# INDUCING DYSLEXIA IN VISION LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Dyslexia, a neurodevelopmental disorder characterized by persistent reading difficulties, is often linked to reduced activity of the visual word form area in the ventral occipito-temporal cortex. Traditional approaches to studying dyslexia, such as behavioral and neuroimaging methods, have provided valuable insights but remain limited in their ability to test causal hypotheses about the underlying mechanisms of reading impairments. In this study, we use large-scale vision-language models (VLMs) to simulate dyslexia by functionally identifying and perturbing artificial analogues of word processing. Using stimuli from cognitive neuroscience, we identify visual-word-form-selective (VWF-selective) units within VLMs and demonstrate that targeted ablation of these units, unlike ablation of random units, leads to selective impairments in reading tasks while general visual and language comprehension abilities remain intact. In particular, the resulting model matches dyslexic humans' phonological deficits without a significant change in orthographic processing. Additionally, the model's VWF-selective units predict human-VWFA neural responses better than random units and the ablated model mirrors dyslexic behavior in font sensitivity. Taken together, our modeling results replicate key characteristics of dyslexia and establish a computational framework for investigating reading disorders. [1]

## 1 INTRODUCTION

Dyslexia is a complex neurodevelopmental learning disorder that impairs a person's ability to decode written language and to spell, despite normal intelligence and educational opportunity. Difficulties arise mainly in reading, spelling, accuracy, fluency, and decoding abilities (Snowling et al., 2020; Kunwar & Sapkota, 2022). Depending on the severity criteria used, dyslexia is estimated to impact 6–17% of children in the school-age population (Pennington et al., 2009; Lyon et al., 2003), with the global prevalence reported to range from less than 5% to 20% of the entire population (Wagner et al., 2020). The brain disorder is associated with atypical neural activity in regions involved in language processing and phonological representation, in particular the visual word form area (VWFA) (Brem et al., 2020; Monzalvo et al., 2012; Shaywitz et al., 2002; Maurer et al., 2007; Boros et al., 2016; Kronbichler & Kronbichler, 2018). While dyslexia has been linked to both genetic and neurocognitive factors such as phonological abilities, its exact causal mechanisms remain unclear (Peterson & Pennington, 2015; Werth, 2023).

Advances in Machine Learning have significantly enhanced our ability to simulate and understand neural processes by leveraging biologically inspired computational frameworks. These developments have enabled the modeling of brain activity with increasing fidelity, capturing not only the structural characteristics but also the precise neural patterns of the human brain. Recent work has demonstrated that artificial neural networks can approximate the brain's mechanisms, aligning closely with neural responses observed in human cortical regions, such as in vision and language (Yamins & DiCarlo, 2016; Schrimpf et al., 2020; 2021).

Building on this foundation, our research aims to extend computational approaches of the healthy brain to the modeling of specific brain disorders — here focusing on dyslexia. In particular, we develop a computational framework to model dyslexia by functionally identifying and perturbing specific units within vision-language models. This approach simulates hypoactivations documented in the VWFA of dyslexic subjects (Brem et al., 2020), abstracting away genetic and other con-

[1] Code available via anonymized repository.

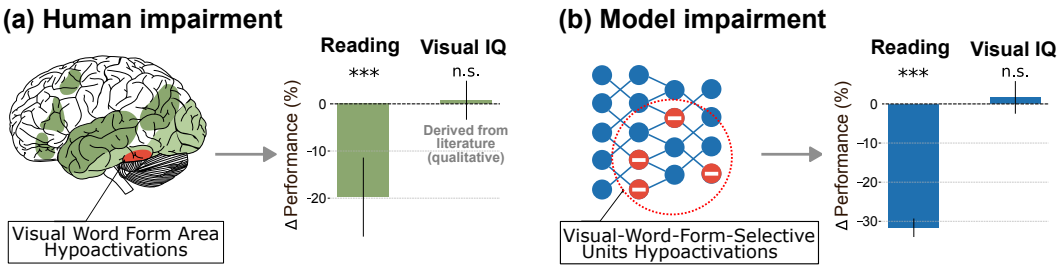


Figure 1: **Modeling dyslexia via visual-word-form hypoactivation. (a)** In humans, reduced activity in the visual word form area is thought to result in diminished performance on reading-related measures while sparing general visual intelligence. **(b)** Testing this hypothesis in vision-language models, we find that ablating visual-word-form-selective units produces the same dissociation.

tributing factors. A successful model simulation of dyslexia should exhibit *selective* impairment of reading performance while preserving other cognitive functions such as general intelligence and reasoning.

We show that identifying and ablating visual-word-form-selective units in a vision language model (such as Qwen (Yang et al., 2024)) indeed leads to diminished performance on dyslexia screening assessments like the Rapid Online Assessment of Reading (ROAR) (Yeatman et al., 2021), while leaving general visual intelligence measures unaffected, including Raven's Progressive Matrices (Burke, 1958; Zhang et al., 2019) and the Kempler Test (Kempler et al., 1998) (Fig. 1). Ablating random units on the other hand does not have a selective effect on reading abilities and instead equally affects visual reasoning performance. An investigation into chosen hyperparameters reveals that the choice of layer type is critical while the effect varies more smoothly with the number and severity of affected units. Testing our model's reading deficits in more detail we find that it mirrors phonological deficits in human dyslexic subjects, but shows no significant impairment on orthographic stimuli. Overall, our computational results mimic empirical findings in dyslexic individuals, exhibiting selective reading-specific impairments without corresponding deficits in general intelligence (Snowling et al., 2020; Peterson & Pennington, 2015), and thus establish a computational framework for modeling the neural mechanisms underlying brain disorders.

## 2 Background & Related Work

**Behavioral effects of dyslexia.** Dyslexia is characterized by persistent difficulties in accurate and/or fluent word recognition and spelling, despite adequate education, normal intelligence, and intact sensory abilities (Lyon et al., 2003). These difficulties are closely linked to deficits in phonological awareness and processing, often manifesting as problems with decoding unfamiliar or nonsense words, rapid naming, and word finding. Further challenges may include variable difficulties with letter–sound learning, oral reading accuracy, written composition, and reading comprehension (Roitsch & Watson, 2019). Importantly, studies indicate that the core mechanisms of dyslexia are consistent regardless of IQ (Stanovich, 2005; Tanaka et al., 2011).

**Hypothesized neural substrate.** A growing body of evidence highlights the central role of the Visual Word Form Area (VWFA) in the neurobiology of dyslexia. The VWFA, located in the left ventral occipitotemporal cortex, is critical for fluent reading as it specializes in the rapid and automatic recognition of written words (Dehaene & Cohen, 2011; McCandliss et al., 2003). In individuals with dyslexia, this region often shows functional and anatomical abnormalities. Longitudinal studies have revealed that dyslexic readers consistently exhibit smaller and less selective VWFAs compared to typical readers, even after significant improvements in reading performance through targeted training (Mitchell et al., 2025; Brem et al., 2020). These persistent differences suggest that VWFA atypicalities are not merely a consequence of poor reading experience but represent a stable neurobiological trait of dyslexia. Empirical studies consistently report VWFA hypoactivation and structural differences in dyslexics across languages and cultures (Paulesu et al., 2001; Silani et al., 2005), reinforcing its cross-linguistic relevance. Moreover, lesions in this region have been causally linked to acquired reading disorders (Turkeltaub et al., 2013), underscoring the VWFA's es-

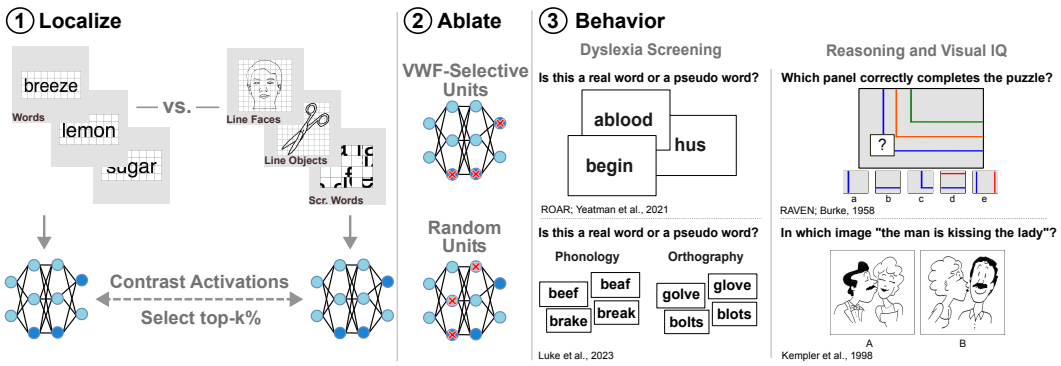


Figure 2: **Identifying visual-word-form-selective units in VLMs. (1)** To identify VWF-selective units, we compare unit activations in response to images of words versus images of non-words, and identify the units that exhibit the strongest word selectivity. **(2)** To model the reduced VWFA activity observed in dyslexic individuals, we ablate the localized units. As a control, we ablate an equal number of randomly selected units. **(3)** To assess the impact of ablations, we evaluate model performance on dyslexia screening tasks (ROAR (Yeatman et al., 2021) and the Lexical Decision benchmark (Luke et al., 2023)) as well as on visual IQ and reasoning tasks (RAVEN (Burke, 1958) and Kempler et al. (1998) sentence comprehension tasks).

sential role in both typical and atypical reading processes. Nevertheless, the causal status of VWFA abnormalities remains debated, with some studies suggesting that these differences may reflect a consequence rather than a primary cause of dyslexia (Olulade et al., 2013; Valdois, 2010).

