procedural Sampling**: Random but reproducible sampling of consonants, clusters, and vowels according to tier-specific constraints.
- **Gold Label Generation**: A central khom_notation() function that applies all rules in the correct order to produce ground-truth notation strings.

All datasets are generated on-the-fly during evaluation to prevent accidental leakage through saved files. The final datasets contain 500 instances for Task A and 150 instances each for Tasks B and C.

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

3.4 Evaluation Tasks

Task A (Rule Generalization): Models receive the full ~800-token rule description (including tables, worked examples, and interaction order) followed by individual test cases. They must output only the correct notation string in lowercase.

Task B (Long-Context Retention): Models are first taught the complete rule system, then asked 10 unrelated general knowledge questions (distractors), and finally tested on the target syllable without restating the rules.

Task C (Belief Updating): Models attempt the task up to three times. After incorrect answers, they receive targeted corrective feedback naming the specific violated rule, followed by a structural hint (prefix of the gold answer) on the third attempt. Success on any attempt counts as a pass.

Response cleaning includes removal of thinking tags, markdown, and common answer prefixes using regular expressions.

3.5 Models, Experimental Setup, and Implementation

3.5.1 Models Evaluated

We evaluated 14 state-of-the-art and strong open-weight models, including:

- Google Gemini series: gemini-2.5-flash, gemini-2.5-pro, gemini-3-pro-preview, gemini-3.1-pro-preview
- Anthropic Claude series: claude-haiku-4-5, claude-sonnet-4-6, claude-opus-4-6
- DeepSeek: deepseek-.2, deepseek-r1-0528
- Qwen series: qwen3-235b-a22b-instruct-2507, qwen3-next-80b-a3b-instruct, qwen3-next-80b-a3b-thinking
- ZAI: glm-5
- Google Gemma: gemma-3-27b

Model selection focused on frontier and near-frontier systems to maximize discriminative power while maintaining computational feasibility.

3.5.2 Experimental Infrastructure

All experiments were conducted using the **Kaggle Benchmarks SDK** (kaggle_benchmarks or kbench), a custom evaluation harness developed for this project. The framework supports:

- Parallel evaluation with configurable n_jobs
- Crash-safe checkpointing (results saved after each model)

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

- Structured chat management for multi-turn tasks (Tasks B and C)
- Automated response cleaning and exact-match scoring
- Detailed per-tier and per-task breakdown reporting

3.5.3 Code Development

We developed a complete, self-contained Python codebase consisting of:

- CONSONANTS, VOWELS, and rule dictionaries
- Core notation engine (build_cluster(), apply_vowel(), khom_notation_())
- Dataset generators (generate_())
- Full rule prompt (FULL_RULES_)
- Response cleaning function (clean_response())
- Targeted feedback generator for Task C (targeted_feedback_())
- Three @kbench.task decorated evaluation functions
- Main runner script with per-model execution and leaderboard generation

The entire pipeline is deterministic and reproducible. All gold labels are generated using fixed seeds. The codebase, evaluation harness, and generated datasets (via seeds) will be made publicly available upon publication to support future research and replication studies.

3.5.4 Evaluation Protocol

- Task A: n_jobs=8, direct prompting with full rules
- Tasks B & C: n_jobs=4, multi-turn chat management via the SDK
- Pass thresholds: ≥40% (Task A), ≥30% (Task B), ≥45% (Task C)
- Performance is reported as accuracy (exact string match after cleaning)

4. Results

This section presents the main experimental findings from evaluating 14 frontier and strong large language models on the three tasks of the KHOM THAI benchmark.

4.1 Overall Performance

Table 4.1 summarizes the performance of all evaluated models across the three tasks. Models are ranked by accuracy on **Task A (Rule Generalization)**, as this task serves as the primary measure of a model's ability to internalize and apply novel symbolic rules.

