TASE: Token Awareness and Structured Evaluation for Multilingual Language Models

Chenzhuo Zhao[1*], Xinda Wang[1*], Yue Huang[1], Junting Lu[1], Ziqian Liu[2]
[1]Peking University
[2]The University of Hong Kong
{cyzcz, nev_settle,huangyue_vl,aidan.lew.37}@stu.pku.edu.cn
liuziqian25@gmail.com

Abstract

While large language models (LLMs) have demonstrated remarkable performance on high-level semantic tasks, they often struggle with fine-grained, token-level understanding and structural reasoning—capabilities that are essential for applications requiring precision and control. We introduce TASE, a comprehensive benchmark designed to evaluate LLMs' ability to perceive and reason about token-level information across languages. TASE covers 10 tasks under two core categories: token awareness and structural understanding, spanning Chinese, English, and Korean, with a 35,927-instance evaluation set and a scalable synthetic data generation pipeline for training. Tasks include character counting, token alignment, syntactic structure parsing, and length constraint satisfaction. We evaluate over 30 leading commercial and open-source LLMs, including O3, Claude 4, Gemini 2.5 Pro, and DeepSeek-R1, and train a custom Qwen2.5-14B model using the GRPO training method. Results show that human performance significantly outpaces current LLMs, revealing persistent weaknesses in token-level reasoning. TASE sheds light on these limitations and provides a new diagnostic lens for future improvements in low-level language understanding and cross-lingual generalization.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities across a wide range of natural language tasks. They excel in high-level semantic understanding such as instruction following, logical reasoning, long-context comprehension, and code generation.(Hua et al. 2025; Kostikova et al. 2025; Wan et al. 2024) These strengths have driven their adoption in various applications, including conversational agents, educational tools, and problem-solving systems.

Despite their success on complex tasks, LLMs often struggle with surprisingly simple, fine-grained tasks that require token-level perception and structural reasoning.(Wang et al. 2025; Hiraoka and Inui 2025) For example, even top-tier models frequently fail to count the number of letter "r"s in *"strawberry"*, or perform basic operations such as detecting spelling errors or manipulating individual characters.

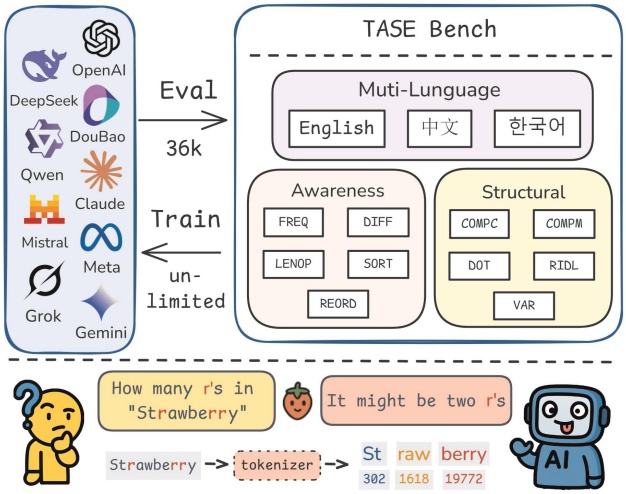

Figure 1: Overview of the TASE benchmark. TASE evaluates LLMs across fine-grained token-level tasks in three languages and two dimensions: token awareness and structural understanding. The strawberry example illustrates common model failures in token-level reasoning.

These shortcomings reveal a persistent gap in token awareness—that is, the model's ability to perceive, reason about, and operate on individual tokens or characters with precision.

A key source of this problem is the reliance on subword tokenization schemes like Byte-Pair Encoding (Shibata et al. 1999), which obscure internal character structures and are not designed for character-level reasoning. This deficiency is especially pronounced in non-English languages like Chinese and Korean, where complex compositional structures pose even greater challenges. Yet, this fundamental blind spot has been largely ignored in mainstream evaluation. Prominent benchmarks like GLUE(Wang et al. 2018), SuperGLUE(Wang et al. 2019), and XNLI(Conneau et al. 2018) almost exclusively target high-level semantic understanding, overlooking tasks that require direct token manipulation or structural analysis. While recent work has begun to probe these limitations(Xu and Ma 2025; Yehudai et al. 2024), a comprehensive, multilingual benchmark

*These authors contributed equally.

focused on these fine-grained abilities is still critically lacking.

To facilitate a systematic evaluation, we design **TASE**, a benchmark with ten fine-grained tasks across three typologically diverse languages: English, Chinese, and Korean, representing alphabetic, logographic, and featural writing systems respectively (see Figure 1). TASE evaluates two core dimensions of token-level understanding. The first, token awareness, directly tests a model's perception of linguistic units through tasks such as counting words in a sentence, generating text of a specified length, identifying minimal token differences, and reordering sentences under strict adjacency constraints. The second, structural understanding, probes the model's ability to analyze the internal form of tokens. This includes tasks such as counting characters within a word, solving composition puzzles, reconstructing text from corrupted representations (e.g., dot-matrix patterns), and recognizing visual patterns in text. Together, these tasks evaluate not just symbolic awareness but also structure-sensitive reasoning.

