Visible Yet Unreadable: A Systematic Blind Spot of Vision–Language Models Across Writing Systems

AI
OpenAI, Anthropic, and Google

Jie Zhang*
CFAR and IHPC, A*STAR

Ting Xu
National University of Singapore

Gelei Deng
Nanyang Technological University

Runyi Hu
Nanyang Technological University

Han Qiu
Tsinghua University

Tianwei Zhang
Nanyang Technological University

Qing Guo†
Nankai University

Ivor Tsang
CFAR and IHPC, A*STAR

Abstract

Writing is a universal cultural technology that reuses vision for symbolic communication. Humans display striking resilience: we readily recognize words even when characters are fragmented, fused, or partially occluded. This paper investigates whether state-of-the-art vision–language models (VLMs) share this resilience. We construct two psychophysics-inspired benchmarks across distinct writing systems—Chinese logographs and English alphabetic words—by splicing, recombining, and overlaying glyphs to yield "visible-but-unreadable" stimuli for models while remaining legible to humans. Despite strong performance on clean text, contemporary VLMs show a severe drop under these perturbations, frequently producing unrelated or incoherent outputs. The pattern suggests a structural limitation: models heavily leverage generic visual invariances but under-rely on compositional priors needed for robust literacy. We release stimuli generation code, prompts, and evaluation protocols to facilitate transparent replication and follow-up work. Our findings motivate architectures and training strategies that encode symbol segmentation, composition, and binding across scripts, and they delineate concrete challenges for deploying multimodal systems in education, accessibility, cultural heritage, and security.

1 Introduction

Reading is one of humanity's most powerful cultural inventions [1, 2]. By mapping arbitrary visual marks onto symbolic meaning, writing systems enable the storage, transmission, and accumulation of knowledge across generations. Despite their vast diversity—from logographic Chinese characters to alphabetic scripts—humans exhibit striking resilience in reading: we readily recognize text even when it is cut, overlapped, occluded, or distorted [3, 4]. This robustness likely reflects deep structural priors in human perception—the expectation that written symbols are composed of parts and follow compositional rules that support recovery from incomplete input [5].

Artificial intelligence, by contrast, often appears to read but lacks the same resilience. Vision–language models (VLMs) can transcribe rendered text, answer questions about documents, and

*`https://zjzac.github.io/`
†The corresponding author

interleave language with images [6, 7, 8]. Yet their reading ability has not been systematically probed under distortions that are trivial for humans. We ask a simple but fundamental question: *Can models read what humans can still read?* Our results indicate that, despite near-perfect performance on clean text, state-of-the-art VLMs show a *sharp degradation* when confronted with perturbed but human-readable constructs, revealing a robust cross-script gap between human and machine literacy.

To probe this gap, we design psychophysics-inspired benchmarks across two distinct writing systems. In Chinese logographs, we render a set of 100 four-character idioms and splice each character along horizontal, vertical, or diagonal axes, recombining fragments into composite glyphs. In English alphabetic words, we select 100 eight-letter words, divide them into two halves, render each half in different colors, and overlay them to form fused stimuli. These manipulations preserve human legibility while consistently reducing VLM accuracy, with models frequently producing unrelated or incoherent outputs. The observed pattern is consistent with a structural limitation: generic visual invariances learned from large-scale training [9, 10] may be insufficient for textual identifiability without stronger symbol-centric priors.

The implications extend beyond technical benchmarks. Robust machine literacy under mild perturbations matters for scientific curation of handwritten notes and historical manuscripts [11, 12], for education and accessibility in non-standard scripts and diverse reader populations [13], for cross-lingual and low-resource contexts [14], and for security-sensitive document analysis where adversaries may exploit brittleness [15]. More broadly, gaps between human and machine reading bear on the reliability of multimodal AI systems [16, 17].

