# The Visual Faithfulness Paradox: Scaling Vision–Language Models Degrades Glyph Recognition in Logographic Scripts

**Anonymous ACL submission**

## Abstract

We identify a *Visual Faithfulness Paradox* in scaling Vision–Language Models (VLMs): although larger models improve aggregate OCR metrics, they can systematically degrade glyph-level accuracy for non-alphabetic scripts. In controlled trilingual OCR experiments using InternVL3 (1B–14B), English performance increases monotonically. Chinese follows a non-monotonic scaling curve: small models suffer visual collapse (1B/2B), the 8B model achieves the best visual–language balance, and the 14B model drifts into language-prior–driven glyph substitution. Japanese exhibits an intermediate pattern, while visually similar kana generate persistent variability. A fine-grained evaluation on a mixed-script Chinese phrase confirms the shift: the 14B model's perfect-match rate falls to 25% (vs. 54% at 8B), while its semantic-deviation error rate is 7.28x higher. We explain these effects using a *Visual Signal-to-Noise Ratio* (VSNR) account: beyond a script-dependent threshold, strengthening language priors override ambiguous visual evidence during decoding. These results expose a fundamental trade-off between linguistic fluency and visual faithfulness, challenging the assumption that "bigger is always better". Future VLMs must explicitly reinforce glyph-structure perception and coordinate the glyph stream with the knowledge stream.

## 1 Introduction

Recent advances in Vision–Language Models (VLMs) have driven substantial progress in document understanding, optical character recognition (OCR), and multimodal reasoning. To accommodate high-resolution inputs and long visual–textual contexts, recent research has increasingly emphasized *visual token efficiency*. Representative systems explore aggressive visual compression to enable long-context OCR under constrained activation budgets (Wei et al., 2025), while others pursue native multimodal pre-training with flexible visual position encoding (Zhu et al., 2025) and dynamic resolution strategies with multimodal rotary position encoding (Bai et al., 2025). Collectively, these approaches aim to reduce computational cost while preserving fine-grained visual perception.

However, as VLMs scale in parameter size and architectural capacity, an underexplored tension emerges. The dominant inference regime gradually shifts from vision-grounded perceptual decoding toward language-prior-driven semantic completion. For alphabetic scripts such as English, where visual form and lexical identity are weakly coupled, stronger language modeling typically yields monotonic OCR improvements. In contrast, for logographic writing systems such as CJK (Chinese, Japanese, and Korean), where visual structure is tightly bound to character identity, this shift can induce systematic degradation in visual faithfulness.

This paper formalizes this phenomenon as the *Visual Faithfulness Paradox*: as VLMs become larger and more linguistically competent, their ability to faithfully transcribe visually grounded glyphs, especially complex and isolated characters such as "漢" or highly similar kana symbols like"フ" vs. "プ", can systematically deteriorate, exhibiting non-monotonic or even reversed scaling behavior.

This paradox reflects a broader pattern of inductive bias in large models. It relates to *the paradox of poetic intent* observed in machine translation, where models achieve high surface-level scores while failing to preserve deeper semantic or cultural structure (Weigang and Brom, 2025). Complementarily, Six-Writings Pictophonetic Coding (SWPC) theory argues that Chinese characters require explicit modeling of graphic–phonetic–semantic structure, rather than being treated as generic visual textures (Weigang et al., 2024). From a physical perspective, Hanzi character image threshold theory further shows that common patch-based resizing strategies can cause stroke fusion and entropy collapse for high-density characters, produc-

| Language | Orthographic Structure | Visual Discriminability Profile | Dominant Error Mode | Severity |
|---|---|---|---|---|
| *English* | Alphabetic script with high inter-letter shape distinctiveness | High visual discriminability; clear shape contrasts among letters mean visual evidence is dominant and language priors are mostly supportive. | Spelling or punctuation errors | Minor: primarily detail-level inaccuracies |
| *Japanese* | Mixed system: Kanji (2,136 – semantic logographs) + Kana (92 – simple, visually similar syllabary) | Kana similarity induces visual ambiguity (e.g., "フ" vs. "プ"); kanji are stable, but kana sequences depend more strongly on language priors. | Prior–driven substitutions and insertions in visually ambiguous kana segments (e.g., "フリ–" → "フリ–カ / フリ–ワ"). | Moderate: localized hallucinations concentrated in kana |
| *Chinese* | Logographic script with high stroke density (many characters in 16px >12 strokes, e.g., 漢) | Visual overload at common UI scales; complex characters exceed perceptual thresholds, degrading glyph features. | *Semantic-deviation substitutions* where visually degraded glyphs are replaced by linguistically plausible characters (e.g., (汉漢→汉汉, 汉漢→ 汉語). | Severe: systematic loss of glyph-level identity |

Table 1: Cross-linguistic contrasts in visual discriminability and dominant OCR error modes.

ing intrinsically ambiguous visual signals prior to decoding (Weigang et al., 2025).

To systematically investigate the *Visual Faithfulness Paradox*, we conduct controlled OCR experiments on aligned English, Chinese, and Japanese Wikipedia screenshots of the *Isotopes of Hydrogen* page[1] using the InternVL3 model family (1B, 2B, 8B, and 14B) (Zhu et al., 2025; Lu et al., 2025). All configurations are evaluated under repeated stochastic sampling to ensure robustness and to disentangle visual effects from decoding variance. The data, code, and detailed replication instructions are available in our GitHub repository[2].

The cross-linguistic scaling behavior (Table 1) diverges sharply. For English, OCR accuracy improves monotonically as model capacity increases, with error rates falling from roughly 0.57 (1B) to 0.04 (14B). This reflects the strong visual distinctiveness of alphabetic letters and the supportive role of language priors. Japanese, with its mixed kana–kanji system, displays a more nuanced pattern: kanji recognition remains stable across scales, whereas visually subtle kana sequences, which exhibit weak visual distinctiveness and weaker lexical constraints.

Chinese presents the starkest case. Although aggregate OCR accuracy appears to improve steadily with scale, fine-grained inspection under simplified–traditional coexistence reveals severe structural distortions at large model sizes. As illustrated in Figure 1, sequence-level OCR accuracy increases monotonically, yet glyph-level visual error follows a U-shaped trajectory, with the *perfect-match error rate* reaching its minimum at 8B (46%) but rising sharply again at 14B (75%). This indicates that, when visual discriminability approaches the perceptual threshold of the encoder, increasingly strong language priors override ambiguous visual evidence.