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

Table 4.1: Overall Performance on KHOM THAI

Rank	Model	Task A (Rule Gen)	Task B (Retention)	Task C (Belief Update)	Verdict
1	Gemini-2.5-Pro	**96.4%**	**96.0%**	**97.3%**	☑ ALL-3
2	Gemini-3-Pro-Preview	91.8%	(quota)	6.7%	⚡ 1/3
3	Gemini-2.5-Flash	84.4%	87.3%	91.3%	☑ ALL-3
4	ZAI GLM-5	84.0%	(quota)	8.0%	⚡ 1/3
5	Qwen3-Next-80B-Thinking	80.4%	85.3%	92.7%	☑ ALL-3
6	Qwen3-235B-A22B-Instruct-2507	70.0%	68.7%	94.7%	☑ ALL-3
7	Claude-Sonnet-4-6	69.4%	79.3%	92.7%	☑ ALL-3
8	DeepSeek-R1-0528	67.2%	43.3%	25.3%	⚡ 2/3
9	Gemma-3-27B	67.2%	56.0%	90.7%	☑ ALL-3
10	DeepSeek-.2	56.4%	52.0%	90.7%	☑ ALL-3
11	Claude-Haiku-4-5	59.2%	67.3%	82.0%	☑ ALL-3
12	Claude-Opus-4-6	54.2%	(quota)	37.3%	⚡ 1/3
13	Qwen3-Next-80B-Instruct	51.4%	56.7%	93.3%	☑ ALL-3
14	Gemini-3.1-Pro-Preview	6.2%	(quota)	0.0%	✗ FAIL

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

Note: “” indicates runs that encountered API quota or rate-limit errors. Pass thresholds: Task A ≥40%, Task B ≥30%, Task C ≥45%. “ALL-3” indicates a model passed all three tasks.

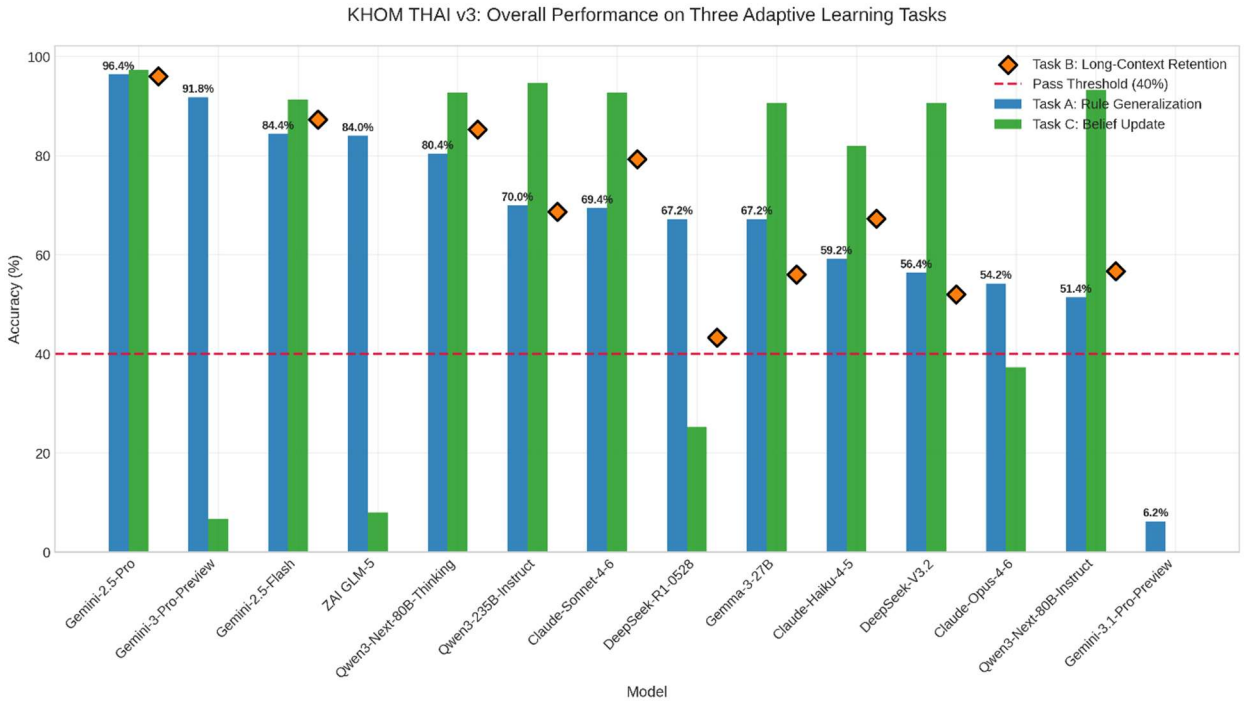


Figure 4.1: Overall Performance on KHOM THAI v3

4.2 Task A: Rule Generalization

Task A represents the core evaluation of a model’s ability to learn and apply the full KHOM THAI rule system from a single comprehensive prompt. Performance varied significantly across difficulty tiers, revealing clear capability gaps.