Contributions. This study makes four contributions:

- **Evidence of a cross-script failure mode:** We show that state-of-the-art VLMs exhibit a marked collapse on "visible-but-unreadable" stimuli—perturbations that remain trivial for humans across logographic and alphabetic scripts.
- **Psychophysics-inspired benchmarks:** We introduce controlled perturbations for Chinese idioms and English words that preserve human readability while systematically challenging machine recognition.
- **Implications for model design:** Findings highlight a fundamental human–AI asymmetry in literacy and suggest that scaling alone may be insufficient, motivating architectures with explicit structural priors for segmentation, composition, and binding [18, 19].
- **Broader impact:** We delineate risks and opportunities for deploying robust machine reading in security, scientific curation, education, accessibility, cultural heritage, and trustworthy multimodal AI.

2 Related Work

2.1 Human Reading, Psychophysics, and Structural Priors

Human readers exhibit remarkable robustness to distortions in writing. Classic psychophysics shows that crowding, occlusion, and fragmentation often impair object recognition, yet humans can still recover meaning from incomplete text [20, 21]. This resilience has been linked to structural priors such as segmentation, binding, and morpheme- or radical-level expectations [5]. Our work situates itself in this tradition: we show that when structural boundaries of glyphs are disrupted, humans maintain readability but vision–language models (VLMs) collapse, exposing the absence of such priors in current architectures.

2.2 Multimodal VLMs and Reading Ability

Large VLMs such as CLIP [22], BLIP-2 [23], Kosmos [24], LLaVA [8], GPT-4V, and Gemini have demonstrated impressive capabilities in multimodal understanding, including text-in-image tasks. Evaluations on document QA, chart understanding, and OCR-VQA benchmarks suggest that these models can "read." However, these evaluations overwhelmingly use natural renderings of text. Our results reveal that once structural perturbations are introduced, VLMs fail systematically—even when humans find the text trivial to recognize. This highlights that current VLM "reading" is more a byproduct of visual invariance than of symbolic identifiability.

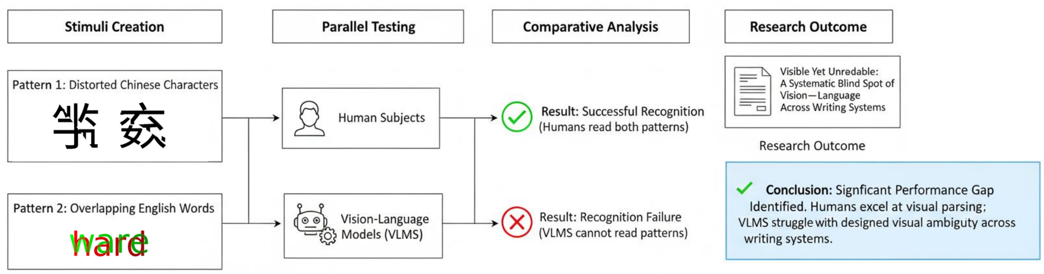

Figure 1: Overview of the experimental framework. Distorted Chinese characters and overlapping English words were created as stimuli and tested in parallel on human subjects and vision–language models (VLMs).

2.3 Psychophysics-inspired Evaluation in Machine Learning

Recent work in machine vision has drawn from psychophysics, using controlled parametric stimuli to generate accuracy–perturbation curves [25, 26]. Such methods have probed frequency sensitivities, crowding effects, and texture–shape biases. We extend this methodology to the domain of reading, introducing controlled perturbations that sever the link between visibility and identifiability. Our cross-script benchmarks expose the tension between generic invariance and symbolic recognition more sharply than prior visual metamer or distortion studies.

2.4 Sub-character Visual Information in Chinese NLP

Complementary to our findings, recent work has examined whether large language and vision–language models can recognize and exploit the sub-character structure of Chinese writing. For instance, Wu *et al.* [27] construct a benchmark to evaluate models' ability to leverage radicals, strokes, and compositional structure in Chinese characters. Their results show that models exhibit a limited sensitivity to these visual features, and that explicitly providing radical information in prompts can improve downstream tasks such as part-of-speech tagging. In contrast, our study does not probe whether models can partially utilize radicals, but instead demonstrates a systematic failure when characters or words are perturbed in ways that remain fully legible to humans. Whereas [27] highlight the potential for models to benefit from sub-character information in Chinese, we highlight a broader architectural blind spot across both logographic and alphabetic scripts, showing that scaling and prompting are insufficient to bridge the human–machine gap in robust reading.

