Developing Efficient Human-Like Memory Systems: Diagnosing the Limitations of Artificial Working Memory under Visual Interference and Out-of-Distribution Inputs

Shi YanZhe
College of Computing and Data Science

Asst Prof Zhang Mengmi
College of Computing and Data Science

Abstract - Developing efficient human-like memory systems requires understanding of not only when artificial working memory matches human working memory but also when it breaks down. The accuracy of a Clean benchmark provides an incomplete measurement as a model can score perfectly but fails under conditions that the human brain can handle naturally. We diagnose these failures that occur in a pretrained recurrent model of LSTM-96, from the Working Memory (WorM) benchmark. The LSTM-96 model is then tested zero-shot on a visual Change Detection (CD) task. In this study, there are three stages that are used to investigate the limitations of the model. Stage 1 introduces visual interference into the retention interval (RI). Stage 2 perturbs the Memory and Probe arrays while keeping RI clean by either recolouring the objects (to novel hues) or the background. Stage 3 tests the model with different contrast of the retention distractor. On clean trials, the model is almost perfectly accurate, however, when retention interference is introduced, accuracy collapses towards chance with a strong “Changed” response bias. When only the novel objects change in colour in Stage 2, the model moderately retains its accuracy with a slight decrease. However, when only the background changes, accuracy collapses towards chance. The Stage 3 experiment reveals a contrast dependent collapse as accuracy of the model decreases proportionally to degree of contrast. These failures suggest that though the model possesses some human working memory traits, it ultimately fails at resisting irrelevant interference and generalization through reasoning. Source code: link.

Keywords - working memory; change detection; out-of-distribution generalisation; visual interference

1 INTRODUCTION

Working memory (WM) is the cognitive system that temporarily holds and manipulates a limited amount of information so as to reason, comprehend and make decisions [1], [2]. To build an artificial memory system that replicates that of a human WM is a long standing goal, and the progress towards it depends on the understanding of their similarities and differences. This study attempts to take a diagnostic approach to investigate the limitations of artificial WM. It does this by asking how, when and why the model fails to replicate human WM instead of asking only how accurately the model performs on clean inputs. The reason for this is because the structure of failure is often more informative about the underlying mechanisms instead of successful outputs.

By a human-like WM system, this study defines it as one that not only stores and retrieves information but also makes a reasonable generalization of the materials provided despite the interference introduced. Such a system should be able to identify distractors that occur and not allow them to affect its judgement. We introduce out-of-distribution (OOD) transformations to evaluate the model's reaction when mild interference is applied. By doing so, we can characterise where the WM model is brittle and design better ones, which is the purpose of this study.

Recent benchmarks make such comparisons concrete. The Working Memory (WorM) benchmark jointly trains recurrent and transformer models on ten WM tasks and shows that these models reproduce qualitative human WM signatures, such as the primacy and recency effects, with perfect accuracy on clean trials [3]. Accuracy on clean trials alone, however, is not a good indicator of robustness, as a large literature shows that high in-distribution accuracy does not guarantee robustness against small adversarial perturbations [4], [5], common corruptions [6], and reliance on non-transferable cues due to shortcut learning [7],[8]. Notably, the WorM study itself already identified one such limitation. By replacing the grey retention frames with random noise distractors, the recurrent model’s Change Detection (CD) accuracy decreases drastically [3]. We take this documented vulnerability of the LSTM-96 model (a recurrent network with a Long-Short-Term-Memory (LSTM) backbone of capacity 96) as our point of departure.

We evaluate LSTM-96 zero-shot (no retraining, fine-tuning, or augmentation) on the visual CD task, using three stages of controlled and label preserving experiments that target different points in the trial. This design allows us to identify the reason for the failure, be it the maintenance of

information during delay or its perceptual encoding. We organise the study around three questions:

1. Does irrelevant visual input during the retention interval (RI) phase disrupt maintenance and, if it does, is the resulting failure structured?
2. When the RI is protected, what is the difference between the model's robustness when a change in a task-relevant object is introduced as compared to a change in task-irrelevant background?
3. What is the main contributor to the failure of the model during Stage 1? Is it the contrast or the saliency of the distractor?

Our contributions are:

I. A three stage, zero-shot diagnostic protocol that differentiates retention interval interference from input-distribution shift on one model and task.
II. Evidence that retention interference collapses the model to chance with a strong bias to respond "Changed".
III. A dissociation in which the model tolerates novel object colours but collapses when background colour changes.
IV. A contrast sweep which shows that the RI failure is graded and saliency dependent.