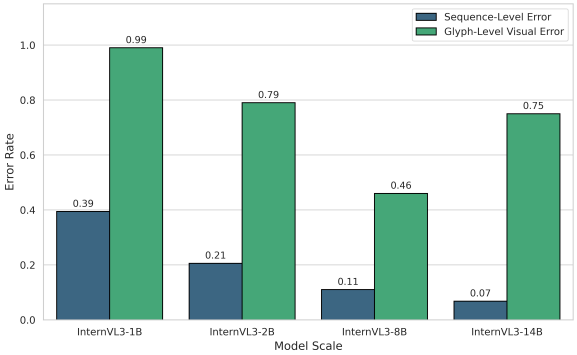



Figure 1: Sequence-level OCR accuracy improves monotonically with model scale, whereas glyph-level visual error on Chinese follows a U-shaped trajectory.

We argue that this paradoxical scaling behaviour arises from four interacting factors: (i) efficiency-oriented patch compression that erodes stroke-level detail near perceptual limits; (ii) the lack of explicit modeling of glyph-internal structure; (iii) scale-dependent amplification of language priors during decoding; and (iv) next-token likelihood objectives that favour statistically fluent continuations over visually faithful reproduction.

Our contributions are threefold. First, we empirically establish a *Visual Faithfulness Paradox* in multilingual OCR: while English improves monotonically with scale, Chinese exhibits a *U-shaped glyph error curve*, where the 8B model achieves the best visual–language balance but the 14B model increasingly substitutes visually plausible yet in-

[1]Page URL: https://en.wikipedia.org/wiki/Isotopes_of_hydrogen

[2]https://github.com/will-be-released-after-review

correct characters. Second, we show that Japanese occupies an *intermediate regime*: kanji remain largely stable, whereas visually similar kana continue to show prior-driven variability rather than catastrophic collapse. Third, we provide a *cross-layer explanation* linking these behaviours to visual signal-to-noise limits and language-prior dominance, clarifying why larger VLMs may become more "intelligent" yet less visually faithful. We situate these findings within the broader development of structure-aware OCR frameworks such as LLM-OCR-SWPC (Weigang and Costa, 2026).

## 2 Related Work

As this section reviews, mainstream VLM and OCR approaches are largely evaluated on alphabetic scripts (primarily English), overlooking the fundamental visual and structural differences of logographic systems such as CJK. This gap motivates the present study, which examines how efficiency-oriented designs interact with character-level visual fidelity in ideographic scripts.

### 2.1 Efficiency-Oriented Visual Token Optimization

A number of recent VLMs have introduced novel mechanisms to improve visual token efficiency. DeepSeek-OCR employs a deep visual encoder to aggressively compress text-dense images, achieving up to 20× token reduction while maintaining long-context OCR capability (Wei et al., 2025). This line of work demonstrates the feasibility of large-scale compression for document images, though its impact on fine-grained glyph structure remains underexplored. The InternVL series emphasizes scalable multimodal alignment through native multimodal pre-training and Variable Visual Position Encoding (V2PE), enabling improved handling of long sequences and multiple images (Chen et al., 2024b; Zhu et al., 2025; Lu et al., 2025).

Qwen2-VL introduces a *Naive Dynamic Resolution* mechanism, dynamically allocating visual tokens based on image content, together with Multimodal Rotary Position Embedding (M-RoPE) to unify positional modeling across text, image, and video modalities (Wang et al., 2024; Bai et al., 2025). This content-aware allocation represents a typical efficiency-driven approach.

Other models pursue more extreme reductions (Liu and Qiu, 2025). LLaVA-OneVision compresses each image frame into a single visual token via modality pre-fusion, substantially shortening input sequences for video understanding (Zhang et al., 2025). Lightweight models such as Monkey prioritize inference speed and deployment under resource constraints by aggressively limiting visual tokens (Chen et al., 2024a). From a different perspective, Donut reformulates document understanding as an OCR-free encoder–decoder problem, bypassing explicit OCR tokenization altogether (Kim et al., 2022).

Collectively, these approaches share a common premise: reducing the number or complexity of visual tokens generally leads to stable or improved downstream performance. This assumption is largely supported by evaluations on alphabetic-script benchmarks such as TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), and ChartQA (Masry et al., 2022).

### 2.2 Limitations: Neglecting Script Heterogeneity

Despite their empirical success, existing efficiency-oriented designs in VLMs exhibit a fundamental limitation: they insufficiently account for heterogeneity across writing systems, particularly in how visual information supports glyph-level identity.

First, current evaluation benchmarks are overwhelmingly alphabet-centric. English-dominated datasets are characterized by relatively simple glyph shapes and a loose coupling between visual form and semantic identity (Singh et al., 2019; Mathew et al., 2021; Masry et al., 2022). Under such conditions, losses in local visual detail caused by token compression or resolution reduction can often be compensated by strong language priors. This bias reinforces the widespread assumption that aggressive visual compression is largely benign, or even desirable, for multimodal OCR (Wang et al., 2024; Zhang et al., 2025). However, such conclusions are drawn primarily from alphabetic scripts and do not necessarily generalize beyond them.

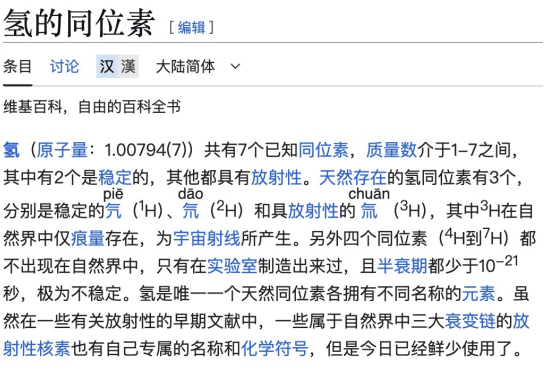
Second, mainstream visual encoders remain largely script-agnostic. Patch-based and dynamically compressed representations do not differentiate between low-complexity alphabetic characters and high-density glyphs found in logographic or mixed writing systems (Chen et al., 2024b; Wang et al., 2024; Wei et al., 2025). The research on *CJK character image processing threshold* demonstrates that once stroke density exceeds a pixel-level limit, irreversible stroke fusion occurs, result-

ing in abrupt increases in structural entropy and ambiguity (Weigang et al., 2025). Uniform patchification or compression strategies therefore introduce systematic structural noise when applied to dense glyphs, even if they remain effective for alphabetic text (Weigang et al., 2024). Mixed scripts such as Japanese with kana and kanji, partially mitigate this effect, but do not eliminate it.