3 Methods

3.1 Overview

Our goal is to test whether state-of-the-art vision–language models (VLMs) can read stimuli that remain legible to humans but violate structural assumptions of conventional text. We design two psychophysics-inspired benchmarks—one based on Chinese logographs and the other on English alphabetic words—and evaluate multiple VLMs under controlled perturbations. Figure 1 illustrates the overall pipeline.

3.2 Stimuli Construction

Chinese logographs. We construct a dataset of 100 four-character idioms (*chengyu*), a canonical and semantically coherent unit in Chinese. Each character image is rendered in a standard font at fixed resolution. As shown in Figure 2, to generate perturbed stimuli, we apply one of three cutting operations:

- **Horizontal cut:** split the glyph into upper and lower halves, then recombine mismatched parts.
- **Vertical cut:** split the glyph into left and right halves, recombining across characters.
- **Diagonal cut:** split along the main diagonal, recombining triangular fragments.

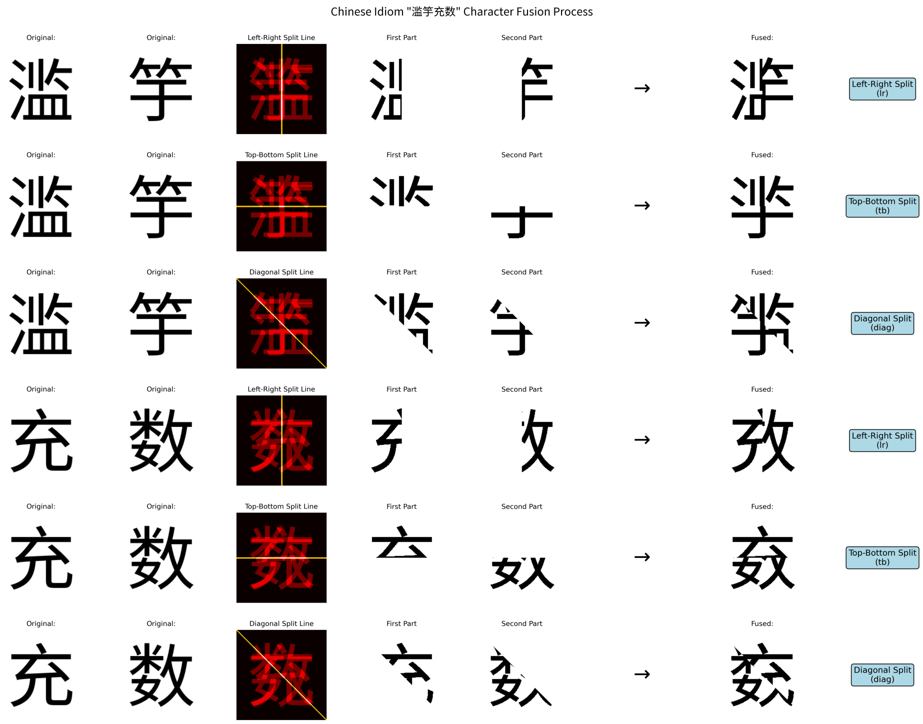

Figure 2: Illustration of the character fusion process for the Chinese idiom "滥竽充数". Each pair of original characters is split along different axes (left–right, top–bottom, or diagonal) and recombined into fused forms. The procedure demonstrates systematic ways of generating visually ambiguous but linguistically interpretable characters, which serve as stimuli for evaluating recognition robustness.

Figure 3: Illustration of the English word fusion process using “hardware” as an example. The original word is split into two segments (“hard” and “ware”), which are then color-coded (red and green) and recombined into an overlapping fused form.

