Relational composition during attribute retrieval in GPT is not purely linear

Michael B. McCoy
Department of Cognitive Sciences
University of California, Irvine
Irvine, CA 92617
michael.b.mccoy@gmail.com

Anna Leshinskaya
Department of Cognitive Sciences
University of California, Irvine
Irvine, CA 92617
anna.leshinskaya@gmail.com

Abstract

The longstanding question of how neural networks could implement relational composition has been buoyed by recent success showing relational abstraction in transformer-based large language models (LLMs). We address recent findings showing some, but imperfect, generalizability in linear composition during knowledge retrieval of attributive triplets such as has-color (banana, yellow) in GPT-J [Hernandez, E. et al, (2024). Linearity of relation decoding in transformer language models (arXiv:2308.09124)]. We report that limitations to relational generalization are explainable by two systematic factors. First, relational combinations that are more accurately retrieved generalize better than uncertain or inaccurate ones. Second, relational generalization scales with the semantic similarity of the entities being bound, showing that it is in fact non-linearly dependent on component meanings rather than being purely invariant. This aligns with longstanding findings that human judgments of adjectival combinations are likewise non-linearly interactive.

1 Introduction

Relational composition is a fundamental capacity of higher-level intelligence, allowing the infinite use of finite means in language and thought. It entails *productivity*, the generation and understanding of novel combinations of simpler parts, and *systematicity*, the ability to meaningfully recognize relations among combinations (Fodor, 1998). For example, systematicity allows us to recognize that *John loves Mary* is analogous to *Amy loves James*, and generalize the same relation to new entities, such as *Gina loves Lisa*. While it remains a longstanding challenge to implement truly compositional operations in neural networks (Dasgupta et al., 2018; Elman, 1991; Fodor & Pylyshyn, 1988; Holyoak & Hummel, 1997; Hummel & Holyoak, 2003; Kriete et al., 2013; Marcus, 2001; Mitchell, 2021; Smolensky, 1990; Socher et al., 2012; Webb et al., 2024), recent transformer-based large language models (LLMs) trained on large corpora of language have demonstrated remarkable emergent relational capabilities, including analogy and function abstraction (Feng & Steinhardt, 2024; Hernandez et al., 2024; Lepori et al., 2023; R. T. McCoy, 2022; Webb et al., 2023). This offers an unprecedented opportunity to empirically discover how neural networks can solve this problem. Yet, debate continues as to whether these networks implement compositionality or some approximation of it (McCoy, 2022; Mitchell et al., 2023). We focus on recent findings suggesting some but imperfect linear composition during knowledge retrieval of attributive relations among concepts in open-source LLM GPT-J (Hernandez et al., 2024) and test several ideas about what might explain this imperfection.

Knowledge retrieval tasks probe a model's stored representation of typical relations among familiar concepts, here investigated as triplets including two concepts bound by an attributive relation. For example, *banana* is expected to be bound to *yellow* with the relation *has-color*. This structure follows the format of classic semantic networks (Collins & Quillian, 1969; McClelland & Rogers,

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

2003). Related work on conceptual combination has long investigated how the meaning of word combinations (such as these) arises out of the meaning of their individual parts (Pustejovsky, 1995; Smith et al., 1988; Smith & Medin, 1981). Whatever the nature of these processes, they are expected to possess some degree of systematicity and productivity so as to allow the limitless and meaningful combinations expressible in natural language and comprehensible to the human mind.

Hernandez et al (2024) tested whether such conceptual combinations in GPT-J are linearly compositional and thus, systematic. A given relational triplet, (s, r, o) is hypothesized to be the product of the model-internal hidden representation of its subject s (e.g., *banana*), some set of operations representing the application of relation r (eg., *has-color*), and a downstream, retrieved hidden representation of its attribute object o (e.g., *yellow*). The combinatorial operation is estimated as a linear function specific to that relation, W_r, that transforms the earlier representation of s into a representation of the correct object o at the output layer. This relational operation is systematic to the degree that it can be re-bound to new concepts while preserving its meaning, such that applying the same W_r to a new item, s' (e.g., *lime*), should yield the correct new tuple, (s', r, o'), in which o' represents *green*. This would show both systematicity and generalization of W_r. Hernandez et al (2024) report that such cross-application has an overall accuracy of 60% across a number of content domains and, on this basis, argue that relational operations in GPT-J attribute knowledge are indeed linearly compositional.