**Models of Brain Function.** Computational neural networks have proven effective at predicting both behavioral performance and neural responses in healthy subjects, making them valuable tools for studying brain function. In the visual domain, deep convolutional as well as transformer-based models have successfully captured neural activity in the primate ventral visual stream across object and scene recognition tasks (e.g., Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018; 2020; Cadena et al., 2019; Spoerer et al., 2020; Zhuang et al., 2021; Wang et al., 2023; Margalit et al., 2024; Gokce & Schrimpf, 2024; Lonnqvist et al., 2025; Tang et al., 2025). In the language domain, transformer-based and recurrent models have accurately predicted neural responses related to semantic, syntactic, and phonological processing (e.g., Schrimpf et al., 2021; Caucheteux et al., 2022; Goldstein et al., 2022; Toneva et al., 2018; Hosseini et al., 2024; Aw et al., 2023; Tuckute et al., 2024; Rathi et al., 2024; AlKhamissi et al., 2025; Du et al., 2025).

Machine learning methods thus provide a powerful platform for capturing the cognitive and neural patterns associated with dyslexia. By enabling controlled experiments and hypothesis testing in silico, such models can offer new insights into the mechanisms underlying the disorder and might support the development of targeted diagnostic and intervention strategies.

Another approach to model brain disorders employs connectivity-based models, which simulate large-scale brain dynamics using empirically derived structural or functional connectomes. In the context of reading and dyslexia, such models have been used to relate altered white matter pathways and disrupted functional connectivity to deficits in phonological processing and reading performance (Müller et al., 2017; Sihvonen et al., 2021), but these comparatively coarse-grain approaches remain limited in capturing the precise underlying neural mechanisms.

To the best of our knowledge, there is no prior work on modeling brain disorders using system-level neural models, particularly in the context of dyslexia. While there are studies that employ computational models to investigate aspects of dyslexia, such as visual information processing (Ogawa et al., 2023) and handwriting anomalies (Alevizos et al., 2024), these approaches do not simulate the complex neural activity changes associated with dyslexia. Therefore, our work represents a first attempt to model brain disorders (here, dyslexia) through hypothesized neural activity changes in state-of-the-art models.

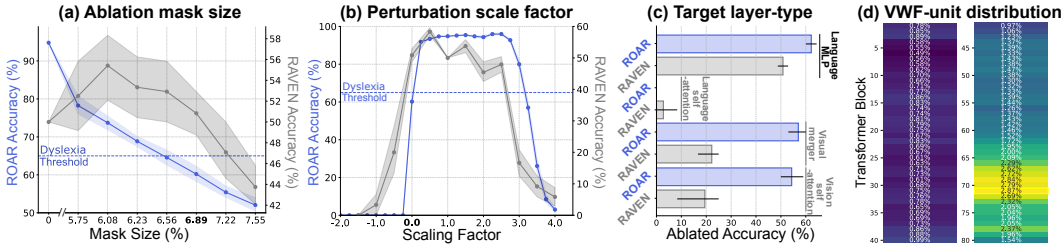


Figure 3: **Ablating visual-word-form-selective units in the model. (a)** Increasing the number of VWF ablated units translates into a severe monotonic performance decline in ROAR (blue), while RAVEN (gray) is only affected at larger mask sizes. We chose the first mask size (bold) where ROAR performance falls below the dyslexia threshold (blue dashed line). Shaded regions represent 95% confidence intervals. **(b)** Beyond full ablation (bold), scaling unit activity (with a fixed mask size of 6.89%) has little effect for positive scaling, while negative scaling severely degrades outputs non-selectively. **(c)** While ablations in all layer types substantially affect performance, only the MLP components of the language decoder showed selective effects, indicating its core involvement in reading (full trends in Fig. 11). **(d)** Distribution of VWF-selective units across the 80 transformer blocks of the language decoder. Ratios are mean across 20 random seeds and resampling of the localizer stimuli; standard deviations never exceed 0.03%.

## 3 Benchmarks

We evaluate model performance using a suite of standardized assessments originally designed for human subjects in clinical psychology and cognitive neuroscience (Fig. 2 right). These include tests of lexical decision-making (ROAR, Yeatman et al., 2021), visual reasoning (RAVEN, Burke, 1958; Zhang et al., 2019), sentence-level syntactic comprehension (Kempler Test, Kempler et al., 1998), and dissociations between orthographic and phonological processing (Luke et al., 2023). By repurposing these human-designed benchmarks for VLMs, we enable a targeted and comparable evaluation of specific cognitive capacities and their impairments, providing a structured framework for interpreting model behavior in relation to established cognitive theories.

**Rapid Online Assessment of Reading (ROAR)** is a browser-based, self-administered lexical decision task that evaluates core reading skills by measuring both accuracy and speed in distinguishing real words (e.g., “able”, “animal”) from pseudo words (e.g., “ablood”, “accastanct”) (Yeatman et al., 2021; more examples in Appendix Table 1). In our model, however, response time is not a consideration, so we evaluate performance solely based on the accuracy of lexical decisions. On each trial, the model is shown an image of a letter string and must decide whether it is a real word or a pseudo word. We present 200 real words and 200 pseudo words drawn from the full ROAR-Word corpus as the train set for finding the minimal VWFA mask, with the remaining 50 real words and 50 pseudo words serving as the test set for lexical evaluation. *Dyslexia Threshold*: We define 65% ROAR performance as the threshold at which subjects are considered as dyslexic. This number is one standard deviation below the mean ROAR-score of the human population, consistent with prior works in humans (Wagner et al., 2020). This choice is also meaningful from an epidemiological perspective as dyslexia is estimated to affect about 5–20% of people (Wagner et al., 2020), and this prevalence aligns well with scores in the range of $[1.65, 0.95]$ standard deviations below the mean.

**Raven’s Progressive Matrices (RAVEN)** is a nonverbal assessment of fluid intelligence in which participants must select, from five candidate panels, the one that correctly completes a 3×3 or 2×2 matrix by preserving an underlying pattern of shapes or spatial relations (Burke, 1958; Zhang et al., 2019; examples in Appendix Fig. 9). Because it minimizes linguistic and cultural biases, RAVEN serves as an ideal control for general visual-spatial reasoning in our study. In our experiments, we administered the easy section of the RAVEN clinical version which consists of twelve items. By assessing the model’s accuracy on these twelve puzzles, we establish a baseline for general visual intelligence against which any reading-specific impairments can be contrasted.

**Kempler’s Sentence Comprehension Test** is a clinical assessment designed to evaluate syntactic comprehension by presenting participants with pairs of images accompanied by a spoken or written sentence (Kempler et al., 1998). The task is to choose the image that correctly corresponds to the

sentence, making it a useful probe of grammatical understanding and sentence-level semantics. In our implementation, we adapt this test for model-based evaluation by using only the image pairs from the original stimulus set, while providing the sentence as a text prompt. The model is asked to decide which image matches a given caption (examples in Appendix Fig. 10). This transforms the task into a visual question answering (VQA) challenge, allowing us to assess the model's capacity for sentence comprehension grounded in visual reasoning.

**Lexical Decision with Orthographic and Phonological Manipulations.** To further dissect reading impairments into orthographic and phonological components, we incorporated stimuli from a lexical decision task developed by Luke et al. (2023). This benchmark includes four stimulus types: (1) homophones (e.g., "brake" and "break"), (2) pseudo-homophones (e.g., "beaf" and "birf"), (3) transposed-letter (TL) neighbour words (e.g., "blots" and "bolts"), and (4) TL non-words (e.g., "golve" and "glove"). These stimuli were originally used to contrast phonological and orthographic theories of dyslexia in human readers. For analysis, we group the stimuli into two categories: phonology-sensitive (homophones and pseudo-homophones) and orthography-sensitive (TL words and TL non-words). The model performs a lexical decision on each item (an image containing a word or a non-word), and accuracy serves as the primary outcome measure. By comparing performance across these two groups, we can determine whether the model finds phonologically demanding items or orthographically demanding items more challenging.

## 4 METHODOLOGY

**Functional Localization of VWF-selective Units.** Given the well-established role of the VWFA in dyslexia, we target visual-word-form-selective (VWF-selective) units in vision-language models (VLMs) as regions of interest for artificial lesions. To reflect the dual involvement of visual and linguistic processes in dyslexia, we focus on computational models that process images (and text) as input which reflect the full processing chain from retina (pixel) input to language. These VLMs, by presenting visual stimuli as image input and task instructions as text tokens, allow us to dissociate visual processing impairments from general language deficiencies, and their integration of visual and linguistic modalities within a single architecture provides a unified framework for examining the cortical processes underlying word recognition and reading impairments in dyslexia. Our method enables a hypothesis-driven investigation of reading-specific mechanisms directly within artificial systems and provides a scalable and manipulable alternative to costly human studies.

Using functional localizers to identify selective units in computational models is a neuroscience-inspired approach rooted in studies like Kanwisher et al. (1997), who used fMRI localizers to identify specialized brain regions. Saygin et al. (2016) functionally localized the VWFA by contrasting responses to written words with visually matched control stimuli, including line drawings of faces, scrambled words, and line drawings of objects, enabling robust identification of word-selective regions in the ventrotemporal cortex. Recently, AlKhamissi et al. (2025) adapted such functional localization methods for large language models and identified the causal role of language-selective units through targeted ablations. We adopt a similar approach by applying a functional localizer paradigm to our vision language model, targeting VWF-selective units.

Specifically, we use a classic fMRI localizer from neuroscience (Saygin et al., 2016) to identify VWF-selective units in computational models. The model is presented with the four stimulus categories from Saygin et al. (2016): written words, scrambled words, faces, and objects. For each considered model unit, we compute a t-statistic comparing responses to word images versus the three non-word control categories. The t-statistic quantifies how strongly a unit prefers words relative to other stimuli by measuring the difference in response means normalized by response variability. Units with higher t-statistics respond more selectively and reliably to words, helping us identify candidate VWF-selective units in the model. Units are then ordered by descending t-statistic, and we define the top $k\%$ of these units to be the model's VWF-selective units (Fig. 2).