These operations preserve local stroke visibility but destroy coherent glyph boundaries. The resulting composites remain readable to human subjects, who can reconstruct the intended idiom with near-perfect accuracy, but appear as novel glyph-like objects to VLMs.

English alphabetic words. We sample 100 eight-letter words from a standard English lexicon. As shown in Figure 3, taking “hardware" as an example, each word is split into two four-letter halves. The first half is rendered in red, the second half in green, with consistent font and resolution. The two halves are then overlaid to form a single composite image. Humans can reliably parse the superimposed text, but the overlapping colors and fused boundaries challenge models that lack explicit segmentation mechanisms.

Table 1: Prompt design for the Chinese idiom fusion task.

Type	Prompt	English Translation
Basic	请识别这张图片中的汉字，直接返回识别出的文字，不需要其他解释。	Identify the Chinese characters in the image and return only the recognized text, without any explanation.
Detailed	这是一张包含合体字的图片，每个字由两个汉字的部分组合而成。请仔细观察并识别出原始的汉字内容。只返回识别出的文字。	This image contains fused Chinese characters, where each character is composed of parts from two different characters. Carefully identify the original characters and return only the recognized text.
Contextual	这张图片显示的是中文成语的艺术字体，其中每个字都是由两个字的部分拼接而成。请识别出这个四字成语。	This image shows an artistic rendering of a Chinese idiom, where each character is formed by fusing parts of two characters. Identify the complete four-character idiom.

Table 2: Prompt design for the English word fusion task.

Type	Prompt
Basic	What text do you see in this image?
Detailed	This is an 8-letter English word where each half has been diagonally fused together. What word is it?

3.3 Models Evaluated

We evaluate a range of widely used VLMs, including open-source models (e.g., Qwen2-VL-7B, [28] LLaVA-Mistral-7B [8], and LLaVA-Next-Vicuna-7B [29]) and proprietary frontier models (e.g., OpenAI GPT-4o [30], GPT-5 [31], Anthropic Claude Opus 4.1 [32], Sonnet 4 [33], Google Gemini 1.5 Pro and 1.5 Flash [34]). All models are accessed through their publicly available APIs or checkpoints, without any additional fine-tuning.

For the Chinese idiom fusion task, we design three prompt settings (see Table 1). For the English word fusion task, we design two prompt settings (see Table 2). To improve reliability and reduce unnecessary token usage in English, we further adopt a **concise prompting strategy**, appending *"Answer with just the word, no explanation or thinking."* to the end of each prompt.

3.4 Evaluation Protocol

For each stimulus, we collect the model's textual output and compare it to the ground truth. Accuracy is measured differently for Chinese idioms and English words:

- **Chinese idioms:** We adopt two complementary evaluation metrics, both applied after preprocessing by removing all non-Chinese characters using the regex.
  - **Strict Match:** A Boolean indicator of whether the prediction is identical to the ground truth.
    - *Definition:* Correct if and only if `prediction == ground_truth`.
    - *Implementation:* Direct string comparison.
    - *Usage:* Used for computing idiom-level accuracy.
  - **Average Similarity:** A soft metric that measures partial correctness.
    - *Computation:* Sequence similarity is computed via `difflib.SequenceMatcher`.
    - *Return value:* A floating-point score between 0 and 1, where 1 indicates a perfect match.
- **English words:** We only evaluate at the whole-word level, using *Exact Match* accuracy (i.e., the predicted word must match the ground truth exactly).
- **Human baseline:** We recruit 10 participants (native speakers for the corresponding script) and measure accuracy under the same stimuli. We randomize order, enforce attention checks, and report aggregate accuracy with confidence intervals, as well as per-item confusion patterns.

Table 3: Recognition accuracy of idiom images with character-splitting rendering under three prompting strategies. The table reports strict matching rate (exact idiom match), average matching rate (similarity-based match), and human evaluation results (all rated as 100%).