Third, existing evaluations rarely assess *visual faithfulness* at the glyph level (Li et al., 2025). In ideographic and mixed scripts, models can produce semantically plausible outputs by relying on language priors, even when the visually grounded character recognition is incorrect, for example, through substitutions between visually similar or script-related characters. Such glyph-level hallucinations are largely invisible under conventional task-level accuracy or sequence similarity metrics, masking systematic failures of visual grounding.

Finally, current architectures lack explicit mechanisms for balancing language priors against character-structure evidence (Weigang and Brom, 2025). Large language decoders are optimized for next-token prediction and naturally favor high-frequency lexical patterns. When visual signals are degraded or ambiguous, this bias can override correct but low-probability glyph realizations (Wei et al., 2025; Zhu et al., 2025). Existing VLMs do not incorporate intermediate representations that explicitly encode character structure, such as stroke- or component-aware features, leaving visual evidence underconstrained during decoding (Weigang et al., 2024). As model scale increases, this imbalance can intensify rather than diminish.

This work challenges the prevailing assumption that improvements in visual token efficiency and model scale generalize uniformly across writing systems. We show that for glyph-dense and structurally complex scripts, efficiency-oriented design choices can induce a non-monotonic trade-off between model scale and visual fidelity, giving rise to the *Visual Faithfulness Paradox*. By foregrounding script heterogeneity and glyph-level faithfulness as first-class concerns, our analysis positions visual fidelity as an independent and necessary evaluation dimension for multilingual, OCR-oriented VLMs. Rather than opposing efficiency or scale, this study highlights the need for structure-aware constraints that ensure visual evidence remains competitive with language priors as models grow larger.


氢的同位素 [编辑]

条目 讨论 汉 漢 大陆简体

维基百科，自由的百科全书

氢（原子量：1.00794(7)）共有7个已知同位素，质量数介于1–7之间，其中有2个是稳定的，其他都具有放射性。天然存在的氢同位素有3个，分别是稳定的氕（piē）（$^{1}$H）、氘（dāo）（$^{2}$H）和具放射性的氚（chuān）（$^{3}$H），其中$^{3}$H在自然界中仅痕量存在，为宇宙射线所产生。另外四个同位素（$^{4}$H到$^{7}$H）都不出现在自然界中，只有在实验室制造出来过，且半衰期都少于$10^{-21}$秒，极为不稳定。氢是唯一一个天然同位素各拥有不同名称的元素。虽然在一些有关放射性的早期文献中，一些属于自然界中三大衰变链的放射性核素也有自己专属的名称和化学符号，但是今日已经鲜少使用了。

Figure 2: Screenshot of the Chinese Wikipedia article Isotopes of Hydrogen, illustrating the coexistence of simplified and traditional Chinese characters (e.g., “汉” vs. “漢”) in real-world documents.

## 3 Experimental Data, Metrics and Results

This section presents a systematic evaluation of the InternVL3 model family on OCR tasks across three *writing systems*: English, Japanese, and Chinese. Rather than treating these as linguistic categories, the experiments focus on their distinct *script-level visual and structural properties*. The central question is whether scaling VLMs improves visual fidelity uniformly, or whether scaling effects are fundamentally modulated by differences in glyph density, character complexity, and visual–semantic coupling that are intrinsic to each writing system, independent of linguistic content.

### 3.1 Data Preparation

Text corpora and corresponding screenshots were collected on November 30 of 2025 from aligned Wikipedia articles on *Isotopes of Hydrogen* in English, Chinese, and Japanese. The selected passage provides structurally comparable content across writing systems while preserving script-specific visual characteristics. The reference texts contain 264 English tokens, 223 Chinese tokens, and 235 Japanese tokens, respectively. For each language, the first-page screenshot of the article was used as the OCR input. The Chinese screenshot is shown in Figure 2. The purpose is not to require linguistic understanding of Chinese, but to provide a direct visualization of glyph-level complexity and stroke density faced by VLM-based OCR systems.

### 3.2 Experimental Setup: Models and Metrics

To evaluate whether scaling effects persist under realistic OCR conditions, we conduct experiments on the InternVL3 model family (1B, 2B, 8B, and

14B) using aligned Wikipedia screenshots in English, Chinese, and Japanese. For each combination of model scale and target language, OCR inference is repeated 100 times on identical screenshots under stochastic decoding. The raw OCR outputs—comprising 1,200 samples generated by four model scales on three writing systems (see our GitHub repository for the complete replication protocol). This sampling-based methodology enables robustness analysis by systematically disentangling genuine visual fidelity effects from stochastic generation variance.

OCR performance is evaluated using three complementary measures: (i) sequence similarity to ground-truth text, (ii) character error rate (ER), defined as one minus similarity, and (iii) token-length ratio, reflecting output stability under each model's native tokenizer. In addition to aggregate statistics, a targeted glyph-level diagnostic is conducted on a mixed simplified–traditional Chinese phrase to expose structural hallucinations that may be masked by sequence-level metrics.

We adopt the Ratcliff–Obershelp similarity measure $S(T_{gt}, T_{gen}) \in [0, 1]$ (Ratcliff, 1981), and define the error rate as:

$$ER = 1 - S(T_{gt}, T_{gen}). \quad (1)$$

Token efficiency is quantified as:

$$R_{\text{token}} = \frac{N_{\text{gen}}}{N_{\text{gt}}}, \quad (2)$$

where $N_{\text{gen}}$ and $N_{\text{gt}}$ denote generated and ground-truth token counts, respectively. This Monte Carlo-style evaluation yields full error distributions rather than single-point estimates, allowing direct comparison of stability and scaling behavior across writing systems.