**Minimal Subnetwork for Dyslexia Simulation.** To induce a dyslexia-like impairment, we gradually increase the fraction of the top k% VWF–selective units that are ablated (i.e., set to zero) within the model's language layers. We begin with no masking (0%) and repeatedly mask a larger proportion of these units, evaluating at each level the model's accuracy on the train subset of ROAR stimuli. We stop increasing the mask size as soon as the ROAR score falls below 65%, which we

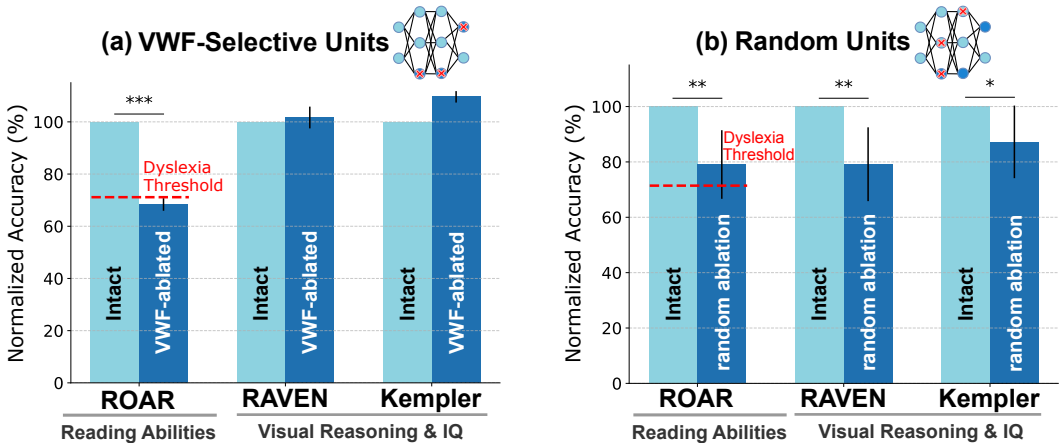


Figure 4: **Reading-selective deficits from ablating VWF-selective units.** **(a)** Ablating VWF-selective units led to a selective reading deficit below the dyslexia threshold (ROAR, $p < 0.012$), while performance on visual IQ and reasoning benchmarks (RAVEN, Kempler) remained intact or was slightly enhanced. **(b)** Ablating an equal number of randomly selected units from the same layers affected performance throughout, with ROAR remaining above the dyslexia threshold and significant impairments to visual reasoning. Dark blue bars indicate ablated model accuracy, relative to the intact model (light blue bars). Error bars denote 95% confidence intervals, and significance was assessed with one-sample, one-tailed Student's t-test.

defined as the ROAR dyslexia threshold (Section 3). The smallest proportion of masked units that meets this criterion defines our VWF–selective units (see also Fig. 3a).

**Model Details.** We focused our analyses on `Qwen2-VL-72B` (Yang et al., 2024), selected for its strong performance across benchmarks, particularly in OCR and visual understanding tasks. Qwen2-VL-72B is a large-scale autoregressive transformer integrating visual and textual modalities. Within the model, we examined the MLP layers of the language decoder where the minimal subnetwork for dyslexia corresponds to approximately 6.89% of all units (see below for hyperparameter analyses). Our approach successfully elicited reading deficits in all tested models (additionally `LLaVA-NeXT-13B`, Liu et al., 2024; and `PixTral-12B`, Agrawal et al., 2024). These deficits selectively affected reading only, without significant differences on visual IQ benchmarks as long as baseline performance was significantly above chance (Appendix Fig. 8).

**Perturbation Strength.** Beyond a full ablation (activity = 0), we additionally experimented with parametric perturbations of VWFA units by varying the scaling factor on the activations of the selected VWF-selective units. Specifically, we tested scaling factors in the range $[-2, 4]$, including sublinear, neutral (1.0), and superlinear amplification. We observed that only complete ablation (scaling factor = 0) reliably produced the selective impairments characteristic of dyslexia (Fig. 3b). Perturbations at other scales either had negligible behavioral effects or resulted in severely degraded output coherence, producing nonsensical or empty responses (Table 3). Consequently, we focus our analysis on full ablation as the most faithful computational analogue of focal VWFA disruption.

**Layer Choice.** To identify VWF-selective units, we evaluated subnetworks from different layer types across the model. The MLP gate projection layers in the language decoder emerged as the most selective, aligning with prior findings that MLP layers exhibit strong knowledge-specific selectivity (Meng et al., 2022; Zhang et al., 2021). To support this interpretation, we ablated minimal subnetworks from other components of the model, the vision encoder (`visual.blocks.{i}.attn.proj`), the visual merger (`visual.merger.mlp.{i}`), and the language decoder's self-attention outputs (`model.layers.{i}.self_attn.o_proj`). These ablations led to sharper declines in RAVEN performance relative to ROAR performance, indicating that the observed effects are not purely reading-selective (Fig. 3c). Based on these results, we define VWF-selective units as those located in the MLP layers (`model.layers.{i}.mlp.gate_proj`) across all 80 transformer blocks of the language decoder (Fig. 3d).

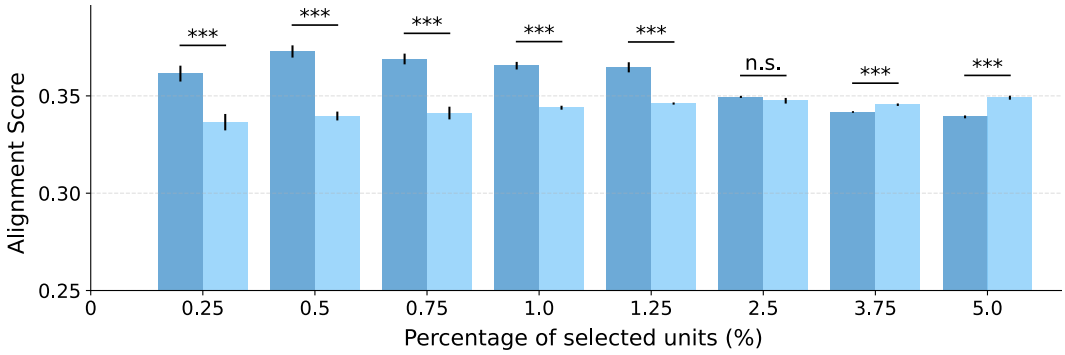


Figure 5: **VWF-selective units align with human VWFA activity.** Fraction of explainable variance (Pearson correlation / noise ceiling) between predicted and actual fMRI responses in the human VWFA, computed across the top-k% of VWF-selective units of the model (dark blue bars) and randomly selected units (light blue bars) for 5 seeds. At small subset sizes, VWF-selective units show significantly higher alignment than the average of five randomly selected subsets. Error bars indicate 95% confidence intervals across seeds.

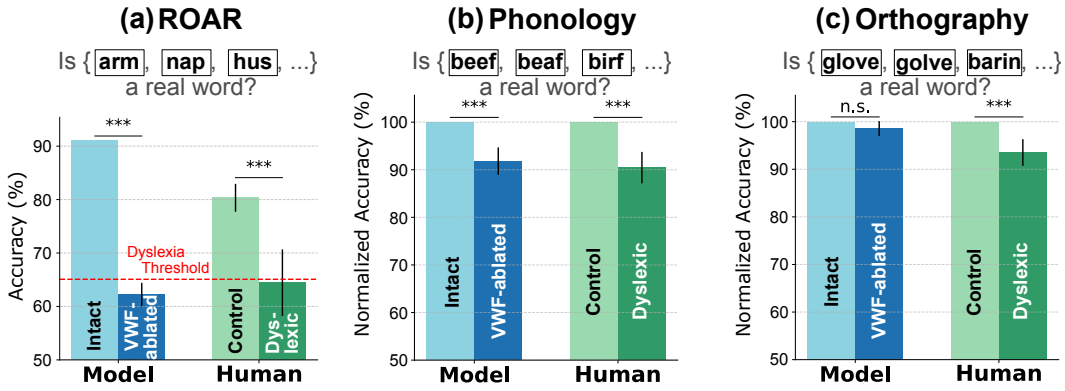


Figure 6: **Model mirrors human phonological reading deficits. (a)** Lexical decision accuracy on the ROAR reading test before (light blue) and after (dark blue) ablation of VWF-selective units. Following ablation, model lexical decision accuracy at real/pseudo word classification drops below the dyslexia threshold, paralleling dyslexic participants (dark green) relative to control subjects (light green). **(b)** Ablated model performance decreases for phonologically confusable stimuli (e.g. "beaf" which sounds the same as "beef" but is a pseudo word), indicating a phonological deficit and mirroring observations in humans. **(c)** Ablated model performance is not significantly affected on orthographically confusable stimuli (e.g. "golve" which looks similar to "glove" but is a pseudo word), whereas dyslexic humans tend to be affected. Error bars denote 95% confidence intervals; model results are averaged over 20 random seeds, each corresponding to a different sample of the localizer; significance assessed via one-tailed Student's t-test (Appendix A.4).

## 5 Results

We evaluated the impact of VWFA-targeted ablation on model performance across tasks that probe reading, language comprehension, and nonverbal reasoning. We examine how selective disruption of functionally localized units affects different cognitive domains.