The remainder of the paper reviews related work (Section 2), describes the model, task and design (Section 3), reports results for the three stages (Section 4), interprets them (Section 5), states limitations (Section 6), and concludes (Section 7).

2 BACKGROUND AND RELATED WORK

2.1 HUMANA AND COMPUTATIONAL WORKING MEMORY

Baddeley and Hitch's multi-component model provides a useful foundation: it describes working memory as a limited-capacity system with separate verbal and visuospatial stores coordinated by a central executive [1], later extended by Baddeley with an episodic buffer that binds information into integrated objects [9]. Human WM is small (on the order of a very limited number of items) [10], and the visual CD paradigm shows that it stores information as integrated objects instead of independent features [11], with a neural correlate that tracks the number of items maintained [12]. The correct binding of these isolated features into an integrated object depends on attention [13], although the exact limits of storage capacity and the extent to which observed behaviour underdetermines underlying mechanisms remain subjects of debate [2], [14]. Despite these theoretical debates, a recurring property of human WM is that it is decently resistant to the impact of task-irrelevant interference visual interference.

This property of human WM motivates the research premise of this paper. Since interference has an insignificant impact on maintenance in human WM, a model that has a human-like memory system should be able to perform well when it is faced with task-irrelevant interference. Our Stage1, 2, and 3 hence test the model's limit and investigate when the system fails to replicate human WM behaviour.

On the computational side, sequence memory models range from simple recurrent network [15] to gated variants such as the LSTM [16] and GRU [17], memory augmented networks [18], and attention based transformers [19], within the broader progress of deep learning [20]. The WorM benchmark places such backbones in a shared framework by using a convolutional encoder, a recurrent or transformer memory module and task specific heads trained jointly across all tasks [3]. The model name in this case is defined by its backbone and capacity (LSTM-96 is the LSTM with 96 hidden units, trained with the Adam optimiser [21]). These models provide a controlled environment such that the same architecture can be probed across many memory tasks.

2.2 ROBUSTNESS AND OUT-OF-DISTRIBUTION GENERALISATION

A model with a human-like memory system must be reliable even under interference and reasonable shifts of its input distribution. Despite this, deep networks are frequently brittle such that imperceptible adversarial perturbations can change their predictions [4], [5], common corruptions such as noise and blur degrade them [6] and they often rely on superficial statistics. For example, they rely on texture rather than shape [7], which does not transfer under distribution shift, indicating an instance of shortcut learning [8]. Our manipulations are designed such that each transformation is applied identically and deterministically to the Memory and Probe arrays, so that the Same/Changed ground truth is preserved, and any failure reflects the model's encoding and maintenance rather than a corrupted label. This diagnosis of the model's failure is a concrete step towards building an efficient human-like memory system.

3 METHODS

3.1 MODEL AND TASKS

We evaluate the pretrained LSTM-96 model from WorM without modification [3]. Each $32 \times 32 \times 3$ frame is encoded by a four layer convolutional neural network (CNN), and is projected to a feature vector, concatenated with a learnable task

embedding, and processed by a single layer LSTM with 96 hidden units. The task specific linear heads then read out the hidden state. The weights of the model are frozen and the network is run in evaluation mode, therefore, every result is zero-shot. In the CD task, a Memory array of bars (varying in Colour, Orientation, Size, and Gap) is shown, followed by a RI (the delay frames between the Memory and Probe arrays) and then a Probe array that is either identical to the Memory array or differs in one bar along a single feature. Finally, the model makes a binary judgement with the encoded "0" indicating "Same" and "1" indicating "Changed". The task has five variants determining which feature changes (Colour, Orientation, Size, Gap, and Conjunction). Our analysis uses the delayed trials (RI>0).

The set size for the task ranges from 2 to 12 items and the retention interval over 0, 6, 12, and 18 frames. In the Conjunction variant the changed bar may differ along any one of the four features, this then forces the model to hold the full feature combination of every bar rather than a single attribute. We restrict all analyses to delayed trials so that a genuine retention interval exists and can be either filled with interference (Stage 1 and 3) or protected while the arrays are perturbed (Stage 2).