Model	Strict Matching Rate	Average Matching Rate	Human Evaluation
OpenAI			
gpt-4o/basic	0.0%	11.1%	100%
gpt-4o/context	0.0%	5.2%	100%
gpt-4o/detailed	0.7%	7.7%	100%
gpt-5/basic	0.0%	10.5%	100%
gpt-5/context	5.0%	11.4%	100%
gpt-5/detailed	1.5%	12.1%	100%
Anthropic			
claude-opus-4-1/basic	1.0%	15.5%	100%
claude-opus-4-1/context	0.0%	1.8%	100%
claude-opus-4-1/detailed	5.2%	14.7%	100%
claude-sonnet-4/basic	0.0%	10.1%	100%
claude-sonnet-4/context	0.0%	1.4%	100%
claude-sonnet-4/detailed	0.8%	7.3%	100%
Gemini			
gemini-1.5-flash/basic	0.3%	7.5%	100%
gemini-1.5-flash/context	5.0%	10.6%	100%
gemini-1.5-flash/detailed	1.5%	11.0%	100%
gemini-1.5-pro/basic	0.0%	8.0%	100%
gemini-1.5-pro/context	0.2%	3.4%	100%
gemini-1.5-pro/detailed	0.3%	7.5%	100%
LLaVA			
llava-mistral-7b/basic	0.0%	0.6%	100%
llava-mistral-7b/context	0.0%	0.5%	100%
llava-mistral-7b/detailed	0.0%	0.6%	100%
llava-next-vicuna-7b/basic	0.0%	0.6%	100%
llava-next-vicuna-7b/context	0.0%	0.5%	100%
llava-next-vicuna-7b/detailed	0.0%	0.6%	100%
Qwen			
qwen2-vl-7b/basic	0.0%	24.4%	100%
qwen2-vl-7b/context	0.0%	13.9%	100%
qwen2-vl-7b/detailed	0.0%	24.0%	100%

4 Experiments

4.1 Overall Results

Across both the Chinese idiom and English word fusion tasks, all evaluated VLMs show a substantial performance gap compared to human recognition (100% across all cases).

For the Chinese idiom task (Table 3), strict matching accuracy (exact idiom recognition) remains extremely low across all models, typically below 5%. Even with similarity-based evaluation, average matching rates rarely exceed 15%, with the exception of Qwen2-VL-7B, which achieves around 24%. Prompt design has only limited impact: detailed prompts provide modest improvements for some models (e.g., GPT-5, Claude Opus 4.1), while context-oriented prompts do not consistently help.

For the English word fusion task (Figure 4), overall performance is likewise poor, with recognition accuracy capped at 20% even under detailed prompts (GPT-5). Proprietary frontier models (e.g., GPT-4o, GPT-5, Gemini 1.5) show slightly better performance than open-source alternatives (LLaVA, Qwen2-VL), yet still fall far below human-level recognition. Detailed prompts consistently outperform basic prompts, indicating that explicit task guidance can mitigate, but not resolve, the recognition challenge.