**Selective Reading Deficits, Preserved Reasoning Capabilities.** Ablating visual-word-form (VWF)-selective units leads to a substantial decline in lexical decision accuracy on the ROAR lexical reading task (Fig. 4a). Table 1 of the Appendix shows example responses of the ablated model. Performance in fact dropped below the dyslexia threshold on the held-out test set, a marked reduction from the model's pre-ablation accuracy ($-32\%$ in normalized accuracy with $p \ll 0.01$ using a one-tailed Student's t-test for whether ablated model performance was significantly lower than

baseline performance on each of the tasks; see Appendix A.4 for more details). In contrast, performance on the RAVEN's Progressive Matrices task remained stable ($p \approx 0.75$), indicating that the lesioned subnetwork selectively impairs reading ability while sparing broader visual reasoning processes. We further evaluated sentence comprehension and visual grounding using the Kempler Sentence Comprehension Test as another visual Q&A task. Following ablation, the model's accuracy on the Kempler task did not decrease compared to the non-ablated baseline ($p \approx 1$, one-tailed test; a two-tailed Student's t-test reveals a statistically significant increase in performance of $+10\%$, $p \ll 0.01$). This pattern parallels empirical findings in some dyslexic individuals, where deficits in phonological or orthographic processing are accompanied by preserved or even enhanced performance in visual-spatial reasoning and comprehension tasks (von Károlyi et al., 2003; Stanovich, 2005; Tanaka et al., 2011).

**Ablating Random Units Does Not Lead to Selective Reading Impairments.** To verify that the observed dyslexia-like effects were not simply due to the number or layer-wise distribution of ablated units, but rather depended on the specific identity of the VWF-selective units, we conducted a control experiment. We preserved both the total number of ablated units and the selected layers, but randomly selected which specific units were ablated within each layer. This approach did not lead to selective reading impairments: the performance on *all* benchmarks dropped ($-21\%$ with $p < 0.003$ for ROAR, $-21\%$ with $p < 0.004$ for RAVEN, and $-13\%$ with $p < 0.042$ for Kempler), and reading performance did not drop below the dyslexia threshold ($p \approx 0.87$; Fig. 4b). Furthermore, when the same number of units was ablated but distributed randomly across the entire network rather than within the same layers, the model's outputs largely degraded, often generating incoherent or empty responses, confirming that the reading-specific impairment depends on targeting VWF-selective units.

**Human-Like Phonological Deficits.** VWFA-targeted ablation selectively impairs reading performance, producing deficits reminiscent of those observed in dyslexic individuals. To assess whether the observed deficit was primarily phonological or orthographic, we evaluated model performance on the lexical decision benchmark from Luke et al. (2023). Post-ablation, model accuracy dropped significantly on phonology-sensitive items (Fig. 6; $-8\%$ with $p \ll 0.01$), but remained relatively stable for orthography-sensitive stimuli ($p > 0.059$). These results suggest that phonological processing was disproportionately disrupted, while orthographic representations were largely preserved. Human behavioral data reveal a related dissociation: Dyslexic participants showed a pronounced accuracy reduction on phonology-sensitive items ($-9\%$ with $p \ll 0.01$), but also demonstrated significant, though slightly smaller, impairments on orthography-sensitive items($-6\%$ with $p \ll 0.01$) *on average*. While the human profile reflects a combination of phonological and orthographic deficits, the ablated model primarily captures the phonological component. This alignment supports the interpretation that targeted suppression of the VWFA induces a phonologically selective reading impairment, consistent with developmental dyslexia associated with hypoactivation in the Visual Word Form Area (McCandliss et al., 2003).

**Alignment Between VWF-selective Units and Human Brain Activity.** Building on the observed behavioral alignment, we next assessed whether selected units correspond to human neural activity. We quantified how VWF-selective units align with human neural responses using a standard encoding-model framework (Naselaris et al., 2011; Schrimpf et al., 2021). For each subject in the Marvi et al. (2025) dataset, ridge-regression models were trained to predict voxel-wise fMRI responses from the model's video-evoked activations using 5-fold cross-validation. Pearson correlations between predicted and actual voxel responses were averaged across folds and normalized by each voxel's noise ceiling to yield the alignment score. We evaluated model alignment using subsets of the most VWF-selective units, varying the subset size from 0.25% to 5% of all units. For each subset size, we generated five VWF-selective masks by resampling the localizer stimuli with five different seeds. Each VWF-selective mask was paired with a size-matched random mask for comparison. At smaller percentages, VWF-selective subsets showed significantly higher neural alignment than random subsets, indicating that these units encode brain-relevant structure rather than arbitrary features and supporting their role in both behavioral and neural correspondence to visual word-form processing (See Fig. 5). For larger percentages (above 1.25%), the same effect is not observed. This likely reflects differences in feature dependence; VWF-selective subsets are drawn from a ranked list and therefore contain correlated units, whereas random subsets contain more heterogeneous, less correlated features. As more units are included, these independent random features improve fMRI prediction, while correlated VWF-selective features add redundant information. This

suggests that VWF-selective units form a specialized subspace that is highly informative when the most selective units are sampled.

**Predicting Dyslexia-Friendly Fonts.** Because the ablated model reliably reproduces behavioral patterns observed in dyslexic readers, it can be used to evaluate reading interventions and design supportive visual stimuli. To illustrate this, we presented both the intact and ablated models with the original ROAR lexical decision items in Arial (the standard font used in the benchmark), keeping the font size as similar as possible, and then systematically replaced the font with alternatives known from prior work and expert recommendations to be either challenging or easier for dyslexic readers. Fonts tested included, Tahoma, OpenDyslexic, Comic Sans, KG Primary Penmanship Chiller, Papyrus, Onyx, and Ecofont. We also tested Arial and OpenDyslexic on a dark orange background, grounded in personal reports from dyslexic individuals that colored backgrounds can aid reading. The ablated model mirrored human sensitivity and performance decreased or remained unchanged on fonts considered difficult, and improved or remained stable on dyslexia-friendly fonts, whereas the intact model showed no systematic differences (See Fig. 7).The ablated model showed significant improvement in lexical decision accuracy relative to Arial for OpenDyslexic, OpenDyslexic on an orange background, Comic Sans, and KG Primary Penmanship ($p \ll 0.001$), and for Orange Arial ($p \ll 0.01$). In contrast, performance significantly decreased for Papyrus ($p \ll 0.01$) compared to Arial. This demonstrates that the model can capture font-specific reading difficulties observed in dyslexic readers. Importantly, this approach opens a novel possibility unique to computational modeling. Future work can design or optimize new fonts by identifying visual configurations that maximize performance in the ablated model while preserving readability in the intact model.

## 6 Discussion

Linking targeted neural ablations in vision language models to selective behavioral deficits opens new opportunities for the study of reading and other brain disorders. By inducing dyslexia-like behavior in a controlled setting, we can examine how specific changes affect performance and generate new ideas about the mechanisms involved. These models, though abstracted from biology, preserve key correspondences with brain-like processing – positioning them as powerful tools to advance our understanding of brain disorders, prototype interventions, and bridge gaps left by the inherent limitations of human data collection.

**Models as Digital Twins.** Models with dyslexia-like behavior serve as digital twins that allow for controlled experiments that are difficult, invasive, or impossible in humans, such as targeted ablations. While there are numerous inherent differences between biological brains and the computational models used here, brain-aligned models can serve as tools or proxies to identify which neural mechanisms are likely responsible for observed behavioral deficits, narrow the search space for human experiments, and design more precise behavioral or neuroimaging tests to validate hypotheses about dyslexia.

**Interpretability.** This work advances interpretability by identifying a reading-selective subnetwork in VLMs reminiscent of the VWFA in humans. Targeted ablations within this subnetwork induce a phonological rather than an orthographic deficit, which is notable given that the model has never processed auditory input. These findings suggest that functional localization can be used to uncover mechanistic and causal links between specialized neural circuits and behavioral outcomes. We also observe a performance increase on nonverbal benchmarks post-ablation, paralleling findings in some human dyslexics (von Károlyi et al., 2003; Lam & Tong, 2021), which we plan to explore further in future work with potential implications for enhancing learning via ablation of non-selective units.

**Development.** Dyslexia is a developmental disorder that emerges as children acquire literacy, and hypoactivation of the VWFA is often linked to phonological deficits (McCandliss et al., 2003), although some studies also argue for orthographic impairments (Luke et al., 2023). In humans it is not possible to selectively ablate the VWFA without broader changes, making this debate difficult to resolve. In contrast, models allow such targeted interventions. By ablating VWFA-selective units, we find that the resulting deficit is phonological, supporting the view that phonological dysfunction is primary, at least within the model. While this captures end-state consequences rather than developmental trajectories, and models lack explicit timing for speed-dependent measures, future work could simulate development under ablations or biased learning environments and approximate temporal dynamics (e.g., logit evolution) to probe reading speed.