3.2 THREE STAGE DIAGNOSTIC DESIGN

The three stages perturb different points in the trial. In Stage 1, the Memory and Probe arrays are unchanged and only the retention frames are replaced at full contrast with one of the four interference patterns (random static, a checkerboard, horizontal or vertical gratings) (Fig. 1). This is tested against a clean grey baseline adopted from the original WorM benchmark dataset. As such the results of the test is therefore attributable to maintenance during delay and nothing else. In Stage 2, only the Memory and Probe array undergo transformations (Fig 2) such that (i) Clean_Grey is the baselines, (ii) OOD_Object_Recolor rotates the object hue by 60°, shifting from the red/green palette used in the WorM benchmark testset to yellow/cyan while leaving the grey background unchanged, (iii) OOD_Background_Recolor recolours the grey background to a steel blue colour while leaving the test objects unchanged. In stage 3, the retention frames are filled with random static whose contrast (amplitude) is scaled by a factor that ranges from 0 (clean grey) to 1 (full contrast static), interpolating between the clean and Stage 1 conditions (Fig 3).

The interference patterns and transformations are simple by design. In Stage 1, the delay frames are filled at full contrast by either regular patterns (checkerboards and gratings) or uniform noise (random static). The Stage 2 object recolouring only mildly changes the colour of the object while keeping the background consistent (same as the training set in the WorM benchmark dataset). Likewise, for background recolouring, only the background is recoloured while the object remains the same as it does in the training set. The Stage 3 distractor blends the full contrast static toward the measured grey background, so its contrast can be changed continuously from zero to one. Because Stage 1 mirrors the noise during the retention probe that is already reported in the WorM benchmark [3], it therefore bridges prior work and the new diagnostics.

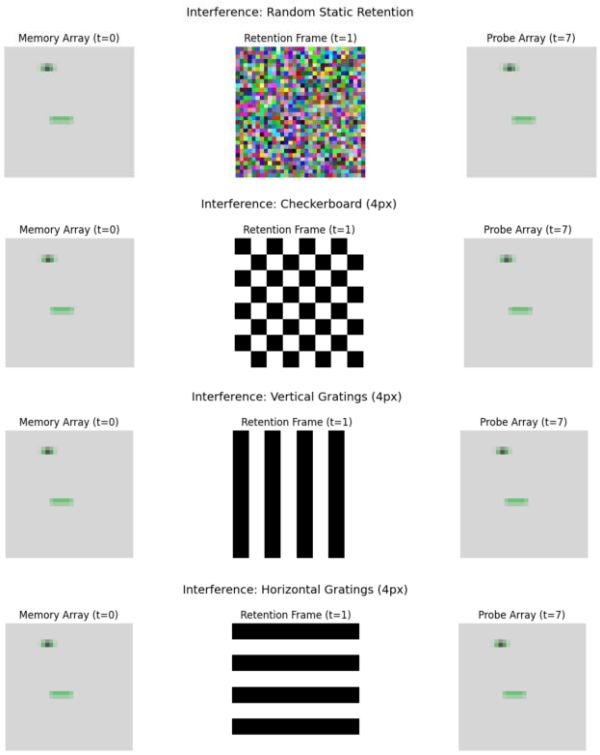
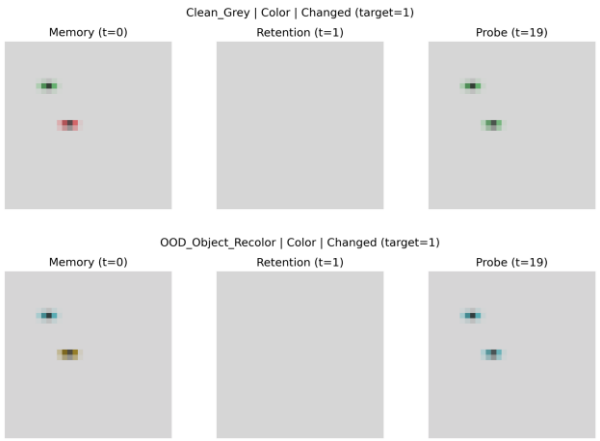


Fig 1 Examples of interference pattern for Stage 1 (random static, a checkerboard, horizontal or vertical gratings)



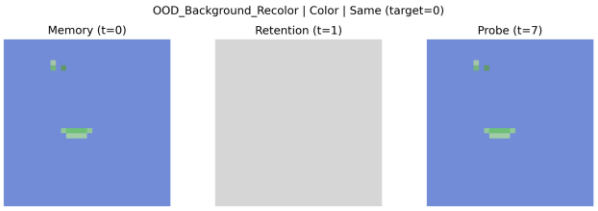


Fig. 2. Stage 2 example trials (Color variant), top to bottom: Clean_Grey, OOD_Object_Recolor (objects hue-shifted to yellow/cyan, grey background kept), and OOD_Background_Recolor (background recoloured, objects kept). The transformation is applied only to the Memory and Probe arrays; the retention frame stays clean grey.