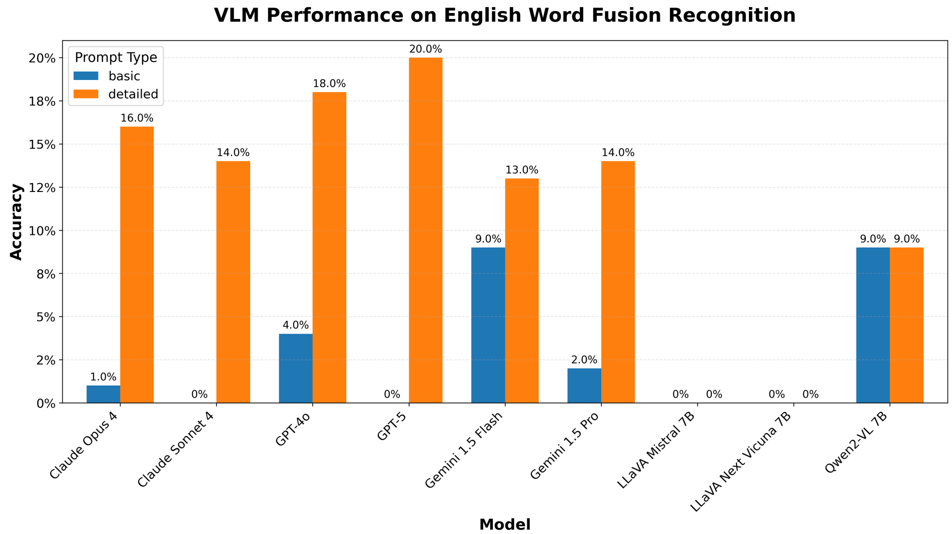

Figure 4: Performance of vision–language models (VLMs) on the English word fusion recognition task. Accuracy is reported under two prompting conditions (basic vs. detailed). Overall performance is low across all models, with proprietary frontier models (e.g., GPT-4o, GPT-5, Gemini) achieving up to 20% accuracy under detailed prompts, while most open-source models (LLaVA variants, Qwen2-VL) perform near chance level.

Figure 5: Word recognition difficulty comparison. Top: hardest words (0% recognition rate). Bottom: easiest words (16.7%-55.6% recognition rate).

Taken together, these results highlight a systematic blind spot in current VLMs: while humans easily parse visually fused characters and words, models fail almost completely, regardless of architecture, scale, or prompting strategy.

4.2 Recognition Difficulty Analysis

To better understand model limitations, we further analyze recognition difficulty across individual stimuli (Figures 5, 6).

For the **English word fusion task** (Figure 5), recognition rates vary widely across words. The hardest cases, such as *hardware*, *checksum*, and *decoding*, exhibit complete failure across models (0% recognition). These words involve highly overlapping letter structures or stroke collisions, which appear to overwhelm the visual parsing mechanisms of current VLMs. In contrast, relatively easier words such as *keyboard*, *alphabet*, and *password* achieve recognition rates between 38.9% and 55.6%, suggesting that clearer segmentation cues or less stroke overlap can partially mitigate the ambiguity.

For the **Chinese idiom fusion task** (Figure 6), the difficulty gap is even more pronounced. Many idioms are never correctly recognized (0%). Even the relatively "easiest" idioms only reach 2.5%–8.0%

Figure 6: Idioms recognition difficulty comparison. Top: hardest idioms (0% recognition rate). Bottom: easiest idioms (2.5%-8% recognition rate).

accuracy. The logographic structure of Chinese characters amplifies this problem: fusing parts of two characters often produces a visually valid but semantically misleading glyph, making it especially challenging for VLMs.

Importantly, however, this gradation of difficulty is *not reflected in human perception*. Human participants reported no meaningful difference between "hard" and "easy" examples, recognizing all items near-perfectly regardless of stroke overlap or fusion style. This divergence highlights that the observed error spectrum is not intrinsic to the stimuli themselves, but rather emerges from architectural blind spots in current VLMs. Whereas humans rely on robust gestalt grouping, contextual priors, and flexible character reconstruction, VLMs lack mechanisms to resolve structured visual ambiguity, leading to stark performance discrepancies that do not exist in human cognition.

In a nutshell, these results reveal that difficulty for VLMs is an artifact of model limitations rather than an inherent property of the task. This underscores a systematic blind spot in visual–linguistic parsing, spanning both alphabetic and logographic writing systems.

5 Discussion and Conclusion

Our study reveals a universal failure mode in vision–language models (VLMs): accuracy that is near-perfect on clean text collapses to near-zero under perturbations that remain fully legible to humans. This gap highlights a fundamental cognitive asymmetry. Humans read by deploying structural priors—segmentation, composition, and symbol binding—while VLMs rely on global invariances that misfire when identifiability is challenged.

The implications are profound. Reading is not mere pattern recognition but structured symbol recovery. Humans achieve this effortlessly; VLMs do not. While larger models and more data may help, our results suggest that scaling alone may be insufficient. Instead, models must incorporate literacy-oriented priors: glyph- or radical-aware representations, mechanisms for segmentation and binding, and cross-script strategies that generalize across writing systems.