Table 4.2: Task A Performance by Difficulty Tier (%)

Model	T1	T2	T3	T4	T5	Overall
Gemini-2.5-Pro	100.0	100.0	95.5	89.3	100.0	**96.4**
Gemini-3-Pro-Preview	98.8	93.8	89.3	85.1	95.6	91.8
Gemini-2.5-Flash	83.8	88.7	83.0	71.1	100.0	84.4

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

Model	T1	T2	T3	T4	T5	Overall
ZAI GLM-5	90.0	86.6	80.4	75.2	92.2	84.0
Qwen3-Next-80B-Thinking	83.8	84.5	74.1	66.9	98.9	80.4
Qwen3-235B-Instruct	92.5	87.6	43.8	55.4	83.3	70.0
Claude-Sonnet-4-6	98.8	87.6	62.5	49.6	58.9	69.4
DeepSeek-R1-0528	72.5	61.9	61.6	52.1	95.6	67.2
Gemma-3-27B	91.2	86.6	45.5	55.4	67.8	67.2
DeepSeek-.2	86.2	58.8	37.5	30.6	85.6	56.4

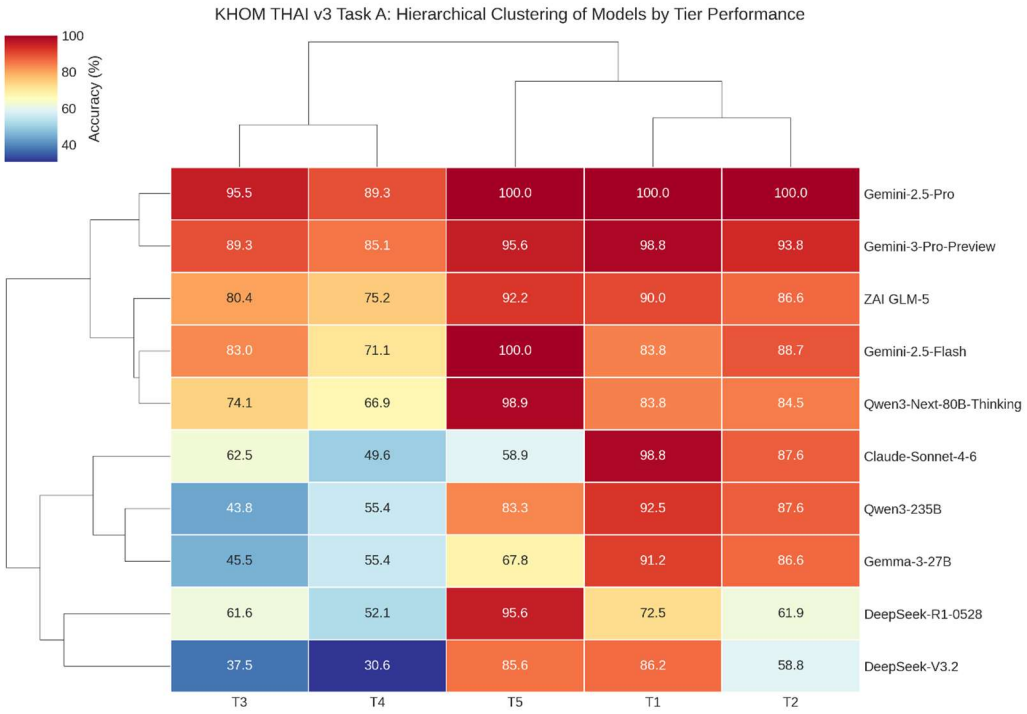


Figure 4.2: Tier-wise Performance Heatmap

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