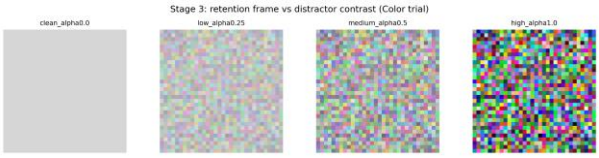


Fig. 3. Stage 3 example retention frames as distractor contrast increases from clean (left) to full (right)

3.3 EVALUATION AND METRICS

We report accuracy and the predicted "Changed" rate (the fraction of trials on which the model responds "Changed"). As the fixed trial sets are approximately balanced, a rate near 0.5 indicates no bias and a value near 1.0 means that a systematic bias to respond "Changed". Overall results are macro-averaged across the five variants. The same fixed delayed trials are reused across all conditions within each stage: 92 (Colour), 86 (Orientation), 90 (Size), 89 (Gap), and 88 (Conjunction), giving 445 trials per condition. All runs used a fixed seed (42), and per variant trial fingerprints were recorded for reproducibility. We macro-average across variants because the per variant trial counts differ slightly, therefore, a trial count weighted average gives effectively identical values. Chance accuracy is 0.5 on these approximately balanced sets, so an accuracy near 0.5 combined with a predicted "Changed" rate near 1.0 implies that the model has stopped discriminating and is producing a near constant response, whereas an accuracy near 0.5 with a balanced rate would indicate uniform guessing.

4 RESULTS

4.1 STAGE 1: RETENTION-INTERVAL INTERFERENCE

With a clean grey RI the model is perfect (accuracy of 1.000) and unbiased (predicted "Changed" rate 0.506). When we replace the retention frames with visual interference, the accuracy collapses toward chance for all four interference types while the predicted "Changed" rate rises sharply (Table 1): accuracy falls to 0.503 (random static), 0.508 (checkerboard), 0.501 (horizontal gratings), and 0.526 (vertical gratings), with predicted "Changed" rates of 0.93-0.97. The collapse is similar across very different interference structures, and because the Memory and Probe arrays are untouched, it suggests the possibility of the LSTM retention state being corrupted by the interfering frames during the delay.

Table 1. Stage 1: accuracy and predicted "Changed" rate for delayed trials (RI>0, 445 trials per condition)

Retention condition	Acc.	Pred. Changed rate
Clean grey	1.000	0.506
Random static	0.503	0.971
Checkerboard	0.508	0.971
Horizontal grating	0.501	0.960
Vertical grating	0.526	0.930

4.2 STAGE 2: OBJECT VS BACKGROUND PERTURBATIONS

When the retention interval is clean and only Memory and Probe arrays are perturbed, the model's robustness depends heavily on what is changed (Table 2). Recolouring the objects to novel hues is largely tolerated (with a macro accuracy of 0.936). The effect of the recolouring affects the Colour variant the heaviest (the accuracy drops to 0.793), which is the variant whose decision depends on the object colour, and is small for others (Gap 1.000, Orientation 0.977, Conjunction 0.966, Size 0.944). The predicted "Changed" rate under object recolour falls slightly below the baseline (0.441), and recolouring makes colour changes harder to detect, so the model misses some of them and therefore, leans toward "Same".

Recolouring only the background produces the opposite and a far more severe result. The macro accuracy collapses to 0.503 (essentially chance) with a near 100% "Changed" rate (0.998). This occurs even though the objects remain clearly visible and, on the "Same" trials, the transformed Memory and Probe arrays are identical. The failure is there not a label artefact. Instead, it is more likely that the model relies on the fixed grey background or the global image statistics that make up most of the 32×32 input. When the majority of the pixels pushes the encoding far out of distribution, the comparisons would then default to "Changed".

Figure 4 shows the per variant accuracies for both conditions.