### 3.3 Scaling Behaviors Across Writing Systems

Figure 3 illustrates the performance distribution (e.g., sequence similarity) across writing systems and model scales, revealing starkly divergent scaling profiles. While aggregate OCR metrics improve with model size, the trajectory and underlying mechanisms differ fundamentally across scripts.

#### 3.3.1 English: Monotonic Scaling as an Aligned Baseline

Table 2 reports OCR performance on English text. As model size increases from 1B to 14B, all evaluation metrics exhibit strictly monotonic improvement. The mean predicted token length converges from 279.81 to 268.02 (reference: 264), while the standard deviation decreases from 287.19 to 11.34, indicating highly stable outputs. Sequence similarity increases from 0.43 to 0.96, with the corresponding error rate dropping from 0.57 to 0.04.

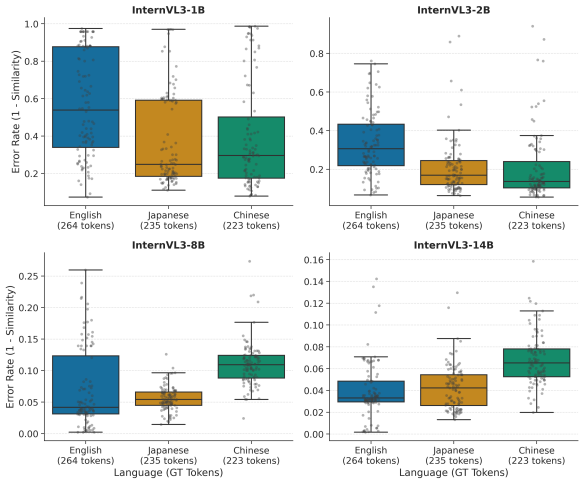



Figure 3: Error-rate distributions across writing systems and model scales for InternVL3. While overall OCR robustness improves with scale, the stability and qualitative behavior of gains differ markedly across scripts.

| Model | Tokens ($\mu \pm \sigma$) | Similarity | Error Rate |
|---|---|---|---|
| 1B | 279.81±287.19 | 0.43 | 0.57 |
| 2B | 326.25±102.40 | 0.67 | 0.33 |
| 8B | 269.00±20.73 | 0.93 | 0.07 |
| 14B | 268.02±11.34 | 0.96 | 0.04 |

Table 2: OCR Performance on English Text.

These results establish English as an *ideal baseline* for VLM-based OCR: in alphabetic writing systems, visual transcription and language modeling objectives are strongly aligned (Table 1). Increased model capacity simultaneously enhances perceptual stability and linguistic accuracy, yielding predictable and monotonic scaling behavior.

#### 3.3.2 Chinese: Full Manifestation of the Visual Faithfulness Paradox

Chinese OCR results provide the definitive evidence for the Visual Faithfulness Paradox, exposing a fundamental contradiction between macro-level fluency and glyph-level visual fidelity. As shown in Figure 3, aggregate indicators suggest continuous improvement from 1B to 14B, including higher sequence similarity (up to 0.93) and lower overall error rates (down to 0.07), approaching English-level performance.

However, direct inspection of model outputs reveals substantial glyph-level failures that are in-

| Error Type | 2B | 8B | 14B |
|---|---|---|---|
| Semantic / Glyph Collapse | 5/5 | 0/5 | 0/5 |
| Empty output | 0/5 | 0/5 | 1/5 |
| Simplified–Traditional Error | 1/5 | 2/5 | 5/5 |
| Glyph Confusion / Neologism | 5/5 | 4/5 | 3/5 |
| Extraneous Symbol Addition | 3/5 | 5/5 | 0/5 |
| Format Inference Error | 4/5 | 3/5 | 2/5 |
| Minor Punctuation Error | 5/5 | 5/5 | 5/5 |

Table 3: Error type distribution across model scales on Chinese OCR (5 samples per model).

| Metric (%) | 2B | 8B | 14B |
|---|---|---|---|
| Avg. Edit Similarity | 85.1 | **92.9** | 82.2 |
| Perfect Match Rate | 30.0 | **50.0** | 20.0 |
| High Similarity (≥90%) | 70.0 | **95.0** | 65.0 |
| Complete Failure | 0.0 | 0.0 | **5.0** |
| S–T Conversion Error | 20.0 | 15.0 | **60.0** |

Obs.: A paired-sample t-test confirms worse performance than the 8B model ($t(19) = 2.94, p < 0.01$)

Table 4: Performance comparison on the mixed-script Chinese phrase (N=20).

visible to surface metrics. We therefore conduct following two complementary analyses.

**(1) Manual Analysis of Five Representative Outputs.** Consistent with the Japanese analysis, manual inspection reveals scale-dependent qualitative error regimes (see Table 3 and Table 4).

*2B: Severe visual insufficiency.* Errors are dominated by glyph-structure collapse and semantic confusion, including malformed or invented characters (e.g., "氣"), confusion among hydrogen isotopes ("氕/ 氘/ 氚"), and unstable recognition of scientific notation.

*8B: Optimal vision–language balance.* Major glyph collapse disappears and factual correctness largely recovers. New error peaks emerge in the form of extraneous symbol insertion (e.g., ∼) and occasional simplified–traditional inconsistencies. These errors reflect transitional instability as improved visual perception begins to capture interface-level symbols without fully stabilized decoding.

*14B: Language-prior-dominated substitution.* Severe visual noise vanishes, but errors shift toward systematic, knowledge-driven glyph replacement. Rare but visually faithful characters are overridden by more frequent alternatives (e.g., "氚" misrecognized as "氫"). Simplified–traditional conversion errors reach 100%, indicating that strong language priors normalize visual input toward dominant lexical distributions.

**(2) Targeted Phrase Analysis over 20 Samples.** We further analyze 20 repeated outputs of the critical phrase "条目讨论汉漢 大陆简体," focusing on edit-distance similarity and exact-match behavior.

Contrary to aggregate trends, glyph-level accuracy follows a clear U-shaped trajectory. The 8B model achieves peak performance, while the 14B model exhibits pronounced degradation. As shown in Figure 1, the perfect match error rate for the character "漢" rises to 75% at 14B, compared to only 46% at 8B.