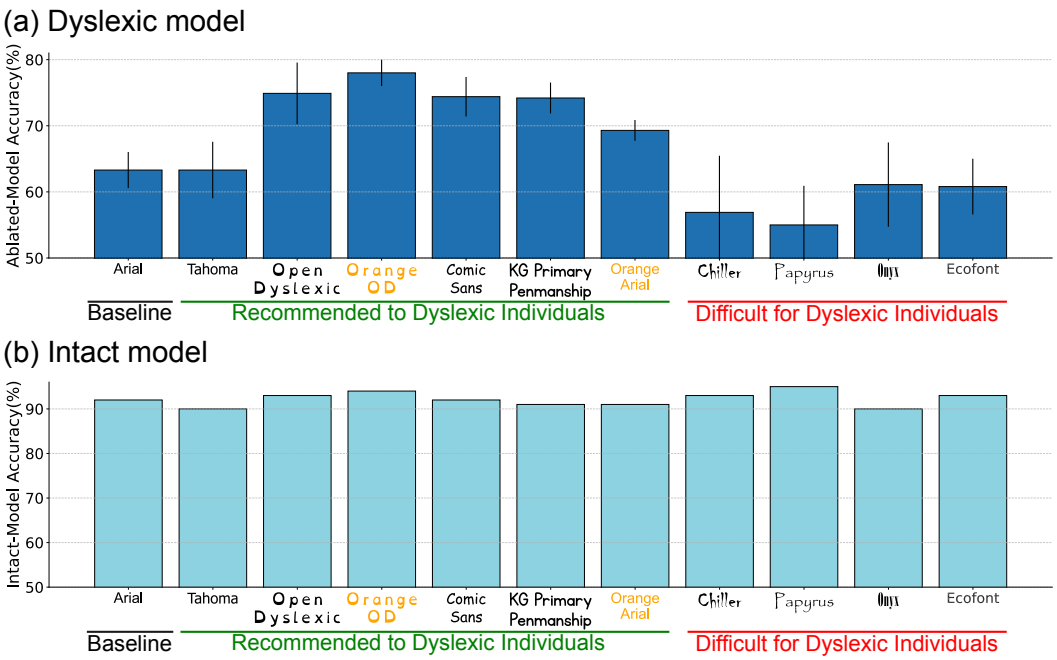


Figure 7: **Model predicts font-specific reading performance.** (a) Accuracy of the ablated ("dyslexic") model on ROAR lexical decision items across different fonts, averaged over 10 seeds. OpenDyslexic, OpenDyslexic on an orange background, Comic Sans, and KG Primary Penmanship showed significant improvement ($p \ll 0.001$), Orange Arial showed improvement ($p \ll 0.01$), and Papyrus showed significantly worse performance ($p \ll 0.01$). Error bars indicate 95% confidence intervals across seeds. (b) Accuracy of the intact model on the same items, averaged over 10 seeds; performance does not change significantly across fonts.

**Broader Applicability.** The approach presented here could, in principle, be adapted to investigate other brain disorders, such as schizophrenia or stroke. By inducing disorder-specific alterations in artificial neural networks, researchers can investigate the potential neural mechanisms underlying behavioral deficits, explore interventions in silico, and generate testable predictions for human studies. The three key steps – localization, ablation, and behavioral testing – can be replicated in brain disorders where contrast stimuli help localize hypothesized neural substrates, and behavioral benchmarks rigorously evaluate model predictions. This framework thus offers a flexible platform for studying psychiatric conditions beyond dyslexia. Disentangling the functional role of units via computational experimentation may further inform strategies for treatment. For example, insights into unit-level sensitivity to visual word features together with model prototyping could guide the design of fonts that improve reading performance in dyslexic readers.

## 7 Conclusion

By localizing and lesioning visual-word-form-selective units in vision-language models, this work provides a biologically inspired simulation of dyslexia, capturing its core reading-specific deficits while preserving broader cognitive functions. Our findings mirror empirical dissociations observed in dyslexic individuals and validate the functional relevance of the visual word form area. Our approach demonstrates how mechanistic manipulations of artificial neural network models can simulate selective cognitive deficits, offering a novel computational framework for testing hypotheses about the causal role of specific neural circuits in reading. Via the identification of model components whose ablation mimics dyslexic-like impairments, such in silico experiments could help test hypotheses about neural targets for early screening, suggest biomarkers for subtyping dyslexia (e.g., based on differential phonological versus orthographic deficits), and inform the development of targeted intervention strategies such as dyslexia-friendly fonts. This framework lays the groundwork for modeling the brain not just in health, but also in disease.

## REFERENCES


Paroma Agrawal, Maria Antoniak, Ethan B. Hanna, Bastien Bout, Devendra Chaplot, Jake Chudnovsky, David Costa, Basile De Monicault, Shrey Garg, Thibaut Gervet, Soumya Ghosh, Antoine Héliou, Pierre Jacob, Angie Q. Jiang, Kartikay Khandelwal, Timothee Lacroix, Guillaume Lample, Diego Lavinia Casas, Thibault Lavril, and Sida Yang. Pixtral 12b (version 2), 2024.

Vasileios Alevizos, Sean Edralin, Akalunga Simasiku, Despoina Malliarou, Antonios Messinis, Georgios Papakostas, Chengjun Xu, and Zhiqiang Yue. Handwriting anomalies and learning disabilities through recurrent neural networks and geometric pattern analysis (version 2). *arXiv*, 2024. DOI: 10.48550/ARXIV.2405.07238.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The llm language network: A neuroscientific approach for identifying causally task-relevant units, 2025. DOI: arXiv preprint.

K. L. Aw, S. Montariol, B. AlKhamissi, M. Schrimpf, and A. Bosselut. Instruction-tuning aligns llms to the human brain. *arXiv*, abs/2312.00575, 2023. DOI: 10.48550/arXiv.2312.00575.

Marta Boros, Jean-Louis Anton, Catherine Pech-Georgel, Jonathan Grainger, Marcin Szwed, and Johannes C Ziegler. Orthographic processing deficits in developmental dyslexia: Beyond the ventral visual stream. *NeuroImage*, 128:316–327, 2016. DOI: 10.1016/j.neuroimage.2016.01.014.

Silvia Brem, Urs Maurer, Martin Kronbichler, Matthias Schurz, Florian Richlan, Vera Blau, Johanna Reithler, Susanne van der Mark, Eva Schulz, Katrin Bucher, Kristina Moll, Karin Landerl, Ernst Martin, Rainer Goebel, Gerd Schulte-Körne, Leo Blomert, Heinz Wimmer, and Daniel Brandeis. Visual word form processing deficits driven by severity of reading impairments in children with developmental dyslexia. *Scientific Reports*, 10(1):1–12, 2020. DOI: 10.1038/s41598-020-75111-8.

H. R. Burke. Raven's progressive matrices: A review and critical evaluation. *The Journal of Genetic Psychology*, 93(2):199–228, 1958. DOI: 10.1080/00221325.1958.10532420.

Sebastian A. Cadena, George H. Denfield, Ethan Y. Walker, Leon A. Gatys, Anthony S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLOS Computational Biology*, 15(4):e1006897, 2019. DOI: 10.1371/journal.pcbi.1006897.

Camille Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, 12(1), 2022. DOI: 10.1038/s41598-022-20460-9.

Stanislas Dehaene and Laurent Cohen. The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6):254–262, 2011. DOI: 10.1016/j.tics.2011.04.003.

Changde Du, Kai Fu, Bo Wen, and et al. Human-like object concept representations emerge naturally in multimodal large language models. *Nature Machine Intelligence*, 7:860–875, 2025. DOI: 10.1038/s42256-025-01049-z.

Ayse Gokce and Martin Schrimpf. Scaling laws for task-optimized models of the primate visual ventral stream (version 2). *arXiv*, 2024. DOI: 10.48550/ARXIV.2411.05712.

Ariel Goldstein, Zohar Zada, Noam Buchnik, Maya Schain, Abigail Price, Bryn Aubrey, Samuel A. Nastase, Assaf Feder, David Emanuel, Asher Cohen, Anne Jansen, Hruday Gazula, Gabby Choe, Akshay Rao, Caroline Kim, Courtney Casto, Laura Fanda, William Doyle, Daniel Friedman, and Uri ... Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. DOI: 10.1038/s41593-022-01026-4.

Ehsan A. Hosseini, Martin Schrimpf, Yu Zhang, Samuel Bowman, Noam Zaslavsky, and Evelina Fedorenko. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63, 2024. DOI: 10.1162/nol_a_00137.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11): 4302–4311, 1997. DOI: 10.1523/jneurosci.17-11-04302.1997.

Daniel Kempler, Amit Almor, Lorraine K. Tyler, Elizabeth S. Andersen, and Mary C. MacDonald. Sentence comprehension deficits in alzheimer's disease: A comparison of off-line vs. on-line sentence processing. *Brain and Language*, 64(3):297–316, 1998. DOI: 10.1006/brln.1998.1980.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11):e1003915, 2014. DOI: 10.1371/journal.pcbi.1003915.

Lisa Kronbichler and Martin Kronbichler. The importance of the left occipitotemporal cortex in developmental dyslexia. *Current Developmental Disorders Reports*, 5(1):1–8, 2018. DOI: 10.1007/s40474-018-0135-4.

Rojina Kunwar and Hari Prasad Sapkota. An overview of dyslexia: Some key issues and its effects on learning mathematics. *Turkish International Journal of Special Education and Guidance & Counselling (TIJSEG)*, 11(2):82–98, 2022.

Joseph Hin Yan Lam and Shelley Xiuli Tong. Drawing a new picture: Children with developmental dyslexia exhibit superior nonverbal creativity. *Research in Developmental Disabilities*, 116: 104036, 2021. DOI: 10.1016/j.ridd.2021.104036.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. Available at: https://llava-vl.github.io/blog/2024-01-30-llava-next/, January 2024.

Björn Lonnqvist, Emmanuel Scialom, Ayse Gokce, Zachary Merchant, Michael H. Herzog, and Martin Schrimpf. Contour integration underlies human-like vision (version 2). *arXiv*, 2025. DOI: 10.48550/ARXIV.2504.05253.

Steven G. Luke, Thomas Brown, Clara Smith, Ana Gutierrez, Christopher Tolley, and Olivia Ford. Dyslexics exhibit an orthographic, not a phonological deficit in lexical decision. *Language, Cognition and Neuroscience*, 39(3):330–340, 2023. DOI: 10.1080/23273798.2023.2288319.

G. Reid Lyon, Sally E. Shaywitz, and Bennett A. Shaywitz. A definition of dyslexia. *Annals of Dyslexia*, 53(1):1–14, 2003. DOI: 10.1007/s11881-003-0001-9.

Erez Margalit, Hyodong Lee, David Finzi, James J. DiCarlo, Kalanit Grill-Spector, and Daniel L. K. Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451.e7, 2024. DOI: 10.1016/j.neuron.2024.04.018.