The two Stage 2 conditions thus separate two different kinds of distribution shifts. By changing the task-relevant objects, which only changes a few pixels, perturbs the exact information the task is about and yet the model seems undisturbed. However, changing the task-irrelevant background, which occupies the majority of the image, seems to be catastrophic to the model's predictions. As such, one can infer that robustness is not determined by if a change is relevant but more of how much is changed.

Table 2. Stage 2 macro-averaged results across the five variants (RI>0, 445 trials per condition)

Condition	Acc.	Pred. Changed rate
Clean_Grey	1.000	0.505
OOD_Object_Recolor	0.936	0.441
OOD_Background_Recolor	0.503	0.998

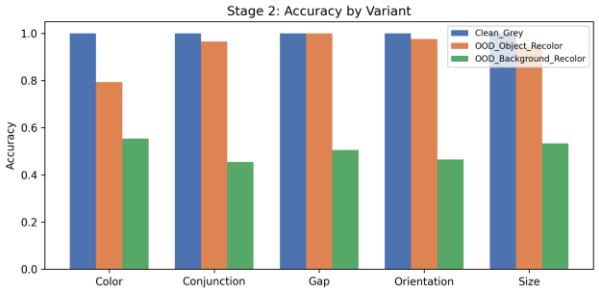


Fig. 4. Stage 2 accuracy by CD variant under the three conditions. Object recolouring is tolerated (lowest for Color); background recolouring collapses every variant to near chance.

4.3 STAGE 3: A CONTRAST DEPENDENT COLLAPSE

Stage 3 sweeps the contrast of the retention distractor (Table 3, Figure 5). At zero and low contrast the model is unaffected (accuracy 1.000 at contrast of 0.0 and 0.1, with a balanced "Changed" rate of 0.505). When contrast is at 0.25, accuracy then falls steeply to 0.693. This trend continues such that when contrast is 0.5, accuracy reaches chance level (0.505) and remains at near chance as contrast increases (0.516 at 0.75, 0.503 at 1.0), while the predicted "Changed" rate rises to 0.99 by mid contrast. The full contrast endpoint reproduces the Stage 1 random static collapse, this then provides an internal consistency check. This relationship indicates that the retention interval failure is driven by the contrast and hence the visual saliency of the distractor rather than just the presence of a structured pattern during the delay.

The collapse is consistent across variants (Appendix, Table A1): every variant reaches near chance (or chance) by mid contrast, and the variant level curves track the macro average closely, so the dose response is a property of the model rather than of any single feature dimension.

Table 3. Stage 3 macro-averaged accuracy and predicted Changed rate as a function of distractor contrast (RI>0)

Distractor contrast	Acc.	Pred. Changed rate
0.00	1.000	0.505
0.10	1.000	0.505
0.25	0.693	0.619
0.50	0.505	0.995
0.75	0.516	0.907
1.00	0.503	0.970

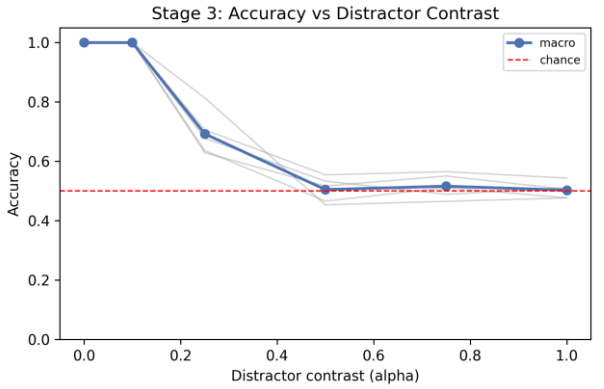


Fig. 5. Stage 3 macro accuracy versus distractor contrast (faint lines: individual variants). Accuracy is preserved at low contrast and falls to chance by mid contrast.

5 DISCUSSION

Across all three stages, the same LSTM-96 that performed consistently well on clean delayed trials can be driven to chance by manipulations that a human observer would handle easily (a noisy delay, or a recoloured background). This then supports the paper's claim: clean benchmark accuracy is an incomplete measure of artificial WM, and a better diagnostic value can be found in the structure of the failures.