The dominant failures are systematic, structured substitutions (e.g., "汉汉", "汉语")—glyph-level hallucinations where strong language priors override visual evidence. This degradation at the 14B scale is statistically significant: a paired-sample t-test confirms worse performance than the 8B model ($t(19) = 2.94, p < 0.01$). The original outputs for the 20 samples, along with the calculation and statistics of their edit-distance similarity to the reference text, are provided in Appendix A Table 6.

### 3.3.3 Japanese: Mixed-Regime and Partially Aligned Scaling

Japanese departs from strictly monotonic scaling, but it does not exhibit the severe glyph-level breakdown observed in Chinese. As shown in Figure 4, English OCR variance decreases smoothly with model size, whereas Japanese token-length variance remains elevated at small scales and re-emerges at 14B, indicating persistent decoding instability in kana–kanji mixtures.

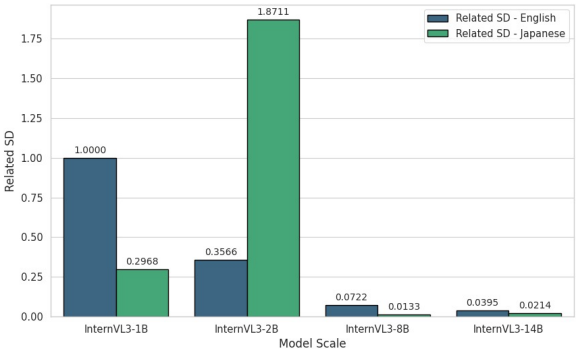



Figure 4: Relative standard deviation of predicted token length for English and Japanese OCR outputs across InternVL3 model scales, normalized to the English 1B baseline.

At the system level, the 8B model achieves the best overall balance, with near-reference length, high edit similarity, and low variance, see Table 5. The 14B model yields slightly higher average

| Metric (%) | 2B | 8B | 14B |
|---|---|---|---|
| Avg. Edit Similarity | 73.5 | **89.6** | 89.4 |
| Perfect Match Rate | 15.0 | **35.0** | 30.0 |
| High Similarity (≥90%) | 25.0 | **70.0** | 65.0 |
| Complete Failure | 0.0 | 0.0 | 0.0 |
| Language-Prior Hallucinations | 35.0 | 15.0 | **25.0** |

Obs.: No significant difference between 8B and 14B models ($t(19) = 0.35, p = 0.73$), both outperform 2B ($p < 0.001$).

Table 5: Performance comparison on the Japanese reference phrase (N=20).

similarity but noticeably larger variance, suggesting that stronger language priors begin to perturb visually grounded decoding.

To verify whether this instability reflects qualitative regime shifts, we conducted a fine-grained analysis of a canonical Japanese phrase, "出典: フリ百科事典『ウィキペディア（Wikipedia）』", comparing 20 sampled OCR outputs across scales (Table 5 and Appendix C Table 8). Both 8B and 14B significantly outperform 2B ($p < 0.001$), but there is no significant difference between 8B and 14B ($t(19) = 0.35$, $p = 0.73$). However, the error types differ. The 2B model suffers from semantic drift and fabrication due to weak visual grounding. The 8B model produces mostly faithful transcriptions with only minor formatting deviations. By contrast, the 14B model introduces isolated but systematic *language-prior substitutions*, alongside stylistic "polishing." These include knowledge-based replacements that are not visually supported in the input, illustrating the same prior–vision tension seen in Chinese, though in milder form, see Appendix C Table 8.

A key observation is that instability concentrates in *kana* sequences rather than kanji. Although kana have low stroke complexity, they are visually similar (e.g., "フ/プ/ブ/ヌ/ワ"), and thus only weakly discriminable at small scales, exactly where language priors become competitive. It yields a *mixed-regime behavior*: scaling improves robustness and fluency, but prior dominance introduces measurable variability before convergence. This pattern supports our broader claim that orthographic structure modulates the balance between visual fidelity and linguistic priors in VLM decoding.

### 3.4 Cross-Lingual Synthesis

Our cross-lingual analysis shows that the *Visual Faithfulness Paradox* is strongly moderated by orthographic structure. Alphabetic English follows a near-ideal scaling pattern, with larger models jointly improving linguistic fluency and visual accuracy. Japanese, which mixes kanji and visually subtle kana, displays moderate instability at small scales but largely stabilizes as capacity increases, reflecting only a partial tension between visual form and prior-driven decoding. Chinese, however, exhibits the most severe manifestation: glyph-level fidelity *declines* beyond a critical scale despite continued gains in aggregate OCR metrics.

This decline is tightly linked to character complexity and perceptual resolution. Learning-based OCR studies show a stroke-density threshold effect in CJK scripts, where complex glyphs become visually ambiguous at common rendering sizes (e.g., >12 strokes at 16 px) (Weigang et al., 2025). For characters such as "漢" (14 strokes), the visual encoder thus operates near or beyond its effective resolution limit, yielding weak or unstable glyph features. At moderate scales, models rely primarily on this degraded but still constraining visual signal. As scale grows, however, stronger language priors increasingly dominate decoding, completing ambiguous glyphs via lexical frequency or semantics rather than perception. What begins as helpful compensation therefore becomes a systematic source of glyph substitution, capturing the core of the paradox, see Table 1.

An optimal operating regime thus emerges for mid-scale models (e.g., ∼8B), where visual evidence remains decisive and language priors offer only limited support. More broadly, these results show that aggregate OCR metrics are insufficient: glyph-level visual fidelity, especially near perceptual complexity thresholds, must be treated as a distinct and essential evaluation dimension for multilingual VLMs, also see Appendix B Table 7.

## 4 Discussion: Visual Signal-to-Noise Evolution

To explain the non-uniform scaling behaviors observed across writing systems, we introduce *Visual Signal-to-Noise Ratio (VSNR)* as a unifying explanatory framework (Sinha et al., 2007; Hendricks et al., 2021). VSNR characterizes the internal competition between meaningful glyph-structure signals extracted by the visual encoder and noise arising from resolution limits, compression artifacts, and representational uncertainty.