Ali I. Marvi, Simon Hutchinson, Evelina Fedorenko, Rebecca R. Saxe, Florian S. Kamps, Tal I. Regev, Emily M. Chen, and Nancy G. Kanwisher. An efficient multifunction fmri localizer for high-level visual, auditory, and cognitive regions in humans. *Imaging Neuroscience (Camb)*, 3: IMAG.a.905, 2025. DOI: 10.1162/IMAG.a.905.

Urs Maurer, Sarah Brem, Kerstin Bucher, Franz Kranz, Rene Benz, Hans-Christoph Steinhausen, and Daniel Brandeis. Impaired tuning of a fast occipito-temporal response for print in dyslexic children learning to read. *Brain*, 130(12):3200–3210, 2007. DOI: 10.1093/brain/awm193.

Bruce D. McCandliss, Laurent Cohen, and Stanislas Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, 7(7):293–299, 2003. DOI: 10.1016/s1364-6613(03)00134-7.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt (version 5). *arXiv*, 2022. DOI: 10.48550/ARXIV.2202.05262.

J. L. Mitchell, M. Yablonski, H. L. Stone, M. Fuentes-Jimenez, M. E. Takada, K. A. Tang, J. E. Tran, C. Chou, and J. D. Yeatman. Small or absent visual word form area is a trait of dyslexia. *bioRxiv*, pp. 2025.01.14.632854, Jan 2025. doi: 10.1101/2025.01.14.632854. Preprint.

Karine Monzalvo, Joanna Fluss, Claire Billard, Stanislas Dehaene, and Ghislaine Dehaene-Lambertz. Cortical networks for vision and language in dyslexic and normal children of variable socio-economic status. *NeuroImage*, 61(1):258–274, 2012. DOI: 10.1016/j.neuroimage.2012.02.035.

V. I. Müller, E. C. Cieslik, I. Serbanescu, A. R. Laird, P. T. Fox, and S. B. Eickhoff. Altered brain activity in unipolar depression revisited. *JAMA Psychiatry*, 74(1):47, 2017. DOI: 10.1001/jamapsychiatry.2016.2783.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fmri. *NeuroImage*, 56(2):400–410, 2011. DOI: 10.1016/j.neuroimage.2010.07.073.

Hiroki Ogawa, Shuntaro Ogoshi, Yuto Ogoshi, and Ayaka Nakai. Can deep generative models explain brain function in people with developmental dyslexia? *Electronics*, 12(10):2305, 2023. DOI: 10.3390/electronics12102305.

O. A. Olulade, E. M. Napoliello, and G. F. Eden. Abnormal visual motion processing is not a cause of dyslexia. *Neuron*, 79(1):180–190, 2013. DOI: 10.1016/j.neuron.2013.05.002.

Eraldo Paulesu, Jean-François Démonet, Ferruccio Fazio, Emily McCrory, Virginie Chanoine, Nicola Brunswick, Stefano F. Cappa, Giuseppe Cossu, Michel Habib, Christopher D. Frith, and Uta Frith. Dyslexia: Cultural diversity and biological unity. *Science*, 291(5511):2165–2167, 2001. DOI: 10.1126/science.1057179.

Bruce F. Pennington, Laurie M. McGrath, Jennifer Rosenberg, Holly Barnard, Sandra D. Smith, Erik G. Willcutt, Abigail Friend, John C. DeFries, and Richard K. Olson. Gene × environment interactions in reading disability and attention-deficit/hyperactivity disorder. *Developmental Psychology*, 45(1):77–89, 2009. DOI: 10.1037/a0014549.

Robin L. Peterson and Bruce F. Pennington. Developmental dyslexia. *Annual Review of Clinical Psychology*, 11(1):283–307, 2015. DOI: 10.1146/annurev-clinpsy-032814-112842.

Nishant Rathi, Jennifer Mehrer, Bassel AlKhamissi, Turki Binhuraib, Nathan M. Blauch, and Martin Schrimpf. Topolm: brain-like spatio-functional organization in a topographic language model (version 3). *arXiv*, 2024. DOI: 10.48550/ARXIV.2410.11516.

J. Roitsch and S. Watson. An overview of dyslexia: Definition, characteristics, assessment, identification, and intervention. *Science Journal of Education*, 7(4):81, 2019. DOI: 10.11648/j.sjedu.20190704.11.

Zeynep M. Saygin, Daniel E. Osher, Elizabeth S. Norton, David A. Youssoufian, Stephanie D. Beach, Jennifer Feather, Nadine Gaab, John D. E. Gabrieli, and Nancy Kanwisher. Connectivity precedes function in the development of the visual word form area. *Nature Neuroscience*, 19(9): 1250–1255, 2016. DOI: 10.1038/nn.4354.

Martin Schrimpf, Jonas Kubilius, Haider Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, K. Kar, Pouya Bashivan, Jennifer Prescott-Roy, Fabian Geiger, Katrin Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *Cold Spring Harbor Laboratory*, 2018. DOI: 10.1101/407007.

Martin Schrimpf, Jonas Kubilius, Malin J Lee, Neel A Ratan Murty, Rani Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020. DOI: 10.1016/j.neuron.2020.07.040.

Martin Schrimpf, Idan A. Blank, Greta Tuckute, Carina Kauf, Evelina A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021. DOI: 10.1073/pnas.2105646118.

Bennett A Shaywitz, Sally E Shaywitz, Kenneth R Pugh, W Eric Mencl, R Kirk Fulbright, Pawel Skudlarski, R Todd Constable, Kelly E Marchione, Jack M Fletcher, G Reid Lyon, and John C Gore. Disruption of posterior brain systems for reading in children with developmental dyslexia. *Biological Psychiatry*, 52(2):101–110, 2002. DOI: 10.1016/s0006-3223(02)01365-3.

A. J. Sihvonen, P. Virtala, A. Thiede, M. Laasonen, and T. Kujala. Structural white matter connectometry of reading and dyslexia. *NeuroImage*, 241:118411, 2021. DOI: 10.1016/j.neuroimage.2021.118411.

Giorgia Silani, Uta Frith, Jean-François Demonet, Ferruccio Fazio, Daniela Perani, Cathy Price, Christopher D. Frith, and Eraldo Paulesu. Brain abnormalities underlying altered activation in dyslexia: a voxel based morphometry study. *Brain*, 128(10):2453–2461, 2005. DOI: 10.1093/brain/awh579.

Margaret J. Snowling, Charles Hulme, and Kate Nation. Defining and understanding dyslexia: past, present and future. *Oxford Review of Education*, 46(4):501–513, 2020. DOI: 10.1080/03054985.2020.1765756.

Christian J. Spoerer, Tim C. Kietzmann, Jennifer Mehrer, Isabelle Charest, and Nikolaus Kriegeskorte. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, 16(10):e1008215, 2020. DOI: 10.1371/journal.pcbi.1008215.

K. E. Stanovich. The future of a mistake: Will discrepancy measurement continue to make the learning disabilities field a pseudoscience? *Learning Disability Quarterly*, 28(2):103–106, 2005.

H. Tanaka, J. M. Black, C. Hulme, L. M. Stanley, S. R. Kesler, S. Whitfield-Gabrieli, A. L. Reiss, J. D. E. Gabrieli, and F. Hoeft. The brain basis of the phonological deficit in dyslexia is independent of iq. *Psychological Science*, 22(11):1442–1451, 2011. DOI: 10.1177/0956797611419521.

Yujia Tang, Ayse Gokce, Khalid J. Al-Karkari, Daniel Yamins, and Martin Schrimpf. Many-two-one: Diverse representations across visual pathways emerge from a single objective. *Cold Spring Harbor Laboratory*, 2025. DOI: 10.1101/2025.07.22.664908.

Mariya Toneva, Alessandro Sordoni, R. T. des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning (version 3). *arXiv*, 2018. DOI: 10.48550/ARXIV.1812.05159.

Gabriela Tuckute, Aniruddh Sathe, Srini Srikant, Matthew Taliaferro, Michael Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behavior*, 2024. DOI: 10.1038/s41562-023-01783-7.

Peter E. Turkeltaub, Eyal M. Goldberg, Warren A. Postman-Caucheteux, Michelle Palovcak, Claire Quinn, Cara Cantor, and H. Branch Coslett. Alexia due to ischemic stroke of the visual word form area. *Neurocase*, 20(2):230–235, 2013. DOI: 10.1080/13554794.2013.770873.

S. Valdois. Dyslexia (developmental). In *Encyclopedia of Behavioral Neuroscience*, pp. 454–460. Elsevier, 2010. DOI: 10.1016/b978-0-08-045396-5.00043-9.

Csaba von Károlyi, Ellen Winner, Wendy Gray, and Gary F. Sherman. Dyslexia linked to talent: Global visual-spatial ability. *Brain and Language*, 85(3):427–431, 2003. DOI: 10.1016/s0093-934x(03)00052-x.

Richard K. Wagner, Frank A. Zirps, Anna A. Edwards, Steven G. Wood, Rebecca E. Joyner, Betsy J. Becker, Guanjie Liu, and Brittany Beal. The prevalence of dyslexia: A new approach to its estimation. *Journal of Learning Disabilities*, 53(5):354–365, 2020. DOI: 10.1177/0022219420920377.

A. Y. Wang, K. Kay, T. Naselaris, M. J. Tarr, and L. Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, 2023. DOI: 10.1038/s42256-023-00753-y.

Reinhard Werth. Dyslexia: Causes and concomitant impairments. *Brain Sciences*, 13(3):472, 2023. DOI: 10.3390/brainsci13030472.