We construct a curated evaluation set of 35,927 instances across all tasks and languages, ensuring a broad and reliable basis for performance measurement. Moreover, we develop a scalable synthetic data generation pipeline capable of producing unlimited training examples with guaranteed correctness. This pipeline allows researchers to train or fine-tune models on these tasks and to conduct controlled experiments on how token-level training affects model behavior.

We benchmark more than 30 leading LLMs, including proprietary systems such as GPT-4.1, Claude 4, and Gemini 2.5 Pro, and state-of-the-art open models like DeepSeek-R1. In addition, using our synthetic data and the GRPO algorithm(Shao et al. 2024), we fine-tuned a new model based on Qwen2.5-14B-Instruct. Our evaluation reveals a clear and consistent gap between human and model performance. Even the strongest models underperform substantially on TASE tasks, especially those requiring structure-aware reasoning such as visual recognition or component-level decomposition. For example, tasks involving character composition or spatial reasoning often produce incorrect or hallucinated outputs, while simple length-controlled generation is frequently violated. Despite these shortcomings, we find that targeted fine-tuning yields measurable improvements, as our trained model surpasses its base on several tasks, demonstrating that enhancing fine-grained capabilities can help narrow the performance gap. Nevertheless, no model achieves human-level performance across all tasks or languages, underscoring that token-level and structure-aware understanding remains a core challenge for current LLMs.

We summarize our main contributions as follows:

- A multilingual benchmark specifically designed to evaluate token-level awareness and structural reasoning in LLMs, spanning English, Chinese, and Korean.
- A reproducible dataset of 35,927 evaluation instances across ten fine-grained tasks, covering both perception and manipulation of linguistic structure.
- A scalable synthetic data generation pipeline for each task, enabling training, fine-tuning, and controlled analysis of model behavior.
- A comprehensive evaluation of over 20 state-of-the-art LLMs and a fine-tuned 14B custom model, revealing persistent weaknesses and quantifying the gap between models and human performance.

2 Related Work

2.1 High-Level Understanding Benchmarks

Most existing benchmarks for large language models (LLMs) focus on high-level semantic tasks. GLUE (Wang et al. 2018) and SuperGLUE (Wang et al. 2019) emphasize sentence-level classification, inference, and question answering. Their multilingual extensions, such as XGLUE (Liang et al. 2020), XNLI (Conneau et al. 2018), and TyDiQA (Clark et al. 2020), apply similar task types to non-English languages. More recent efforts like P-MMEval and BenchMAX expand the coverage to multilingual reasoning, coding, and instruction following. However, these benchmarks do not evaluate fine-grained capabilities such as character counting, token alignment, or structural analysis. Furthermore, current benchmarks (Wang et al. 2023; Singh et al. 2024; Lai et al. 2023) lack support for testing token-level or language-agnostic skills, particularly in low-resource settings.

2.2 Token Awareness and Fine-Grained Evaluation

Several studies have recently highlighted that LLMs exhibit surprising failures on basic token-aware tasks. Xu and Ma (2025) show that even top-tier models frequently miscount letters within a word, despite understanding their semantic context. Fu et al. (2023) find that models recognize letters but often fail to count them accurately. Yehudai et al. (2024) provide a theoretical framework suggesting that fixed-size transformers struggle with simple counting operations due to architectural limitations. Benchmarks like LMEntry (Efrat, Honovich, and Levy 2022) and CUTE test models on elementary character-level operations (e.g., identifying first/last letters, swapping characters) and expose consistent performance gaps. These findings reveal fundamental weaknesses in LLMs' token-level reasoning, often obscured in high-level evaluations. However, current studies remain fragmented, with evaluations scattered across isolated tasks, and lack a unified or systematic framework for assessing token-level capabilities.

2.3 Structural Reasoning and Text Manipulation

Beyond token awareness, structural reasoning tasks such as text editing or visual token recognition remain underexplored. CWUM introduces a bilingual benchmark targeting letter-level editing and reordering, reporting large gaps between human and model accuracy. LLMs often fail to enforce formatting constraints, satisfy length conditions, or recover corrupted character structures (Zhou et al. 2023). Attempts to address these issues through synthetic data augmentation or decomposition strategies (for example, spelling out words or breaking tokens into components) offer partial improvements. Still, systematic evaluation suites targeting such structural capabilities across languages remain limited. TASE represents the first systematic multilingual benchmark