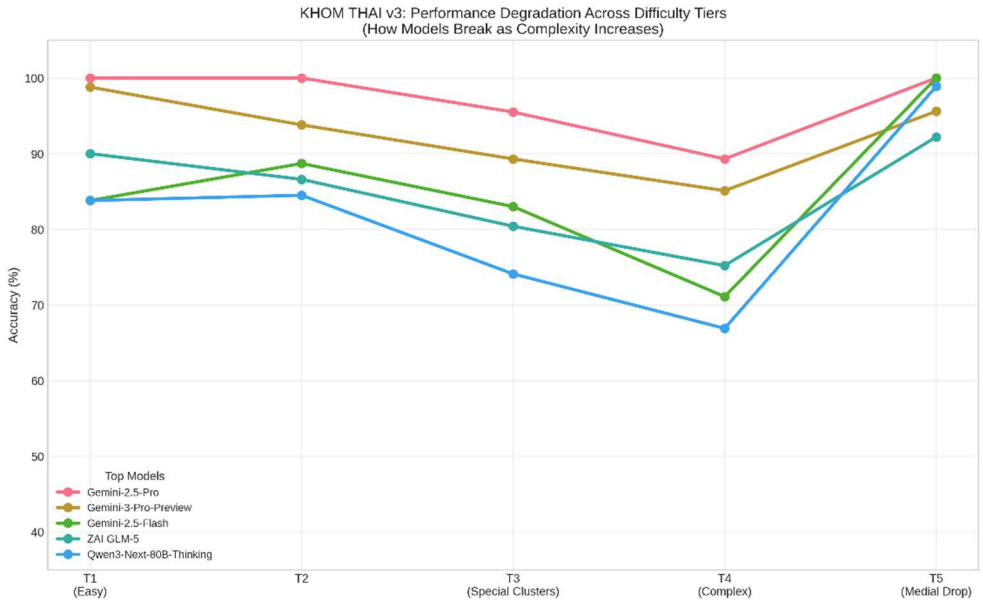


Figure 4.3: Performance Degradation Across Tiers (Top Models)

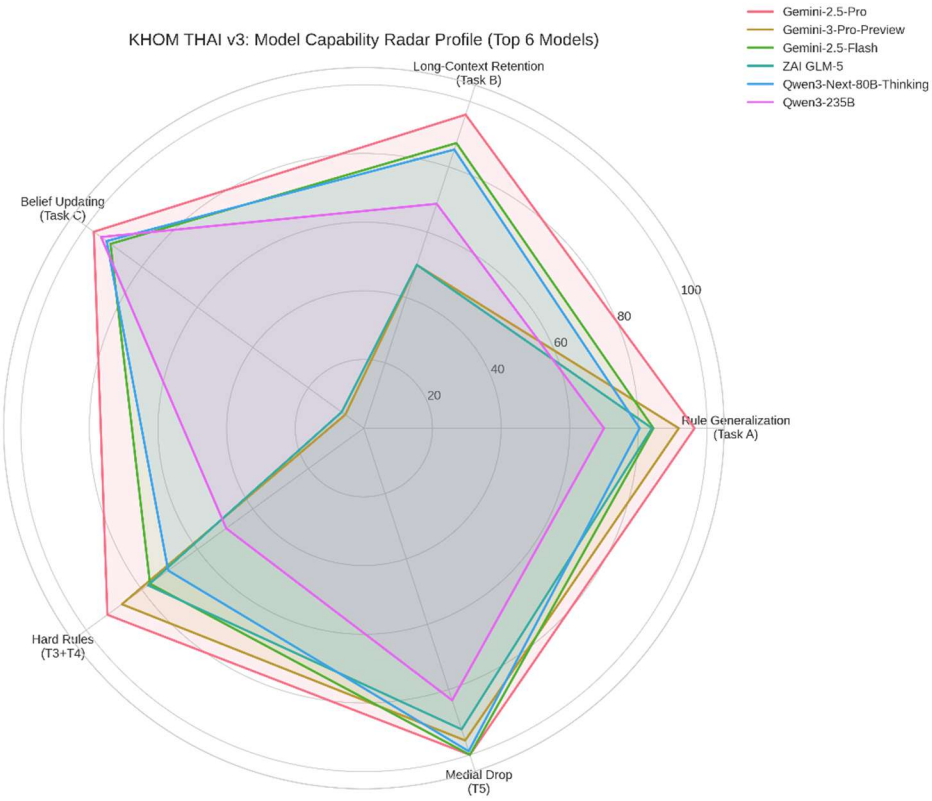


Figure 4.4: Model Capability Radar Profile

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

Key Observations:

- **Tier 1 & 2** (simple rules): Most models performed strongly, indicating good basic rule following.
- **Tier 3** (special clusters: rwrap, ras, msub): This tier proved highly discriminative. Several models showed sharp performance drops.
- **Tier 4** (hsub + ~ modifier, hh clusters, new vowels, diphthongs): The most challenging tier for the majority of models.
- **Tier 5** (medial /a/ omission with class-sensitive doubling): Surprisingly strong performance by several models (Gemini series and Qwen-thinking reached near or above 95%).