The "visible-but-unreadable" blind spot matters well beyond safety. Robust reading under perturbation is essential for science, accessibility, cultural heritage, and trustworthy multimodal AI. Addressing this gap is therefore not a minor refinement but a prerequisite for building AI that can partner with humans in domains where literacy is indispensable.

Our study is limited to two scripts and controlled perturbations. Expanding to more languages, fonts, and distortions, combined with systematic human studies, will help map the full scope of the problem and inspire design. Exploring symbolic-neural hybrids and explicit segmentation architectures remains an open and promising frontier.

In summary, our work documents a striking divergence between human and machine reading and frames it as both a vulnerability and an opportunity. Achieving human-like resilience will require rethinking how structure, priors, and compositionality are embedded in multimodal learning.

6 AI Agent Setup

This paper was developed with assistance from multiple AI agents. GPT-5 was used for idea exploration, literature synthesis, and drafting the manuscript text, which was then iteratively refined through human–AI collaboration. Claude Code contributed to experiment implementation, generation of results, and polishing of technical sections including code, tables, and figures. The framework figure was created using Gemini 2.5 Pro, based on conceptual input from the authors. All human authors reviewed, edited, and approved the final content. The AI agents were used as drafting and productivity tools; all research design, analysis, and interpretations were directed by the human authors.

References

[1] Jack Goody. *The Logic of Writing and the Organization of Society*. Cambridge University Press, 1986.

[2] Stanislas Dehaene. Reading in the brain: The science and evolution of a human invention. *Penguin*, 2009.

[3] Michael H Herzog and Mauro Manassi. Uncorking the bottleneck of crowding: A fresh look at object recognition. *Current Opinion in Behavioral Sciences*, 1:86–93, 2015.

[4] Jonathan Grainger and Arthur M Jacobs. Orthographic processing in visual word recognition: a multiple read-out model. *Psychological review*, 103(3):518, 1996.

[5] Stanislas Dehaene, Laurent Cohen, Mariano Sigman, and Fabien Vinckier. The neural code for written words: a proposal. *Trends in cognitive sciences*, 9(7):335–341, 2005.

[6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[7] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[9] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021.

[10] Alec Radford et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, 2021.

[11] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007.

[12] Akram Bennour, Merouane Boudraa, Imran Siddiqi, Mohammed Al-Sarem, Mohammed Al-Shabi, and Fahad Ghabban. A deep learning framework for historical manuscripts writer identification using data-driven features. *Multimedia Tools and Applications*, 83(33):80075–80101, 2024.

[13] Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. *Psychology of reading*. Psychology Press, 2012.

[14] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp community. *Proceedings of ACL*, 2020.

[15] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 407–425. IEEE, 2024.

[16] Rishi Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[17] Laura Weidinger et al. Taxonomy of risks posed by language models. *Proceedings of FAccT*, 2022.

[18] Uri Hasson, Samuel A Nastase, and Ariel Goldstein. Direct-fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, 2020.

[19] Peter W Battaglia et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[20] Denis G Pelli and Katharine A Tillman. Crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, 2008.

[21] Margaret J Snowling, Charles Hulme, and Kate Nation. *The science of reading: A handbook*. John Wiley & Sons, 2022.

[22] Alec Radford and et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[24] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023.

[25] Robert Geirhos and et al. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[26] Katherine L Hermann and et al. The origins and prevalence of texture bias in convolutional neural networks. *NeurIPS*, 2020.

[27] Xiaofeng Wu, Karl Stratos, and Wei Xu. The impact of visual information in chinese characters: Evaluating large models' ability to recognize and utilize radicals. *arXiv preprint arXiv:2410.09013*, 2024.

[28] Qwen Team. Qwen2-vl: Enhancing vision-language understanding with better alignment and data. *arXiv preprint arXiv:2407.10372*, 2024.

[29] Haotian Liu, Zirui Yang, Zhiqiang Wang, et al. Llava-next: Stronger visual instruction tuning with better data and model design. *arXiv preprint arXiv:2406.08494*, 2024.