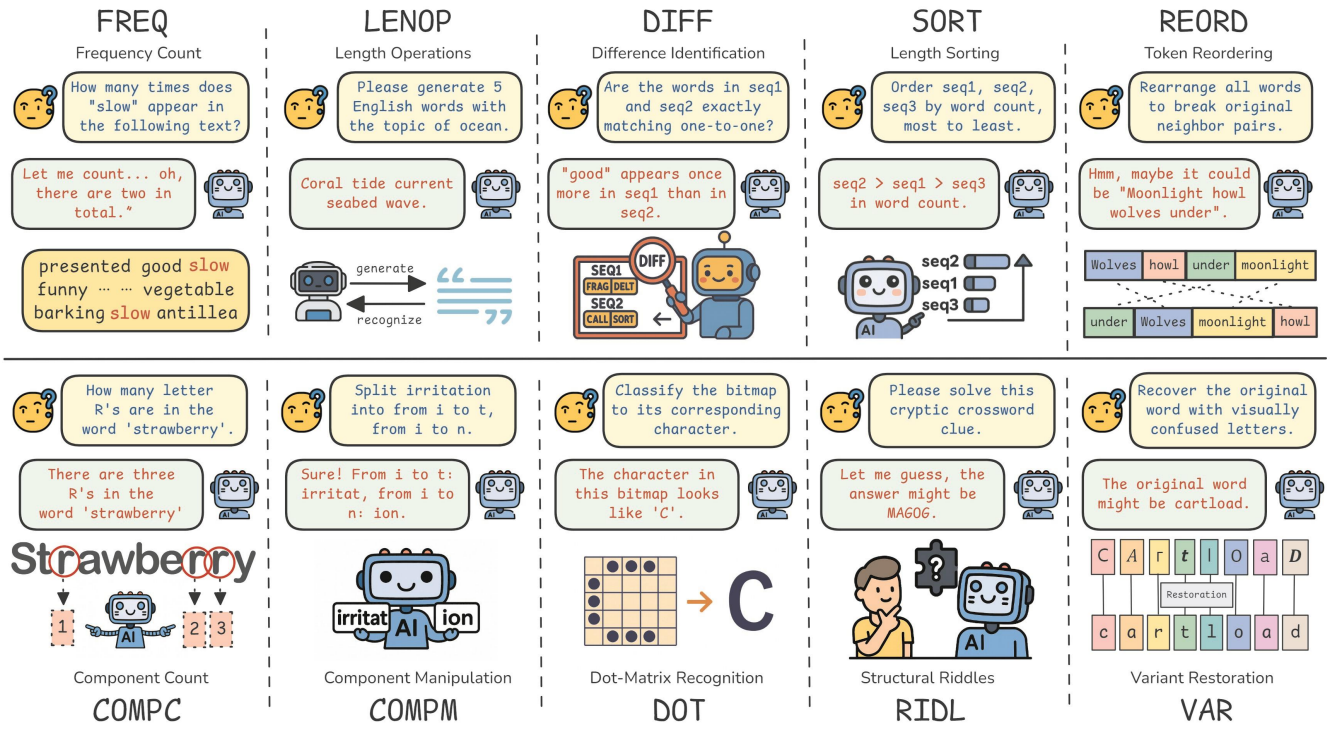

Figure 2: Illustration of the ten TASE tasks, grouped by two core capabilities: **Token Awareness** (top row) and **Structural Understanding** (bottom row). Each cell shows an example highlighting the specific reasoning or perception skill evaluated by the task.

designed to evaluate both token-level awareness and structural reasoning capabilities.

3 TASE

Three characteristics differentiate TASE from existing LLM benchmarks: (1) TASE shifts the evaluation focus from abstract semantic comprehension to concrete, low-level text processing skills, directly probing a model's awareness of tokens and their internal structure. (2) TASE systematically investigates cross-lingual generalization by incorporating parallel tasks across English (alphabetic), Chinese (logographic), and Korean (featural syllabary) to reveal biases associated with different linguistic structures. (3) TASE is designed for scalability and reproducibility, providing not only a fixed evaluation set for fair comparison but also a synthetic data generation pipeline for creating a virtually unlimited volume of training examples.

3.1 Taxonomy of Language Capabilities

We define 10 fundamental low-level language capabilities, organized into two distinct categories: token awareness and structural understanding. These tasks are designed to assess a model's perception and manipulation of the fundamental components of text. A detailed overview and visual examples of each task are provided in Figure 2.

Token Awareness This category focuses on direct operations over discrete token sequences, treating each token as an atomic unit without considering its internal structure.

We formalize all tasks in this category as functions operating over token sequences:

$$F_{\text{token}} : S \text{ or } (S_1, S_2) \text{ or } \mathcal{S} \to A$$

where $S = \langle t_1, t_2, \ldots, t_n \rangle$ is a token sequence, $\mathcal{S}$ is a set of such sequences, and A is a scalar, token, or reordered sequence depending on the task (e.g., count, difference set, sorted list).

- **Frequency Count (FREQ)**: the ability to accurately count all occurrences of a specific token within a text.
- **Length Operations (LENOP)**: the ability to count the number of tokens in a sentence and to generate a sentence with a precise number of tokens.
- **Difference Identification (DIFF)**: the ability to compare two sets of tokens and identify the single differing token between them.
- **Length Sorting (SORT)**: the ability to sort a list of sentences in descending order based on their token count.
- **Token Reordering (REORD)**: the ability to reorder a sequence of tokens so that no token remains next to its original neighbors.

Structural Understanding This category focuses on the internal composition of tokens, such as characters, radicals, strokes, or subwords, and their manipulation or reasoning.

We formalize all tasks in this category as functions that analyze or reconstruct sub-token structure:

$$G_{\text{struct}} : t \text{ or } S \text{ or } R \rightarrow B$$

where t is a token, S a sequence, R a corrupted or visual form (e.g., matrix), and B a component count, token, pattern, or restored form depending on the task.