Gemini-2.5-Pro demonstrated near-perfect mastery across nearly all tiers, while many other models struggled particularly with complex cluster-vowel interactions in Tiers 3 and 4.

4.3 Task B: Long-Context Retention

Task B tested whether models could maintain the learned KHOM THAI rules after processing ten unrelated distractor questions. Top models (Gemini-2.5-Pro and Gemini-2.5-Flash) showed excellent retention with minimal degradation. In contrast, reasoning-oriented models like DeepSeek-R1-0528 exhibited more noticeable contextual drift.

4.4 Task C: Belief Updating

Belief updating emerged as a relative strength for most models. Even models that performed modestly on Task A often achieved high correction rates when given targeted rule-specific feedback. Qwen and Claude models, in particular, showed strong performance in this task, with several exceeding 92%. This suggests that current frontier models are relatively capable of revising their outputs when provided with explicit corrective signals.

4.5 Error Analysis

Common failure modes observed in Task A included:

- Missing the ~ modifier after high-class subscripts (hsub)
- Incorrect order for above-vowels with rwrap clusters
- Failure to apply class-sensitive doubling (ldbl/mdbl/hdbl/rdbl) in Tier 5
- Confusion between ras(r) and rwrap for h+r clusters
- Perseveration on incorrect cluster notations despite feedback (more common in weaker models)

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

These errors highlight that while models can learn basic rules, handling exceptions and interaction effects remains challenging.

5. Discussion and Future Work

5.1 Key Findings

The results of the KHOM THAI benchmark reveal both promising capabilities and significant limitations in current frontier large language models regarding adaptive in-context learning.

The strongest model, Gemini-2.5-Pro, achieved near-ceiling performance on Task A (96.4%), maintained excellent rule retention even after ten distractor questions (96.0%), and demonstrated outstanding belief updating (97.3%). This suggests that certain frontier models have made substantial progress in rapidly internalizing complex, multi-layered symbolic rule systems when provided with clear, structured instructions.

However, performance varied dramatically across models and difficulty tiers. While most models handled basic rules (Tiers 1–2) reasonably well, Tier 3 (special clusters such as rwrap, ras, and msub) and Tier 4 (high-class subscript modifiers, hh clusters, new vowels, and interaction effects) emerged as major bottlenecks. These tiers require not only memorizing individual rules but also understanding and correctly applying their interactions a capability that still challenges many models.

Notably, belief updating (Task C) proved to be a relative strength across the board. Even models that performed modestly on initial rule generalization often achieved high correction rates when given targeted, rule-specific feedback. This indicates that current LLMs are reasonably capable of revising their outputs when explicitly guided, but they still struggle with proactive error detection and robust rule application without external scaffolding.

Long-context retention (Task B) showed more variability. While top models maintained performance well, others exhibited noticeable degradation, suggesting that contextual interference remains a challenge for stable rule maintenance over extended sequences.

5.2 Implications for LLM Evaluation and AGI Progress

KHOM THAI makes several important contributions to the field:

First, it demonstrates the value of zero-leakage benchmarks constructed through synthetic dataset generation and procedural content generation. By creating an entirely novel symbolic system with complex interactions, we reduce the risk of data contamination that plagues traditional benchmarks.

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

Second, the benchmark successfully isolates three distinct cognitive abilities rule generalization, long-context retention, and belief updating within a single coherent domain. This unified approach provides a more holistic view of a model's cognitive plasticity than fragmented evaluations across unrelated tasks.

Third, the tiered difficulty structure and rich interaction rules expose capability gaps that are often masked in simpler benchmarks. The sharp performance drops observed in Tiers 3 and 4 suggest that current models still rely heavily on pattern matching rather than deep structural understanding of rule systems.