Stage 1 shows that the recurrent memory state is highly vulnerable to irrelevant interference during maintenance, and Stage 3 sharpens this into a dose-response: the failure is graded and contrast

dependent, appearing only once the distractor is sufficiently obvious. Because the low contrast distractor with the same spatial structure leaves performance intact, the collapse is best explained by strong, task-irrelevant visuals that overwrites the maintained representation, rather than by the presence of the structure. This is where the model differentiates itself from human-like WM as human maintenance would be comparatively resistant to such interference.

Stage 2 reveals an even more striking dissociation. The model tolerates a significant amount of change to task-relevant feature (novel object colours) but collapses when the task-irrelevant background changes, even though the objects stay visible and the arrays remain identical. The two conditions also produce opposite biases such that the object recolouring makes the model more likely to respond "Same" (it misses some changes), whereas background recolouring makes it respond "Changed" almost always. This pattern is consistent with an encoder whose representation is dominated by background level statistics rather than the small foreground objects. This behaviour aligns with previous research showing that neural networks frequently exploit non-transferable low-level visual cues [7],[8]. A memory system that depends heavily on a fixed background context is nowhere near as robust as the way human memory is.

These failures also answer the efficiency side of the goal. As the model attains its precision by training on stable regularities (constant grey background and a fixed contrast range), it is therefore extremely brittle when those regularities change. Developing memory systems that are both efficient and human-like therefore require representations organised around objects and their relations, as in human binding accounts [11], [13], or architectures that separate maintenance from perception more explicitly, such as memory augmented or attention based designs [18],[19]. Our diagnostics do not establish which method is best, but they identify the specific weaknesses any new method implemented must address.

It is worth situating these results against the benchmark from which the model is taken. WorM reports that recurrent models produce several qualitative human signatures on clean tasks [3]. Our findings do not contradict this, but they show that agreement on clean inputs does not extend to the robustness properties examined here. A more human-like model would be expected to degrade gracefully under interference rather than collapsing abruptly, and to remain near invariant to a change of background rather than shifting its decision entirely. The distance between these expectations and the observed behaviour is precisely what a diagnostic evaluation (as opposed to a single clean accuracy score) is designed to reveal.

Taken together, a recurring bias toward responding "Changed" emerges whenever the input is pushed far enough out of distribution (by delay interference in Stage 1 and 3 or by a background shift in Stage 2). Yet the object recolour result shows this is not a single universal failure as the direction of the bias depends on the manipulation. Reporting decision bias alongside accuracy, and localising perturbations to specific phases of the trial, are therefore simple but informative diagnostics. These observations connect directly to the project's central aim. Developing efficient human-like memory systems requires the model to not only store information but also to be able to generalise the materials when reasonable perceptual changes occur. Identifying precisely when the model fails at these requirements (as the three stages do here) is a prerequisite for building architectures and training schemes that satisfy them.

6 LIMITATIONS

Several limitations exist in this project. We evaluated a single model (LSTM-96) zero-shot, so the results can only be applied to this specific model and should not be generalized across all other artificial working memory. The delayed test set is modest (with 445 trials per condition and 86-92 per variant), this is sufficient for large effects reported but is unable to identify small differences. The object recolouring is a 60° hue rotation that tries to shift colours towards yellow and cyan but not exactly hitting them. Yellow is a novel colour to the whole model, whereas blue appears in other WorM tasks, therefore, colour novelty is only partial. The background change is a large global shift on a small 32×32 image and should be interpreted as such. Finally, we analyse behavioural outputs only (not internal CNN or LSTM representations), so the mechanistic interpretations above remain hypotheses to be tested.

7 CONCLUSION

Using a three stage, zero-shot diagnostic examination, we found that a pretrained LSTM-96 with perfect clean accuracy collapses to chance under retention interval interference, tolerates novel object colour changes but fails catastrophically when only the background colour changes, and exhibits a graded, contrast dependent collapse under delay interference. This failure was later investigated and was found to be systematic. This paper suggests that clean benchmark accuracy alone cannot certify an artificial working memory model as robust. By pinpointing why this model fails, the study contributes a small but concrete step towards developing efficient human-like memory systems. For future studies, a suggestion would be to test on multiple backbones, using larger balanced out-of-distribution test sets, and analysing

internal representations to confirm the mechanisms suggested here.

ACKNOWLEDGMENT

I would like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project. I thank my supervisor Asst Prof Zhang Mengmi for guidance, and the authors of the Working Memory (WorM) benchmark for releasing the dataset, tasks, and the pretrained models on which this work builds.