- **Component Count (COMPC)**: the ability to count sub-token units, such as radicals in Chinese characters, letters in words, or jamo in Korean.
- **Component Manipulation (COMPM)**: the ability to combine constituent parts into a valid token or decompose a token into its fundamental components.
- **Dot-Matrix Recognition (DOT)**: the ability to recognize and classify characters from their visual representation as a binary matrix.
- **Structural Riddles (RIDL)**: the ability to solve riddles based on the orthographic or structural properties of words, not their semantic meanings.
- **Variant Restoration (VAR)**: the ability to identify and correct characters that have been replaced by visually similar homoglyphs.

3.2 Dataset Construction

The TASE benchmark is built upon a core evaluation set of 35,927 instances, among which three tasks under the dot category contain only 977 instances each. We designed the dataset with a specific emphasis on probing the foundational, structural capabilities of LLMs—a domain often overlooked by traditional benchmarks. The dataset is comprehensive, with 1,000 instances per language for most of our 10 defined tasks, ensuring robust measurement. Only the riddle task in English and Chinese leverages public data, while the vast majority of instances are programmatically generated by our synthetic pipeline. This approach guarantees not only diversity and scale but also ground-truth correctness for every instance, as the data is created with a known solution.

3.3 Evaluation Methodology

To enable a fair and objective assessment, our evaluation methodology is designed to be rigorous and straightforward. The tasks are deliberately structured to have unambiguous, close-ended answers, such as a specific number, a single word, or a precise ordering. The benchmark does not include tasks that require creative or open-ended generation.

This evaluation format is crucial as it minimizes the influence of a model's particular language generation style, allowing for a more direct and accurate measurement of its core reasoning abilities. The focus on verifiable answers simplifies automated scoring and ensures that TASE provides a level playing field to compare the low-level skills of different models across the multiple languages (English, Chinese, and Korean) in the dataset.

3.4 Data Generation Pipeline

Our benchmark is supported by a scalable synthetic data generation pipeline that creates evaluation instances using a set of fixed, programmatic rules. For Token Awareness tasks, we constructed questions by sampling from large, standardized word and character lists across English, Chinese, and Korean. For Structural Understanding tasks, we generated examples by systematically decomposing tokens into their fundamental components, such as breaking down Chinese characters into radicals or Korean syllables into phonetic elements, following established linguistic rules. This automated approach ensures the ground-truth correctness and consistency of our dataset. The complete details regarding the specific resources, tools, and methods for each task are available in the **appendix**.

4 Experiments

4.1 Experimental Setup

Benchmark Overview. Our evaluation is conducted on the TASE benchmark, a comprehensive suite designed to assess low-level language capabilities. TASE comprises 10 distinct tasks organized into two core categories: *Token Awareness* and *Structural Understanding*. The benchmark is multilingual, with tasks spanning Chinese, English, and Korean, and contains a total of 35,927 evaluation instances. For all model evaluations, we employed a consistent set of generation parameters: temperature of 0.7, maximum token limit (MaxTokens) of 16384, Top_p of 0.95, and Top_k of 50.

Evaluated Models. We evaluated over 20 leading large language models to provide a comprehensive view of the current landscape. These models fall into three categories. **Leading proprietary models** refer to top-tier commercial systems known for their state-of-the-art performance, such as GPT-4.1(Achiam et al. 2023), Claude Opus 4, Gemini 2.5 Pro(Comanici et al. 2025), and O3. **Mainstream open-source models** include a wide range of powerful, publicly available alternatives, such as DeepSeek-R1, the Qwen2.5 series (7B, 14B, 32B, and 72B)(Qwen et al. 2025), and Llama-3.3. Finally, we also developed a **custom fine-tuned model** based on Qwen2.5-14B-Instruct, trained using the GRPO algorithm. This approach leverages a finer-grained reward function and synthetic training data generated through the TASE pipeline, aiming to improve task-specific performance.

Evaluation Metrics and Human Baseline. The primary metric for evaluation across all TASE tasks is accuracy, or a normalized score derived from it. Performance is measured against a human baseline, which serves as the gold standard for these tasks. To establish this baseline, we recruited three native speakers for each language (Chinese, English, and Korean). For each task type, 200 questions were uniformly sampled and assigned to the evaluators. To affirm the validity of using a 200-item sample, we conducted a rigorous statistical analysis comparing sampled evaluations against full-dataset evaluations. The results confirm that this sample size is a highly reliable and accurate proxy for the full dataset's results (for a detailed statistical breakdown,see the **Appendix**). As noted in the introduction, humans achieve near-perfect accuracy on these fine-grained token manipulation and reasoning challenges. This human performance level represents the upper bound and the target for which

models should aim.