[30] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[31] OpenAI. Gpt-5 system card. *OpenAI Technical Report*, 2025. Available at `https://cdn.openai.com/gpt-5-system-card.pdf`.

[32] Anthropic. Claude 4.1 model family: Opus, sonnet. *Anthropic Technical Report*, 2025. Available at `https://www.anthropic.com/news/claude-4-1`.

[33] Anthropic. Claude 4 model family: Opus, sonnet, haiku. *Anthropic Technical Report*, 2025. Available at `https://www.anthropic.com/news/claude-4`.

[34] Google DeepMind. Gemini 1.5: Unlocking multimodal understanding across long context. *DeepMind Technical Report*, 2024. Available at `https://deepmind.google/discover/blog/gemini-1-5/`.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[B]**

   Explanation: The research idea and framing were primarily human-driven. The motivation—testing whether VLMs can read human-legible but perturbed text—was formulated by the human researcher, drawing on background knowledge in psycholinguistics and AI safety. AI tools provided assistance in refining the wording of hypotheses, brainstorming possible perturbation types, and clarifying methodological framing, but the core direction came from the human.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[C]**

   Explanation: AI models were used extensively for generating stimuli (e.g., rendering idioms with character-splitting, overlaying English words with colors, creating figure prototypes) and for producing code snippets to automate data processing and evaluation. The human researcher guided the overall design, validated outputs, and ensured scientific rigor, but much of the coding and figure-generation was handled with AI assistance. Thus, the majority of the execution work came from AI under human supervision.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[C]**

   Explanation: AI helped organize recognition accuracy results, generate tables and plots, and draft interpretations of trends across models and prompt types. The human researcher critically evaluated these analyses, drew the central conclusions (e.g., the universal gap between human and AI literacy), and ensured the arguments connected to cognitive science and AI architecture. Therefore, AI performed a large share of the data summarization and visualization, while humans provided conceptual interpretation and validation.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[C]**

   Explanation: AI was heavily used for drafting, polishing, and restructuring text—including the Introduction, Discussion, and figure captions. The human researcher provided the core ideas, checked factual correctness, ensured alignment with scientific standards, and made final editorial decisions. Thus, while the narrative flow and sentence structure benefited from AI generation, the intellectual substance and framing remained human-led.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: 1) Surface-level reasoning: AI often produced plausible but shallow explanations, which required human correction to ensure conceptual depth and technical accuracy. 2) While AI tools were effective for generating analysis figures and result plots (e.g., through Python code for automated visualization), they showed clear limitations in producing complex schematic diagrams such as methodological flowcharts. These tasks often required significant manual adjustment or external design tools. Among the models tested, Gemini 2.5 Pro provided the most useful support for figure drafting, but even so, the quality and flexibility were below what is required for final publication standards.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions—identifying a universal failure mode in VLMs, designing cross-script benchmarks, and analyzing implications. These claims are fully supported by the experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated discussion of limitations, noting that the study is restricted to two scripts and controlled perturbations, and that further research is needed across more languages, fonts, and distortions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper is empirical and does not contain formal theorems or proofs. Therefore, this item is not applicable.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the benchmark construction, prompts, evaluation metrics, and tested models in detail. This information is sufficient for independent reproduction given access to the listed VLM APIs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: we will release the code upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: As the study evaluates existing models rather than training new ones, the relevant details focus on prompt design, dataset construction, and evaluation procedures, which are fully documented in the manuscript.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Recognition accuracies are averaged over 2 runs, and we report both strict and similarity-based metrics. This conveys the robustness of results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments relied on API access to proprietary VLMs and open-source checkpoints (7B scale) run on a single H200 GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: The work conforms to the Agents4Science Code of Ethics. No private or sensitive data are used, and experiments focus on widely available benchmarks and open-source/public APIs.

   Guidelines:

   - The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper explicitly discusses broader impacts, including positive applications in science, accessibility, and cultural heritage, as well as risks of adversarial misuse in content moderation and security contexts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.