VSNR is not defined at the pixel level; rather, it reflects the balance within a model's internal visual feature space between deterministic structural representations and ambiguous or unstable feature

components. As model scale increases, this balance shifts, altering the relative influence of visual evidence and language priors during multimodal OCR decoding. Conceptually, VSNR can be approximated as:

$$\text{VSNR} \approx \frac{\mathbb{E}[\|\mathbf{f}_g\|^2]}{\mathbb{E}[\|\mathbf{f}_n\|^2]}, \tag{3}$$

where $\mathbf{f}_g$ denotes glyph-correlated visual features and $\mathbf{f}_n$ denotes noise-dominated components.

*Stage A: Visual Failure (1B–2B).* At small scales, the visual encoder fails to extract stable glyph representations, yielding truncation, unreadable symbols, or visually unrelated characters. Error rates exceed 79–99%, and VSNR remains near zero, preventing reliable decoding by either visual evidence or language priors. This regime is most evident for dense logographic characters but also affects low-resolution kana strings through uncontrolled phonetic variation.

*Stage B: Visual Dominance (8B).* At intermediate scale, VSNR exceeds unity, allowing visual evidence to dominate decoding. Error rates drop to 46% for Chinese, with remaining mistakes limited to rare visual neighbor confusions. Japanese OCR likewise reaches peak stability, though residual variance persists for kana sequences with low inter-character distinctiveness, reflecting partial decoupling between visual form and lexical identity.

*Stage C: Language-Prior Dominance (14B).* Beyond a critical scale, error rates rebound sharply to 75% for Chinese. Although visual representations remain strong, decoding becomes dominated by language priors, producing systematic semantic substitutions (e.g., “漢” → “汉” or “汉语”). In Japanese, this dominance manifests differently: instead of semantic collapse, strong priors amplify phonetic smoothing, yielding clustered kana substitutions (e.g., “フリ” → “フリワ/ フリラ/ フリカ”) despite visually correct kanji.

Together, these stages reveal a nonlinear transition from visual insufficiency to visual dominance and ultimately to language-prior hegemony. The resulting failures differ by script: threshold-induced ambiguity for dense logographs and low-discriminability drift for phonetic kana. Performance degradation at large scales thus reflects not insufficient capacity, but a systematic reweighting of decision cues under scaling. This diagnosis establishes the theoretical basis for structure-aware corrective frameworks, such as LLM-OCR-SWPC (Weigang and Costa, 2026), aimed at preserving glyph-level visual faithfulness under model scaling.

## 5 Conclusion

The *Visual Faithfulness Paradox* is best understood as a continuum governed by orthographic structure: it intensifies when visual discriminability is low and linguistic constraints are strong. Thus, Chinese characters show the sharpest paradox, Japanese kana reveal moderate creative errors, while the Latin alphabet remains largely stable.

This paper demonstrates that larger VLMs may become *less* visually faithful even when their overall OCR fluency improves. Across InternVL3 (1B–14B), English performance scales monotonically as expected. In contrast, Chinese exhibits a *U-shaped* glyph outcome: the perfect-match error rate is hight at 2B (79%), lowest at 8B (46%), but rises again to 75% at 14B, reflecting a shift toward language-prior–dominated decoding in which visually ambiguous yet semantically important characters are replaced by statistically plausible alternatives (e.g., simplified for traditional forms). Japanese occupies an intermediate position: kanji remain largely stable, while visually similar kana continue to induce prior-driven variability and creative substitutions.

We interpret these effects through a *Visual Signal-to-Noise Ratio (VSNR)* lens. Scaling initially improves VSNR, but beyond a critical point language priors overpower weak visual evidence, shifting decoding from perception-driven transcription to prior-driven completion—the essence of the paradox.

Therefore, “bigger is better” does not universally hold for multilingual OCR. Robust scaling requires structure-aware designs that respect script-specific perceptual limits. This work highlights that “vision” in VLMs is not merely a pixel encoder, but a dynamic equilibrium between perceptual evidence and linguistic priors whose balance is script-dependent. Promising directions include adaptive visual resolution for dense glyphs, explicit glyph-structure anchoring (e.g., SWPC), and benchmarks that measure multilingual visual faithfulness beyond edit distance. Although our case studies focus on Chinese and Japanese, the same mechanisms are likely to extend to other scripts with low visual separability or high ligaturing complexity (e.g., Vietnamese Hán–Nôm, Arabic, or Devanagari), suggesting a broader research agenda. Restoring strong visual grounding is therefore essential for the next generation of large-scale vision–language models.

## Limitations

This study has several limitations. First, our analysis is based on a deliberately focused dataset consisting of three screenshots from English, Chinese, and Japanese Wikipedia pages. This case-study design enables controlled, fine-grained observation of glyph-level behavior (e.g., distinguishing "汉" from "漢", but necessarily limits statistical breadth and visual diversity. Second, all results are obtained from a single model family, InternVL3, evaluated across four parameter scales (1B–14B). Although the observed "Visual Faithfulness Paradox" appears internally consistent, confirming whether it generalizes across architectures, training regimes, and data sources requires broader replication. Finally, our interpretation is grounded in the visual–semantic binding properties of logographic and mixed-script systems. While the analyses of Chinese characters and Japanese kana likely extend to other visually confusable scripts, a comprehensive cross-script survey and larger-scale evaluation on varied real-world imagery remain important directions for future work.

## Ethical Considerations

This study uses only publicly available Wikipedia screenshots and model-generated text. No personal or sensitive data are included, and no content falls into harmful or offensive categories.

## Acknowledgements

We thank the developers of publicly available LLM services (ChatGPT, DeepSeek, Grok) for assistance with language refinement. We are also grateful to the OpenGVLab team for open-sourcing the InternVL3 model series, which made this controlled scaling study possible.

## References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Wei Chen, Zhiyuan Li, and Shuo Xin. 2024a. Omnivlm: A token-compressed, sub-billion-parameter vision-language model for efficient on-device inference. *arXiv preprint arXiv:2412.11475*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *2024 IEEE/CVF-CVPR)*, pages 24185–24198. IEEE.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Zongqian Li, Yinhong Liu, Yixuan Su, and Nigel Collier. 2025. Prompt compression for large language models: A survey. In *Proc. of 2025 NAACL-HLT (vol. 1)*, pages 7182–7195.