REFERENCES

[1] A. D. Baddeley and G. J. Hitch, “Working memory,” in The Psychology of Learning and Motivation, vol. 8, G. H. Bower, Ed. New York: Academic Press, 1974, pp. 47–89.

[2] W. J. Ma, M. Husain, and P. M. Bays, “Changing concepts of working memory,” Nature Neuroscience, vol. 17, no. 3, pp. 347–356, 2014.

[3] A. Sikarwar and M. Zhang, “Decoding the Enigma: Benchmarking Humans and AIs on the Many Facets of Working Memory,” in Proc. 37th Conf. on Neural Information Processing Systems (NeurIPS 2023), Datasets and Benchmarks Track, 2023. arXiv:2307.10768.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in Proc. Int. Conf. on Learning Representations (ICLR), 2014. arXiv:1312.6199.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proc. Int. Conf. on Learning Representations (ICLR), 2015. arXiv:1412.6572.

[6] D. Hendrycks and T. Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations,” in Proc. Int. Conf. on Learning Representations (ICLR), 2019.

[7] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,” in Proc. Int. Conf. on Learning Representations (ICLR), 2019.

[8] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, “Shortcut learning in deep neural networks,” Nature Machine Intelligence, vol. 2, no. 11, pp. 665–673, 2020.

[9] A. Baddeley, “The episodic buffer: a new component of working memory?,” Trends in Cognitive Sciences, vol. 4, no. 11, pp. 417–423, 2000.

[10] N. Cowan, “The magical number 4 in short-term memory: A reconsideration of mental storage capacity,” Behavioral and Brain Sciences, vol. 24, no. 1, pp. 87–114, 2001.

[11] S. J. Luck and E. K. Vogel, “The capacity of visual working memory for features and conjunctions,” Nature, vol. 390, no. 6657, pp. 279–281, 1997.

[12] E. K. Vogel and M. G. Machizawa, “Neural activity predicts individual differences in visual working memory capacity,” Nature, vol. 428, no. 6984, pp. 748–751, 2004.

[13] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” Cognitive Psychology, vol. 12, no. 1, pp. 97–136, 1980.

[14] K. Oberauer, S. Lewandowsky, E. Awh, G. D. A. Brown, A. Conway, N. Cowan, C. Donkin, S. Farrell, G. J. Hitch, M. J. Hurlstone, W. J. Ma, C. C. Morey, D. E. Nee, J. Schweppe, E. Vergauwe, and G. Ward, “Benchmarks for models of short-term and working memory,” Psychological Bulletin, vol. 144, no. 9, pp. 885–958, 2018. doi: 10.1037/bul0000153.

[15] J. L. Elman, “Finding structure in time,” Cognitive Science, vol. 14, no. 2, pp. 179–211, 1990.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in Proc. EMNLP, 2014, pp. 1724–1734.

[18] A. Graves, G. Wayne, and I. Danihelka, “Neural Turing Machines,” arXiv:1410.5401, 2014.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 2017.

[20] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[21] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. on Learning Representations (ICLR), 2015. arXiv:1412.6980.

APPENDIX

A. PER VARIANT STAGE 2 RESULTS

Table A1 reports Stage 2 accuracy and predicted Changed rate for every variant under the object- and background-recolour conditions (RI > 0).

Table A1. Stage 2 per-variant accuracy (Acc) and predicted Changed rate (PC).

Variant	Obj Acc	Obj PC	Bg Acc	Bg PC
Color	0.793	0.348	0.554	1.000
Orientation	0.977	0.442	0.465	1.000
Size	0.944	0.478	0.533	1.000
Gap	1.000	0.506	0.506	1.000
Conjunction	0.966	0.432	0.455	0.989

B. REPRODUCIBILITY

All evaluations used a fixed global seed (42) and the same fixed delayed trials within each stage (Color 92, Orientation 86, Size 90, Gap 89, Conjunction 88; 445 per condition), with per-variant stimulus, response, sequence-length, retention-interval, and set-size checksums recorded. The CD background grey value was measured from the data as 0.839 (not assumed). A label-preservation check confirmed, for every Stage 2 condition, that transformed Memory and Probe frames were identical on Same trials, differed on Changed trials, and left the retention frames unchanged. Object recolouring used a 60° luminance-preserving hue rotation; background recolouring set achromatic (background) pixels to a steel-blue (RGB ≈ 0.45, 0.55, 0.85).