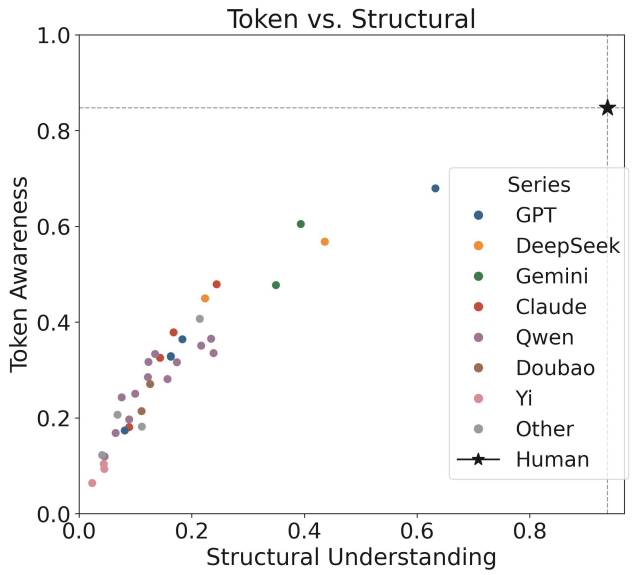

Figure 3: Token Awareness vs. Structural Understanding across major LLM families. This scatter plot illustrates the performance of various Large Language Models (LLMs) compared to human benchmarks on the TASE evaluation. The data reveals a strong positive correlation between a model's token-level awareness and its structural comprehension.

4.2 Overall Performance

Our evaluation highlights a key finding: all language models fall significantly short of human-level performance on the TASE benchmark. As shown in Table 1, humans achieve the highest scores across all metrics (average 89.24%), while the best model, **O3**, reaches only 65.60%, underscoring a persistent gap in fine-grained, high-precision language tasks.

As shown in Figure 3, most models remain distant from human-level capabilities across both Token Awareness and Structural Understanding. Only a few exhibit moderate competence. Structural tasks such as DOT, RIDL, and COMPM remain especially challenging, with models performing far below humans.

Token Awareness Evaluation. The Token Awareness tasks, which assess a model's ability to perceive, count, and manipulate basic textual units, revealed a highly polarized performance trend. While state-of-the-art models like **O3**, **DeepSeek-R1**, and **Gemini 2.5 Pro** showed foundational perceptual skills by achieving near-human performance on direct subtasks like Frequency Count (FREQ) and simple counting, their capabilities sharply decline when faced with complex constraints. This weakness is particularly stark in the Token Reordering (REORD) task, where nearly all models, including **GPT-4.1** and **Claude Opus 4**, score close to zero. Even the best-performing model, **O3**, only managed 31.60%, underscoring a critical deficiency in precise, constrained text manipulation.

Structural Understanding Evaluation. The Structural Understanding tasks, which are designed to probe a model's grasp of the internal visual form of tokens like components, radicals, and strokes, expose the most significant weakness across all evaluated LLMs. Even top-tier models such as O3 and Gemini 2.5 Pro demonstrated uniformly low scores on specific challenges including Component Count (COMPC), Dot-Matrix Recognition (DOT), and Structural Riddles (RIDL). This widespread failure lends strong empirical support to the "tokenizer blindness" hypothesis, which posits that because LLMs depend on subword tokenization, they are deprived of direct access to complete character-level or visual information. Consequently, these models are fundamentally ill-equipped to handle tasks that require structural discrimination or visual pattern recognition, leading to a collapse in their performance within this category.

4.3 Cross-lingual Performance

Pervasive Linguistic Imbalance. As shown in Table 2, a consistent pattern of linguistic imbalance emerges across all evaluated models. Performance is generally strongest in English, followed by Chinese and then Korean, establishing a common trend of **English >Chinese >Korean**. This hierarchy is evident even among the top-performing models. For instance, **O3** achieves 86.71% accuracy in English but drops to 69.12% in Chinese and 67.83% in Korean, resulting in a cross-lingual gap exceeding 18 percentage points. Likewise, **GPT-4.1** obtains 39.89% in English, while only reaching 27.87% in Chinese and 25.75% in Korean. Such discrepancies illustrate a persistent and systemic bias favoring English across diverse LLM architectures.

In-depth Discussion. We identify three key, interconnected factors behind the observed performance gap. First, the **imbalance in pre-training data** skews model capabilities toward English, as most LLMs are trained on corpora rich in high-quality English text. Second, the **tokenizer effect** hinders CJK processing—tokenizers like BPE or SentencePiece, optimized for alphabetic scripts, often fragment meaningful CJK units into subwords, disrupting structural and semantic learning.

Third, the challenge is compounded by the **linguistic complexity** of CJK languages. Korean's featural syllabic system and Chinese's logographic structure require holistic modeling of sub-character components, such as jamo or radicals. Current models lack the capacity to fully capture these features, leading to degraded performance, as seen in the Korean and Chinese results in Table 2. Overcoming these issues calls for tokenizer innovations and more balanced, linguistically diverse pretraining.