These findings align with broader discussions on measuring progress toward Artificial General Intelligence (AGI). As argued by Chollet [2019], true intelligence involves the efficient acquisition of new skills on novel problems. KHOM THAI contributes to this vision by providing a challenging, human-interpretable task that remains difficult for today's most advanced systems.

5.3 Limitations

Several limitations should be acknowledged. First, some Task B runs for preview models encountered API quota limitations, resulting in incomplete retention data. Second, while we made every effort to ensure zero leakage through fictionalized rules, future models trained on this paper or its associated artifacts could potentially benefit from exposure. Third, the benchmark currently focuses on a single symbolic domain; broader generalization across different types of rule systems remains to be tested.

Additionally, we did not collect human baselines in this study. Comparing model performance with educated humans after studying the same rule set would provide valuable context regarding the remaining gap to human-level adaptive learning.

5.4 Future Work

We envision several promising directions for extending this research:

1. Human Baseline Study: Conduct a controlled experiment with human participants to establish performance benchmarks and analyze differences in error patterns and learning strategies between humans and LLMs.
2. Multi-Syllable and Compositional Tasks: Extend the benchmark to full words and short phrases, requiring models to handle rule interactions across syllable boundaries.
3. Dynamic Rule Evolution: Introduce tasks where rules change or accumulate across multiple turns, testing meta-learning and continual adaptation.
4. Visual Modality Integration: Pair the romanized notation with actual Khom Thai-like glyph rendering to create a multimodal version of the benchmark.
5. Adversarial Rule Variants: Generate "rule-breaking" or highly counterintuitive variants to further stress-test models' flexibility and robustness.

International Journal of DeepTech in Medical Science and Technology,
Vol 1, No. 1, September-October 2025
ISSN: 3088-2370 (Online) | © Geoinformatics International

CC BY

6. Open-Source Benchmark Suite: Release an evolving version of KHOM THAI with new rule sets and procedurally generated instances to prevent future contamination as models improve.

5.5 Conclusion

KHOM THAI demonstrates that while frontier models have made impressive strides in adaptive in-context learning, significant challenges remain in mastering complex, exception-rich symbolic systems particularly when rules involve subtle interactions and exceptions.

By providing a clean, reproducible, and highly discriminative benchmark, we hope this work contributes to more rigorous evaluation methodologies that better reflect the requirements of genuine intelligence. As LLMs continue to advance, benchmarks like KHOM THAI will play a crucial role in tracking meaningful progress toward robust, flexible, and human-like artificial intelligence.

References

[1] C. Xu, S. Guan, D. Greene, and M.-T. Kechadi, "Benchmark Data Contamination of Large Language Models: A Survey," Jun. 2024.

[2] C. Deng, Y. Zhao, X. Tang, M. Gerstein, and A. Cohan, "Investigating Data Contamination in Modern Benchmarks for Large Language Models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, Association for Computational Linguistics (ACL), 2024, pp. 8698–8711. doi: 10.18653/v1/2024.naacl-long.482.

[3] S. Golchin and M. Surdeanu, "TIME TRAVEL IN LLMS: TRACING DATA CONTAMINATION IN LARGE LANGUAGE MODELS," in *12th International Conference on Learning Representations, ICLR 2024*, International Conference on Learning Representations, ICLR, 2024.

[4] O. Sainz, J. A. Campos, I. García-Ferrero, J. Etxaniz, O. Lopez de Lacalle, and E. Agirre, "NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics (ACL), 2023, pp. 10776–10787. doi: 10.18653/v1/2023.findings-emnlp.722.

[5] S. Balloccu, P. Schmidtová, M. Lango, and O. Dušek, "Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs," in *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2024, pp. 67–93. doi: 10.18653/v1/2024.eacl-long.5.

[6] F. Chollet, “On the Measure of Intelligence,” Nov. 2019.

[7] T. Brown *et al.*, “Language Models are Few-Shot Learners,” https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[8] Y. Cheng, Y. Chang, and Y. Wu, “A Survey on Data Contamination for Large Language Models,” *arXiv preprint arXiv:2502.14425*, 2025.

[9] F. Chollet, M. Knoop, G. Kamradt, and B. Landers, “ARC Prize 2025: Technical Report,” Jan. 2026.

[10] “Khom Thai alphabet,” https://www.omniglot.com/writing/khomthai.htm.