Fanfan Liu and Haibo Qiu. 2025. Context cascade compression: Exploring the upper limits of text compression. *arXiv preprint arXiv:2511.15244*.

Dongchen Lu, Yuyao Sun, Zilu Zhang, Leping Huang, Jianliang Zeng, Mao Shu, and Huo Cao. 2025. Internvl-x: Advancing and accelerating internvl series with efficient visual token compression. *arXiv preprint arXiv:2503.21307*.

Ahmed Masry, Do Long, Jie Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for document visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209.

Roger Ratcliff. 1981. A theory of order relations in perceptual matching. *Psychological review*, 88(6):552.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF-CVPR*, pages 8317–8326.

Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. 2007. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Li Weigang and Pedro Carvalho Brom. 2025. Paradox of poetic intent in back-translation: Evaluating the quality of large language models in chinese translation. *Frontiers of Information Technology & Electronic Engineering*, 26(11):2176–2203.

Li Weigang and Joao Paulo Vieira Costa. 2026. Llm-ocr-swpc: Hanzi's cognitive advantage and token efficiency in multimodal ai. In *Proceedings of 2026 IEEE 18th International Conference on Computer Research and Development*. IEEE.

Li Weigang, M. C. Marinho, and Danilo Lelin Li. 2024. Six-writings multimodal processing with pictophonetic coding to enhance chinese language models. *Frontiers of Information Technology & Electronic Engineering*, 25(1):84–105.

Li Weigang, Rafael Marconi Ramos, Pedro Carvalho Brom, and Danilo Lelin Li. 2025. Threshold study for hanzi image recognition: Defining character and component limits in chinese, japanese, and korean script processing. *International Journal of Asian Language Processing*, 35(1):2450011.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

# Appendix

## Appendix A: Complete OCR outputs and edit-distance similarity for the Chinese phrase across model scales (N=20)

## Appendix B: A Unified Theoretical Parallel between the Paradox of Poetic Intent and the Paradox of Visual Fidelity

## Appendix C: Complete OCR outputs and edit-distance similarity for the Japanese phrase across model scales (N=20)

| 2B Model | | 8B Model | | 14B Model | |
|---|---|---|---|---|---|
| **OCR Output** | **Sim.** | **OCR Output** | **Sim.** | **OCR Output** | **Sim.** |
| 条目讨论汉漢 大陆简体跑 | 0.900 | 条目讨论汉漢 大陆简体 | 1.000 | (空输出) | 0.000 |
| 条目讨论汉汉大陆简体海条目 | 0.643 | 条目讨论汉漢 大陆简体 | 0.900 | 条目讨论汉汉大陆简体 | 0.900 |
| 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉汉大陆简体 | 0.900 | 条目讨论汉漢 大陆简体chuan | 0.643 |
| 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉漢 大陆简体C | 0.900 | 条目讨论汉语大陆简体 | 0.900 |
| 题目讨论汉漢 大简体陆简体 | 0.400 | 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉汉大陆简体 | 0.900 |
| 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉汉大陆简体X | 0.800 |
| 条目讨论汉 | 0.444 | 条目讨论汉漢 大陆简体∨ | 0.900 | 条目讨论汉漢 大陆简体✓ | 0.900 |
| 条目讨论汉漢 大陆简体↓ | 0.900 | 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉漢 大陆简体 | 1.000 |
| 条目讨论汉大陆简体 | 0.900 | 条目讨论汉漢 大陆简体→ | 0.900 | 条目讨论氢漢 大陆简体∨ | 0.778 |
| 条目讨论汉漢 大陆简体↓ | 0.900 | 条目讨论汉漢 大陆简体 | 0.889 | 条目讨论汉漢 大陆简体 | 0.900 |
| 条目讨论汉漢 大简体 | 1.000 | 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉汉大陆简体↓ | 0.800 |
| 条目讨论汉大陆简体↓ | 0.889 | 条目讨论汉汉大陆简体 | 0.900 | 条目讨论汉漢 大陆简体 | 1.000 |
| 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉汉大陆简体 | 0.778 |
| 条目讨论汉漢 大陆简体↓ | 0.900 | 条目讨论汉漢 大陆简体⇄ | 0.900 | 条目讨论汉汉大陆简体 | 0.900 |
| 条目讨论汉漢 大陆简体↓ | 0.900 | 条目讨论汉漢 大陆简体 | 0.900 | 条目讨论汉语大陆简体 | 0.900 |
| 条目讨论汉漢 大 | 0.889 | 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉漢 大陆简体 | 1.000 |
| 条目讨论汉大陆简体↓ | 0.778 | 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉漢 大陆简体 | 0.900 |
| 条目讨论汉：漢 大陆简体↓ | 0.900 | 条目讨论汉演大陆简体 | 0.667 | 条目讨论汉汉大陆简体 | 0.900 |
| 栏目讨论汉大陆简体 | 0.667 | 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉漢 大陆简体 | 1.000 |
| 条目讨论汉大陆简体 | 0.900 | 条目讨论汉漢 大陆简体 | 1.000 | 条目讨论汉漢 大陆简体 | 1.000 |
| Mean ± SD | 0.851 ± 0.145 | Mean ± SD | 0.929 ± 0.020 | Mean ± SD | 0.822 ± 0.208 |
| | | Wikipedia Source Phrase: | 条目讨论汉漢 大陆简体 | | |

*Obs.:* A paired-sample $t$-test confirms the 14B model performs significantly worse than the 8B model ($t(19) = 2.94$, $p < 0.01$).