4.4 Effect of Model Scale

Figure 4 shows the relationship between model size and overall TASE accuracy across three families: **Qwen3**, **Qwen2.5-Instruct**, and **Yi-1.5**. While performance generally increases with the number of parameters, the magnitude of this improvement varies significantly across different model series. Notably, **Qwen3** models consistently outperform their Qwen2.5 and Yi-1.5 counterparts—even at

Model	Structural Understanding					Token Awareness					Avg.
	COMPC	COMPM	DOT	RIDL	VAR	FREQ	LENOP	DIFF	SORT	REORD	
Human	98.17	97.56	95.61	46.33	85.94	89.06	98.17	92.89	96.06	92.67	89.24
o3	75.47	**85.17**	**26.95**	**52.17**	**48.87**	96.37	**93.67**	**55.87**	**89.83**	**31.60**	**65.60**
deepseek-r1	75.03	71.70	18.70	38.90	46.63	94.60	52.17	24.10	77.50	2.87	50.22
gemini-2.5-pro	**77.69**	70.08	23.66	28.20	42.73	**98.90**	37.58	32.27	77.30	10.97	49.94
gemini-2.5-flash	64.89	58.72	8.52	20.60	34.23	83.00	46.15	23.50	67.40	6.27	41.33
claude-opus-4	56.37	52.62	19.16	20.53	26.50	81.40	28.65	30.13	42.70	3.73	36.18
deepseek-v3	71.30	59.93	9.84	22.87	30.43	63.13	22.88	20.70	34.23	1.23	33.66
grok-3	66.73	55.75	6.22	23.93	23.53	62.93	30.83	11.77	28.33	0.50	31.05
qwen3-32b	56.75	37.22	9.06	16.33	15.47	61.57	32.93	18.03	47.70	4.63	29.97
qwq-32b	50.27	30.77	12.33	16.03	15.37	54.80	31.97	19.43	44.30	11.73	28.70
qwen3-14b	55.03	32.38	5.16	12.30	13.13	62.27	33.10	20.60	46.40	3.27	28.36
gpt-4.1	49.29	52.77	5.90	22.47	24.03	58.17	20.27	16.03	24.70	0.17	27.38
claude-sonnet-4	41.90	43.97	19.06	13.73	23.70	54.93	15.25	29.43	29.93	1.20	27.31
gpt-4o	40.52	52.43	4.51	14.17	23.80	45.33	18.58	21.63	24.40	0.30	24.57
o1-mini	46.33	41.87	8.36	14.63	18.00	41.10	16.65	26.23	31.10	0.93	24.52
qwen3-8b	52.47	27.02	5.43	6.53	8.20	56.63	30.97	16.63	39.43	1.67	24.50
claude-3-sonnet	38.05	38.87	13.72	11.87	18.40	49.10	14.65	23.20	25.80	1.03	23.47
qwen-max	48.01	49.13	6.14	12.63	13.17	46.33	11.73	17.23	29.27	0.50	23.41
qwen-turbo	44.88	46.73	5.95	11.83	12.27	44.80	9.05	16.03	28.00	0.37	21.99
qwen3-4b	46.05	21.12	3.68	4.03	6.17	53.13	26.88	16.63	39.50	1.73	21.89
qwen2.5-14b-grpo	32.25	26.23	11.52	4.40	3.57	49.90	27.22	22.60	25.67	0.17	20.35
doubao-pro-32k	30.16	45.27	4.67	16.87	21.37	50.47	2.98	4.77	21.53	0.23	19.83
qwen-plus	33.06	31.20	4.86	6.40	5.80	38.10	9.68	18.03	27.57	0.30	17.50
doubao-lite	23.76	36.55	3.96	17.03	17.40	26.87	3.72	16.10	17.27	0.03	16.27
qwen2.5-72b	26.30	26.02	4.51	5.30	4.70	39.20	3.70	25.63	23.97	0.13	15.95
llama-3.3-70b	19.65	9.70	4.21	1.50	3.80	37.77	16.52	19.73	33.83	0.00	14.67
qwen2.5-32b	22.56	20.52	3.55	4.10	3.83	35.77	12.62	16.20	23.37	0.40	14.29
dots.llm1	42.24	33.68	3.87	11.97	9.80	21.30	6.45	2.13	5.90	0.07	13.74
claude-3-haiku	27.69	21.27	2.47	7.80	5.57	32.03	10.47	7.03	20.17	0.30	13.48
gpt-3.5-turbo	26.72	23.02	2.82	6.53	6.00	26.20	10.18	8.30	17.73	0.10	12.76
qwen2.5-14b	20.57	17.25	3.18	3.47	4.03	23.90	4.87	19.37	19.93	0.17	11.67
yi-1.5-34b	17.21	11.08	4.58	1.17	0.80	14.80	9.42	4.23	10.57	0.03	7.39

Table 1: Evaluation Results of Token Awareness and Structural Understanding (%)

smaller scales—indicating that architecture design and training data quality play a more critical role than sheer scale.

The **Yi-1.5** series shows only marginal gains despite a nearly sixfold increase in size, while **Qwen2.5-Instruct** exhibits smoother scaling but still lags behind Qwen3. These patterns suggest that beyond a certain point, increasing model size alone is insufficient for improving fine-grained reasoning; the performance ceiling is largely dictated by pretraining methodology and inductive biases encoded in the model architecture.