Daniel L. K. Yamins, Haider Hong, Charles F. Cadieu, Evan A. Solomon, Daniel Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. DOI: 10.1073/pnas.1403112111.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. DOI: 10.1038/nn.4244.

An Yang et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Jason D. Yeatman, Kevin A. Tang, Patrick M. Donnelly, Max Yablonski, Mohit Ramamurthy, Iliana I. Karipidis, Sendy Caffarra, M. Elizabeth Takada, Kaitlyn Kanopka, Michal Ben-Shachar, and Benjamin W. Domingue. Rapid online assessment of reading ability. *Scientific Reports*, 11 (1), 2021. DOI: 10.1038/s41598-021-85907-x.

Chuang Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning (version 1), 2019. DOI: 10.48550/ARXIV.1903.02741.

Z. Zhang, Y. Lin, Z. Liu, P. Li, M. Sun, and J. Zhou. Moefication: Transformer feed-forward layers are mixtures of experts. *arXiv*, abs/2110.01786, 2021. DOI: 10.48550/arXiv.2110.01786.

Chaithanya Zhuang, Song Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), 2021. DOI: 10.1073/pnas.2014196118.

# A Appendix

## A.1 Other Models

We applied our VWFA localization and ablation methodology across two other vision-language models: LLaVA-NeXT-13B(Liu et al., 2024), and PixTral-12B(Agrawal et al., 2024). LLaVA-NeXT-13B showed particularly low accuracy on Raven's Progressive Matrices ($16.67\%$, below chance-level performance which is $20\%$), making it unsuitable for evaluating the selective preservation of visual reasoning after ablation. While VWFA ablation in LLaVA-NeXT did produce the expected ROAR performance reduction, the model's unreliable baseline on nonverbal reasoning tasks prevented meaningful interpretation of whether this deficit was reading-specific (Fig. 8a).

PixTral-12B demonstrated robust baseline performance on ROAR: $86.75\%$. After ablating the VWF-selective units, ROAR performance decreased, whereas the change in RAVEN performance was not statistically significant, suggesting that the method can be generalized to other models.

## A.2 Visual IQ and Reasoning Benchmarks

Figure 9 shows an example of the non-verbal benchmarks. More examples of Kempler are shown in Figure 10.

## A.3 Prompts Used Across All Experiments

For transparency and reproducibility, we include the exact prompts used across all lexical, phonological, and nonverbal reasoning benchmarks. These prompts were created with the explicit goal of mimicking how human participants are typically instructed to perform these tasks, ensuring that the model receives directions that parallel standard experimental protocols.

**Lexical Decision Prompt.**

```
A real or pseudo word will be presented to you in
an image.  The pseudo words might look like English
words, but they don't mean anything in English.  For
example, laip, bove or cigbert are pseudo words.  The
real words will be ones you recognize.  They are real
English words like is, or basket, or lion.  Please
answer the following question:  Is the word in the
image a real word or a pseudo word?
```

**KEMPLER Prompt.**

```
In which image [DESCRIPTION OF TARGET IMAGE]? After
providing the reason, give your final answer in this
format:  "The answer is picture a" or "The answer is
picture b".
```

**RAVEN Prompt.**

```
You will be presented with a nonverbal Raven's
Progressive Matrices IQ puzzle.  The puzzle will
consist of a visual pattern with one missing element,
indicated by a question mark.  Your task is to
identify which of the five provided options best
completes the pattern.  Please choose the option that
logically fits the sequence.  Write the answer in one
digit in [1, 2, 3, 4, 5].
```

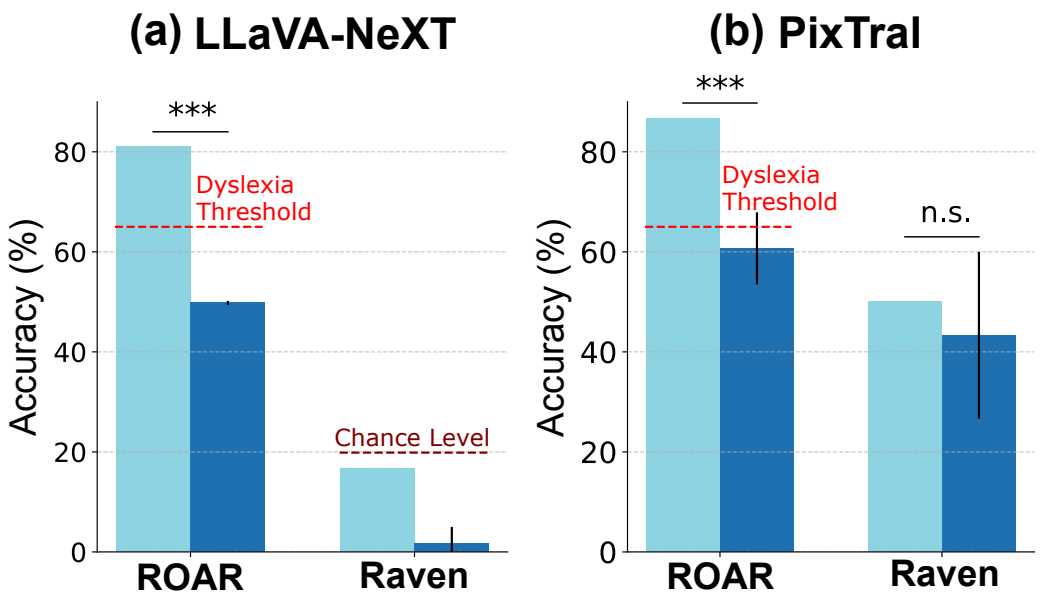


Figure 8: **Ablating VWF-selective units in other vision-language models.** Plots show accuracy of the intact and ablated models on the ROAR and RAVEN benchmarks. Bars indicate mean accuracy over 5 random seeds, with error bars representing 95% confidence intervals. For LLaVA-Next **(a)**, ablation of self-attention and MLP layers from the language decoder reduces ROAR performance below the dyslexia threshold (red dashed line); however, RAVEN performance was already below chance before ablation, so no conclusions about reasoning selectivity can be drawn. For PixTral **(b)**, ablation of MLP layers from the language decoder significantly impairs ROAR performance, while the change in RAVEN performance is not statistically significant, i.e. the deficits are selective for reading.

### A.4 Statistical Testing

All model results are reported as averages over 20 random seeds, each corresponding to a different random sample of the localizer stimuli. To evaluate whether ablations reduced performance, we compared the 20 ablated values (one per seed) against the intact model's baseline using a one-tailed Student's $t$-test. The null hypothesis was that mean ablated performance equals baseline, and the alternative was that mean ablated performance is lower. A one-tailed test was chosen because our hypothesis was directional: ablations were expected to impair, but not improve, task performance.

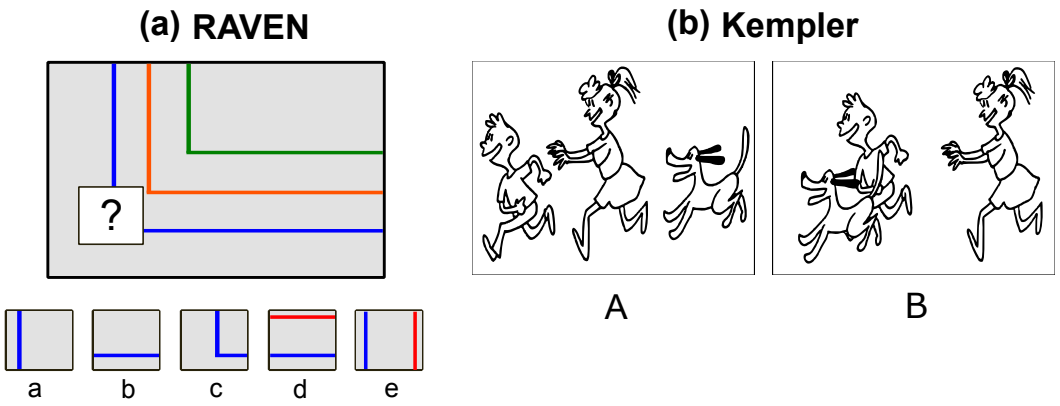


Figure 9: **Examples of IQ and visual reasoning benchmarks (a)** Example item from the RAVEN visual reasoning test. The task is to identify the panel (from five candidates) that correctly completes the matrix by following the underlying visual pattern. In this example, the correct answer is Option 3. RAVEN serves as a measure of general visual-spatial reasoning, independent of language ability. **(b)** Example item from the Kempler's Sentence Comprehension Test used in our model evaluation. Given a caption and a pair of images (A: left, B: right), the model is prompted to identify which image correctly depicts the described event. In this example, the model is asked: "In which image (A or B) the dog that chases the girl is following the boy?" (correct answer option A.)

Reported $t$-statistics and $p$-values reflect this comparison for each benchmark. Intriguingly, a two-tailed Student's $t$-test, testing bidirectional changes of value distributions, revealed a statistically significant increase on the Kempler visual reasoning task.

### A.5 Computational resources

All experiments were run on two NVIDIA A100 GPUs with 80GB memory each. The ablation implementation does not introduce meaningful overhead.

### A.6 Use of LLM Assistance

We used *ChatGPT (GPT-5, OpenAI, 2025)* to aid in structuring and polishing the writing of this document. The authors reviewed, edited, and are fully responsible for the final content.