Table 6: **Appendix A**: Complete OCR outputs and edit-distance similarity for the Chinese phrase across model scales (N=20).

| **Dimension** | **Paradox of Poetic Intent** (Weigang and Brom, 2025) | **Paradox of Visual Fidelity** |
|---|---|---|
| Task Level | Sentence- and discourse-level back-translation of poetry | Character-level OCR in multimodal vision–language models |
| Observed Phenomenon | High BLEU scores coexist with loss of metaphor and cultural intent | High linguistic confidence coexists with glyph-level visual errors |
| Surface Metric Optimized | Lexical overlap (BLEU) | Token-level similarity / semantic plausibility |
| Core Paradox | Surface textual fidelity $\not\Rightarrow$ poetic understanding | Language fluency $\not\Rightarrow$ visual faithfulness |
| Typical Failure Mode | Literal reproduction with shallow interpretation | Semantic substitution overriding ambiguous visual evidence |
| Trigger Condition | Dense metaphor, cultural allusion, poetic ambiguity | High stroke density, isolated glyphs, script-level ambiguity |
| Role of Model Scaling | Larger models amplify memorization and prior-driven generation | Larger models amplify language-prior dominance over vision |
| Underlying Mechanism | Optimization toward lexical probability over interpretive depth | Language-dominant decoding under degraded visual signal |
| Unifying Explanation | **Strong model priors dominate when task-relevant signal is weak or ambiguous** | |
| Signal Perspective | Low semantic signal-to-prior ratio | Low visual signal-to-noise ratio (VSNR) |
| Theoretical Implication | Evaluation based on surface metrics is fundamentally insufficient | Visual fidelity requires independent, structure-aware assessment |
| Proposed Direction | Semantics-aware evaluation of back-translation | Structure-aware OCR with glyph-level anchoring (LLM-OCR-SWPC) (Weigang and Costa, 2026) |

Table 7: **Appendix B**: A Unified Theoretical Parallel between the Paradox of Poetic Intent and the Paradox of Visual Fidelity.

| 2B Model | | 8B Model | | 14B Model | |
|---|---|---|---|---|---|
| **OCR Output** | **Sim.** | **OCR Output** | **Sim.** | **OCR Output** | **Sim.** |
| 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 | 出典: フリカド辞典『ウィキペディア（Wikipedia）』 | 0.885 | 出典: フリ百科事典「ウィキペディア（Wikipedia）」 | 0.962 |
| 出典: フリワイド百科事典『ウィキペディア(Wikipedia)』 | 0.885 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』 | 0.962 | 出典: フリ百科事典「ウィキペディア(Wikipedia)」 | 0.962 |
| 出典: フリレス百科事典『ウィキペディア(Wikipedia)』 | 0.885 | 出典: フリ百科事典『ウィキペディア』*(Wikipedia)* | 0.769 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』 | 0.923 |
| 出典: フリワルドセンタスパディア（Wikipedia）] | 0.538 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』 | 0.923 | 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 |
| 出典: フリヒ百科事典『ウィキペディア（Wikipedia）』 | 0.846 | 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 | 出典: フリ百科事典「ウィキペディア(Wikipedia)」 | 0.962 |
| 出典: フリワルド百科事典['ウィキペディア（Wikipedia）'] | 0.731 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』 | 0.962 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』 | 0.962 |
| 出典: フリラ百科事典『ウィキペイデタ(Wikipedia)』 | 0.692 | 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 | 出典: フリ百科事典「ウィキペディア（Wikipedia）」 | 0.962 |
| 出典: フリパ百科事典『ウィキペディア(Wikipedia)』 | 0.885 | 出典: フリキ百科事典「ウィキペディア（Wikipedia）」 | 0.885 | 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 |
| 出典: フリマウンドイリア（Wikipedia） | 0.577 | 出典: フリカイ百科事典『ウィキペディア(Wikipedia)』 | 0.885 | 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 |
| 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 | 出典: フリ百科事典『ウィキペディア』(Wikipedia) | 0.885 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』 | 0.962 |
| 出自: フリリッチ百科事典『ウィキペディア（Wikipedia）』 | 0.846 | 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』() | 0.885 |
| 出典: フリリ百科事典『ウィキペディア（Wikipedia）』 | 0.923 | 出典: フリキ百科事典『ウィキペディア（Wikipedia）』 | 0.885 | 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 |
| 出典: フリワルド事典『ウィキペディア（Wikipedia）』 | 0.808 | 出典: フリキ百科事典『ウィキペディア（Wikipedia）』 | 0.885 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』 | 0.962 |
| 出典: フリラ百科事典『ウィキペディア（Wikipedia）』 | 0.923 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』 | 0.962 | 出典: フリ百科事典「ウィキペディア(Wikipedia)」 | 0.962 |
| 出典: フリワルドセ事典『ウィキペディア(Wikipedia)』 | 0.769 | 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 | 出典: フリエ百科事典「ウィキペディア(Wikipedia)」 | 0.885 |
| 出典: フリワルド·ウィキメディア.ペス（Wikipedia）[1] | 0.615 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』 | 0.962 | 出典: フリペディア「ウィキペディア(Wikipedia)」 | 0.731 |
| 出典: フリワイヤメディア（Wikipedia）[1] | 0.577 | 出典: フリビル百科事典『ウィキペディア（Wikipedia）』 | 0.885 | 出典: フリ百科事典「ウィキペディア（Wikipedia）」 | 0.962 |
| 出典: フリワ百科事典『ウィキペディア(Wikipedia)』 | 0.885 | 出典: フリゾノ学事典『ウィキペディア（Wikipedia）』 | 0.846 | 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 |
| 出典: フリ百科事典『ウィキペディア（Wikipedia）』 | 1.000 | 出典: フリ百科事典『ウィキペディア』（Wikipedia） | 0.885 | 出典: フリモ百科事典「ウィキペディア(Wikipedia)」 | 0.885 |
| 出典: フリワルド事典『ウィキペディア（Wikipedia〕』 | 0.808 | 出典: フリ百科事典『ウィキペディア(Wikipedia)』 | 0.962 | 出典: フリ百科事典「ウィキペディア(Wikipedia)」 | 0.962 |
| Mean ± SD | 0.735 ± 0.184 | Mean ± SD | 0.896 ± 0.071 | Mean ± SD | 0.894 ± 0.084 |

Wikipedia Source Phrase: 出典: フリ百科事典『ウィキペディア（Wikipedia）』

Paired-sample $t$-tests show no significant difference between 8B and 14B models ($t(19) = 0.35$, $p = 0.73$), both outperform 2B ($p < 0.001$).

Table 8: **Appendix C**: Complete OCR outputs and edit-distance similarity for the Japanese phrase across model scales (N=20).