4.5 The Effect of GRPO's Fine-tuning

The **qwen2.5-14b-grpo** model, fine-tuned via the GRPO method, showcases the efficacy of targeted fine-tuning in addressing fine-grained reasoning gaps. As shown in Table 3, it nearly doubles the average TASE score of the base **Qwen2.5-14B-instruct**, achieving over threefold accuracy gains on tasks like `LENOP`. Despite its smaller 14B size, it even surpasses larger models (e.g., **qwen2.5-32b**, **qwen2.5-72b**), with an average score of 20.40%, particularly excelling in the **Awareness** dimension. These results underscore the strength of our synthetic data pipeline and

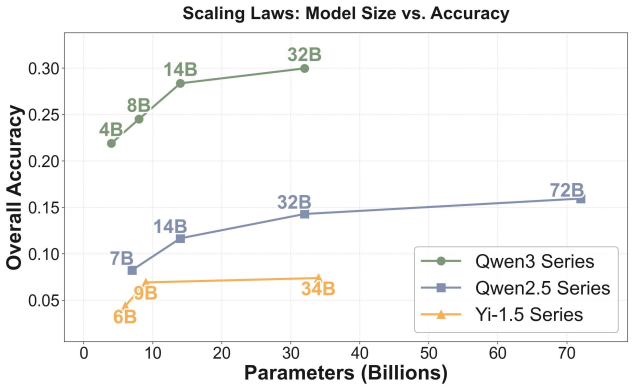

Figure 4: Effect of model size on TASE accuracy.

fine-tuning strategy in endowing smaller models with specialized capabilities. Nonetheless, average performance still trails top-tier models like **qwen-max** and **qwen-turbo**, suggesting that post-hoc fine-tuning cannot fully offset limitations from pretraining or model scale.

Reward Function Design. GRPO exhibits strong gener-

Model	Chinese	English	Korean	Avg.
Human	87.54	91.45	91.71	90.23
o3	**69.12**	**86.71**	**67.83**	**74.55**
deepseek-r1	59.71	60.10	48.32	56.04
gemini-2.5-pro	52.52	58.39	51.34	54.08
gemini-2.5-flash	44.50	51.10	46.35	47.31
claude-opus-4	36.38	45.03	36.19	39.20
deepseek-v3	45.81	39.07	27.57	37.48
grok-3	32.31	48.14	25.96	35.47
Qwen3-32B	32.53	47.68	16.76	32.32
gpt-4.1	27.87	39.89	25.75	31.17
Qwen3-14B	32.13	45.32	14.80	30.75
qwq-32b	28.45	45.22	17.14	30.27
claude-sonnet-4	21.28	36.48	29.30	29.02
gpt-4o	27.98	33.85	24.18	28.67
o1-mini	23.50	35.41	22.15	27.02
Qwen3-8B	28.28	40.04	11.19	26.50
qwen-max	32.39	30.49	15.69	26.19
claude-3-sonnet	18.18	32.11	26.08	25.45
qwen-turbo	31.65	26.92	15.03	24.53
Qwen3-4B	21.99	38.99	8.89	23.29
qwen2.5-14b-grpo	18.26	32.91	15.97	22.38
doubao-pro-32k	33.16	18.41	15.28	22.28
qwen-plus	21.88	24.13	11.36	19.12
doubao-lite	29.85	14.26	10.65	18.25
qwen2.5-72b	17.24	20.20	13.49	16.98
dots.llm1	22.71	16.21	8.35	15.76
qwen2.5-32b	14.26	22.39	10.21	15.62
llama-3.3-70b	6.97	27.71	10.78	15.15
claude-3-haiku	10.24	21.70	13.12	15.02
gpt-3.5-turbo	9.49	21.16	12.77	14.48
qwen2.5-14b	13.88	14.58	8.78	12.41
Yi-1.5-34B	5.11	13.65	5.41	8.06

Table 2: Performance comparison on Chinese, English, and Korean tasks (%).

ality and adaptability. Even with a coarse-grained reward aligned with evaluation (i.e., binary signals), it substantially boosts performance on structural and awareness tasks e.g., † **qwen2.5-14b-grpo** improves average score from 11.72% to 16.08%. Incorporating a fine-grained reward capturing subtle quality differences further enhances performance, with * **qwen2.5-14b-grpo** reaching 20.40%. Thus, while fine-grained rewards are not essential for GRPO to be effective, they amplify performance by increasing the model's sensitivity to subtle distinctions, enabling more precise fine-grained reasoning specialization, with specific training differences provided in the **appendix**.

4.6 The Effect of Chain-of-Thought

Chain-of-Thought(Wei et al. 2023; Kojima et al. 2023) prompting systematically improves model performance on the TASE benchmark across the board. As shown in Table 4, all evaluated models, from capable systems like **o1-mini** to smaller ones like **doubao-lite**, demonstrate a clear performance uplift when employing a CoT strategy. This suggests

Model	Structural	Awareness	Average
qwen-max	25.89	21.01	23.45
qwen-turbo	24.42	19.65	22.04
* qwen2.5-14b-grpo	15.68	25.11	20.40
qwen-plus	16.35	18.74	17.54
† qwen2.5-14b-grpo	14.77	17.40	16.08
qwen2.5-72b	13.43	18.53	15.98
qwen2.5-32b	11.00	17.67	14.33
qwen2.5-14b	9.79	13.65	11.72

Table 3: Performance comparison of different Qwen models on structural and awareness tasks (in percentages).
*: fine-grained version; †: coarse-grained version.

that prompting models to "think step-by-step" helps decompose complex fine-grained tasks into more manageable subproblems, thereby enhancing their reasoning capabilities.