Table 1: **Common error patterns of the VWF-ablated model.** *Blank*: Model gave no output at all (completely empty response). *Misclassification*: Model explicitly labeled a real word as pseudo, or a pseudo word as real. *Contextual Over-Interpretation*: Model invoked external references (languages, acronyms, domains) to justify the wrong answer. *Ambiguous Hedging*: Model refused to commit, saying the word could be real or pseudo depending on context. *Gibberish / Corrupted*: Output was meaningless fragments, random characters, non-English text, or truncated nonsense. More examples of each pattern in Table 2.

| Response type | Sample Word Image | Example Model Output |
|---|---|---|
| Blank | `accustomed` (real) | – |
| Misclassification | `yammerring` (real) | The word "yammerring" looks like a pseudo word. |
| Contextual Over Interpretation | `hus` (pseudo) | The word "hus" in the image is a real word in English. It is a noun that means "husk" which is a type of edible corn. |
| Ambiguous Hedging | `dood` (pseudo) | The word "dood" appears in the image. It seems to be a pseudo word because it has no meaning in English. However, it could potentially be a real word if it were part of a specific context or phonetic representation, for which further investigation would be necessary. |
| Gibberish / Corrupted | `imeyits` (pseudo) | image of the image of the image of the image of the following. |

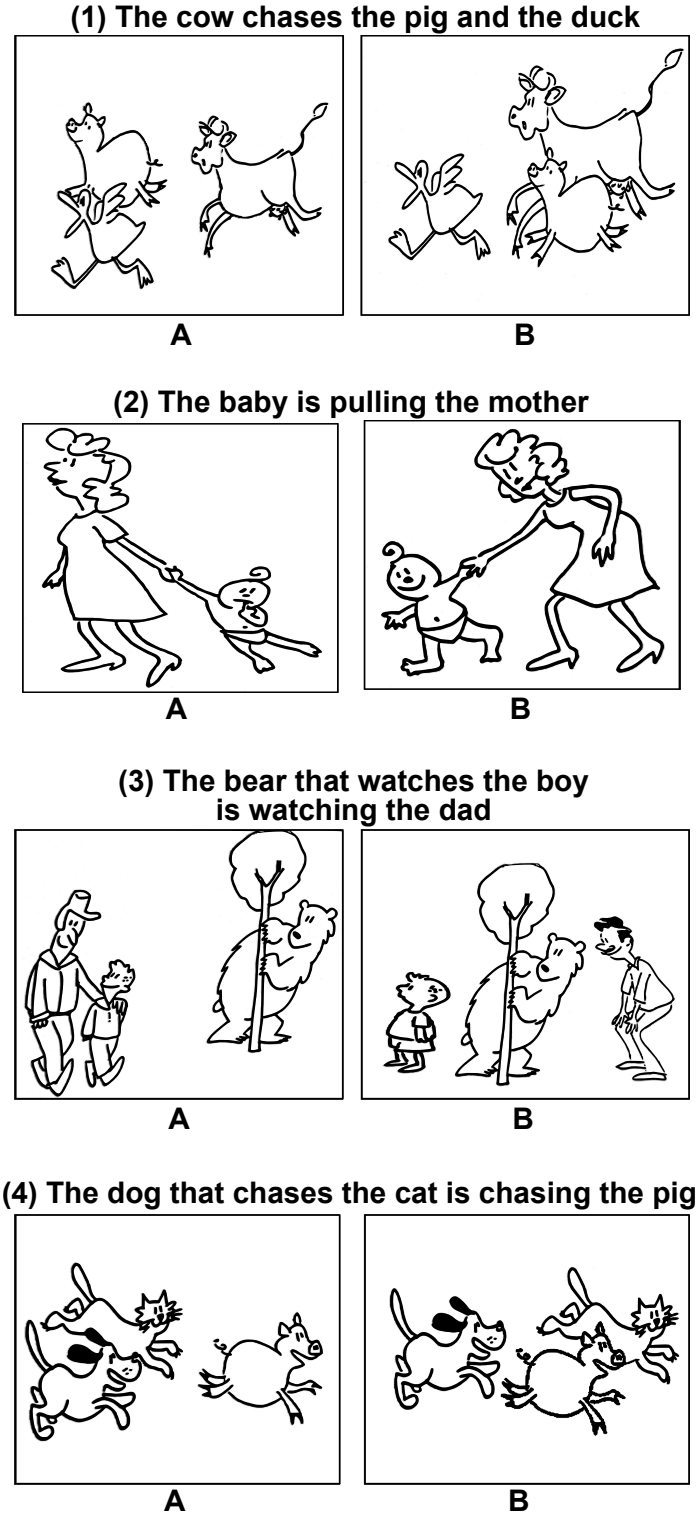


Figure 10: **Examples of IQ and visual reasoning benchmarks** Several items from Kempler's Sentence Comprehension Test, which evaluates the ability to map linguistic descriptions onto visual scenes. In each case, the model is given a caption and two candidate images (labeled A: left, B: right) and must decide which picture matches the sentence. Shown here are four sample items with increasing syntactic complexity. The correct answers for Items 1–4 are: A, B, A, B.

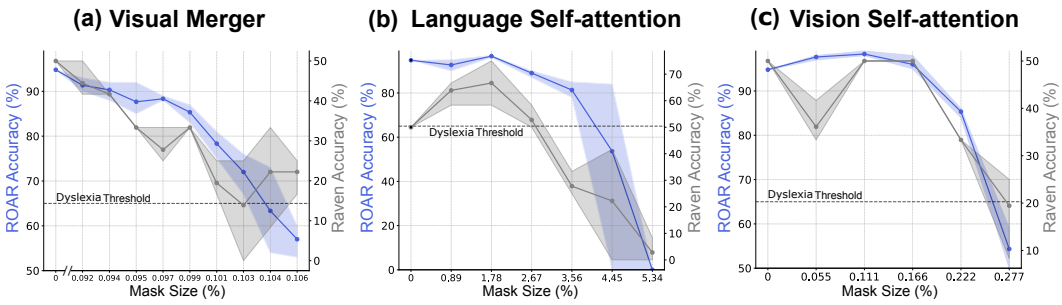


Figure 11: **Ablating units in non-selective layers does not result in selective impairments.** Testing other layers in the **(a)** visual merger (projection layers), **(b)** language decoder, and **(c)** vision encoder shows that RAVEN performance declines faster than ROAR performance indicating the effect is not reading-selective. Intriguingly, the self-attention layers show an inverse effect, with potential selectivity for RAVEN over ROAR. Shaded regions around each curve represent the 95% confidence intervals.

Table 2: **Further examples of wrong model responses to ROAR when VWF-selective units are ablated.** (Continued from Table 1.) Errors span misclassification, contextual over-interpretation, ambiguous hedging, and gibberish/corrupted outputs.

| Response type | Sample word image | Example model output |
|---|---|---|
| Misclassification | charring(real) | The word in the image is a pseudo word. |
| | creat(pseudo) | The word "creat" is a real word in English. It can be used as a verb, such as "to creat something." |
| | rackle(pseudo) | The word in the image is a real word. |
| Contextual Over-Interpretation | che(pseudo) | The word "che" is a real word in English. It is used as a conjunction in Italian grammar and also as a noun in English, referring to a type of cheese. |
| | hu(pseudo) | The word "hu" can be considered a real word in some contexts. It is used in Hungarian names and titles, such as Budapest University of Economics (HU). |
| | mich(pseudo) | The word "mich" in the image appears to be a real word... It seems to be a play on the word "Michigan." |
| | albust(pseudo) | The word "albust" is a real word in English, but it doesn't have any meaning associated with it. It appears to be a play on the word "albus." |
| Ambiguous Hedging | bi(pseudo) | The word "bi" can be a real word or a pseudo word depending on the context. It could mean "two" as a prefix, but alone it may not count as a word. |
| | ase(pseudo) | The word "ase" can be interpreted as a real word or a pseudo word depending on its context. If it stands alone, it is a pseudo word, but in words like "phrase" or "database" it is real. |
| | yos(pseudo) | The word "yos" appears to be a real word. It seems to be a part of a sentence or phrase, suggesting it could be meaningful in English. However, its full meaning is unclear. |
| Gibberish / Corrupted | napery(real) | (Chinese characters). |
| | pumer(pseudo) | (Chinese characters) |
| | cowbim(pseudo) | real words will bepseudo words will bepseudo words |
| | seashell(real) | the image of the image words |

Table 3: **Examples of wrong model responses to ROAR when the VWF-Selective units are hyperactivated by 3.25 times.** Sentences collapse with random numbers, irrelevant fragments, or broken logic, confirming that hyperactivating VWF-selective units does not replicate the patterns of impairment observed in humans.

| Model input | Model output |
|---|---|
| `find(real)` | The word "find" cannot be found in the image, according see me more words 123 6th of March 3, 2027, 5:42: :00—05 no-13 2016 12:00 am I am very well-/.c_0_2020_160_50_00000_132 5028_0_27_001_0_0000000. |
| `in(real)` | The word "in" is actually 1 letters. To get the maximum number of letters, you would have to look at the middle of the word. The text in the middle of this, we hope, can be as much as a as the s. The machine can be working part of its body (p) or a part.pellp the verb 'to' in[Ref#p=truly/500 word no string of fine, in which case it becomes a verb in the text of the image content. H/s=10000001/ <t>the act of |
| `never(real)` | The word "neve" completing the word 'never' is 5 characters long and 100000000 times as long as the word 'I never, by it welcome back after the 'cat' 'never'} |
| `telescoped(real)` | The word "telescopically" is a word only one time. The word "telescoped" can be seen in a different order. How would you support this statement with less than 10000 words at a website with a word count factoring the perfect total of each 4 and 2 before it. . 4,500 E-17. The last time with a similar tag and its own 2000.3999999999993 times [2000.1804.25000.8 |
| `everyone(real)` | The word "everyone" is not a real word. It was just a fun thing to think about. If a word is a real word and it's a word we've all seen before and the "to" means that we don't have to right now but it's a real word because it can be seen and not for other people to be used, they can't be naturally occurring as it if even where none of the time. It's then the word "eves" can be a word with one or two syllate, e.g., "be" can be used by the "best" and "you" (2 |