Varying Degrees of Improvement. The magnitude of the performance gain from CoT varies significantly across different models. The effect is most pronounced for **o1-mini**, which sees its average accuracy skyrocket by over 50% (from 24.56% to 37.27%). Similarly, **gpt-3.5-turbo** achieves a substantial relative improvement of nearly 45% (from 12.80% to 18.45%). However, for **doubao-lite**, the gain is marginal, indicating that the effectiveness of CoT may be correlated with the model's inherent capabilities. While CoT can unlock latent reasoning skills, it cannot create abilities that are fundamentally absent in a less powerful model.

Model	CoT	Structural	Awareness	Average
o1-mini	w/o cot	25.91%	23.20%	24.56%
	w cot	37.08%	37.45%	37.27%
doubao-pro	w/o cot	23.74%	16.00%	19.87%
	w cot	25.17%	20.70%	22.94%
gpt-3.5-turbo	w/o cot	13.10%	12.50%	12.80%
	w cot	17.16%	19.75%	18.45%
doubao-lite	w/o cot	19.81%	12.80%	16.31%
	w cot	19.97%	12.90%	16.44%
qwen2.5-14B	w/o cot	9.79%	13.65%	11.72%
	w cot	12.56%	14.81%	13.69%

Table 4: Performance comparison with and without Chain-of-Thought (CoT) prompting (%).

5 Conclusion

We introduce TASE, a comprehensive cross-lingual benchmark for evaluating LLMs' fine-grained capabilities, and show that while current models excel at high-level semantics they still lag far behind humans on low-level precision tasks, especially character-structure control, strict constraint satisfaction, and cross-lingual generalization in Chinese and Korean, supporting the "tokenizer blindness" hypothesis; moreover, scaling yields diminishing returns, whereas targeted GRPO fine-tuning effectively improves fine-grained reasoning, making TASE both a diagnostic and a roadmap toward models that combine high-level intelligence with low-level precision.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Clark, J. H.; Choi, E.; Collins, M.; Garrette, D.; Kwiatkowski, T.; Nikolaev, V.; and Palomaki, J. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8: 454–470.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Conneau, A.; Lample, G.; Rinott, R.; Williams, A.; Bowman, S. R.; Schwenk, H.; and Stoyanov, V. 2018. XNLI: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Efrat, A.; Honovich, O.; and Levy, O. 2022. Lmentry: A language model benchmark of elementary language tasks. *arXiv preprint arXiv:2211.02069*.

Fu, T.; Ferrando, R.; Conde, J.; Arriaga, C.; and Reviriego, P. 2023. Why Do Large Language Models (LLMs) Struggle to Count Letters? *CoRR*, abs/2412.18626.

Hiraoka, T.; and Inui, K. 2025. Spelling-out is not Straightforward: LLMs' Capability of Tokenization from Token to Characters. arXiv:2506.10641.

Hua, T.; Hua, H.; Xiang, V.; Klieger, B.; Truong, S. T.; Liang, W.; Sun, F.-Y.; and Haber, N. 2025. ResearchCodeBench: Benchmarking LLMs on Implementing Novel Machine Learning Research Code. arXiv:2506.02314.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916.

Kostikova, A.; Wang, Z.; Bajri, D.; Pütz, O.; Paaßen, B.; and Eger, S. 2025. LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models. arXiv:2505.19240.

Lai, V. D.; Van Nguyen, C.; Ngo, N. T.; Nguyen, T.; Dernoncourt, F.; Rossi, R. A.; and Nguyen, T. H. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Liang, Y.; Duan, N.; Gong, Y.; Wu, N.; Guo, F.; Qi, W.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; et al. 2020. XGLUE: A new benchmark dataset for cross-lingual pretraining, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shibata, Y.; Kida, T.; Fukamachi, S.; Takeda, M.; Shinohara, A.; Shinohara, T.; and Arikawa, S. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching.

Singh, S.; Romanou, A.; Fourrier, C.; Adelani, D. I.; Ngui, J. G.; Vila-Suero, D.; Limkonchotiwat, P.; Marchisio, K.; Leong, W. Q.; Susanto, Y.; et al. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Wan, Z.; Wang, X.; Liu, C.; Alam, S.; Zheng, Y.; Liu, J.; Qu, Z.; Yan, S.; Zhu, Y.; Zhang, Q.; Chowdhury, M.; and Zhang, M. 2024. Efficient Large Language Models: A Survey. arXiv:2312.03863.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, B.; Liu, Z.; Huang, X.; Jiao, F.; Ding, Y.; Aw, A.; and Chen, N. F. 2023. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Wang, D.; Li, Y.; Jiang, J.; Ding, Z.; Luo, Z.; Jiang, G.; Liang, J.; and Yang, D. 2025. Tokenization Matters! Degrading Large Language Models through Challenging Their Tokenization. arXiv:2405.17067.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Xu, N.; and Ma, X. 2025. LLM The Genius Paradox: A Linguistic and Math Expert's Struggle with Simple Word-based Counting Problems. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 3344–3370. Originally released as arXiv:2410.14166 (Oct 2024).

Yehudai, G.; Kaplan, H.; Ghandeharioun, A.; Geva, M.; and Globerson, A. 2024. When Can Transformers Count to n? *arXiv preprint arXiv:2407.15160*.

Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.