However, it is unclear what exactly an accuracy of 60% implies about this hypothesis and what factors might explain the remaining inaccuracy. Although this accuracy is above certain notions of chance, what accounts for the remaining variance? One possibility is random noise; another is model knowledge uncertainty; and a third is systematic non-linearity. The objective of the present work is to test these alternatives by quantifying the generalization of W_r on an item-by-item basis and comparing predictors of generalization success across items.

Model knowledge accuracy (vs uncertainty) is a likely factor because the retrieval of o was below ceiling accuracy at baseline for many triplets in Hernandez et al.'s report. Uncertainty in the retrieval of an attribute would contribute to noise in estimating W_r and limit the extent of possible generalization success at test. We thus expect that relational triplets with higher accuracy will generalize best.

Of greater theoretical interest, we hypothesized that the extent to which the relational operation W_r generalizes might be systematically non-linear as a function of the semantic similarity of the bound entities, s and s', or o and o', across pairs. This predicts that the relational operation W_r will be more similar among two fruit (e.g., banana and apple) than among more different items (e.g., banana and barn). It may also be more similar for two entities that are both the same color, such as banana and canary. Broadly, this predicts that relational generalization scales smoothly as a function of the similarity between the bound concepts, rather than being a purely invariant operation. If so, then the relational function is not linear, but rather, systematically non-linear and interactively dependent on its operands. This would not be without precedent in empirical work on conceptual combination, which has long observed that human conceptual combination is non-linearly interactive, as opposed to being tractably, linearly compositional (Fodor, 1998; Medin & Shoben, 1988; Pustejovsky, 1995). We test this prediction here as an account of the variability in generalization of relational functions in GPT-J, closely following the methodology in Hernandez et al. (2024).

2 Methods

2.1 Stimuli

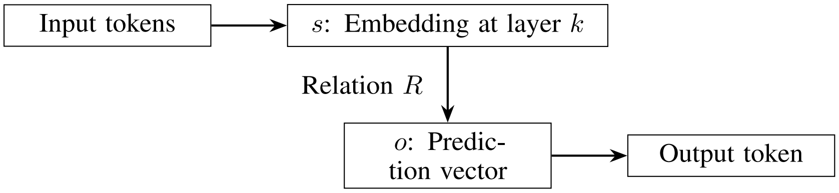
Stimuli were 44 triplets consisting of a subject (s), relation (r) and attribute object (o); Supp. Table 1. We focused on the attributive relation has-color, building on materials in Hernandez et al. We sourced items with reference to human production norms, selecting items for which no other color attribute was produced (Cree & McRae, 2003) to increase the chance of model certainty. For similar reasons, we used common objects (vs unique entities), as these have better representation in language corpora. Color attributes also apply to a wide range of entities, permitting a broad test of generalization, which we performed across the domains of fruit, vegetables, animals, and inanimate objects.

2.2 Model knowledge accuracy

Throughout, in line with the main results in Hernandez et al 2024, we used GPT-J (Wang & Komatsuzaki, 2021), an open-source transformer model. We quantified baseline relational retrieval accuracy for each triplet by prompting the model with s and r and evaluating the probability of returning the expected o, as per Supplemental Table 1. We used 4 synonymous prompts as follows, where s was a given subject: 1) "What color is s on the outside? It is usually"; 2) "The usual color of s is which color? It is", 3) "The outside color of s is typically", 4) "On the outside, s most often has the color". We scored responses by querying whether the expected answer appeared within the first 5 returned tokens. Accuracy was summed across the 4 prompt trials. When using accuracy as a predictor for the amount generalization between a pair of triplets, we averaged the accuracy of the two triplets.

2.3 Estimation of W, s, and o

A large language model is a machine that maps a sequence of input tokens or *prompts* into a single output token. Internally, large language models embed prompts into vectors, then predict next-token probabilities, and the output of the machine is generally the most likely predicted token.



Thus, the relation is represented by the functional form $R(s) = o$ induced by the specific neural network of interest. In our work, we use the activations at layer 5 of GPT-J as the embedded vector s, and the unnormalized output of the final layer as the vector representation of o.

Following (Hernandez et al., 2024), we use calculus to motivate a linear approximation of the relation operator R. For $s' \approx s$, a first-order Taylor series approximation shows that the relationship R can be well-approximated by an affine function via the Jacobian matrix $W := J_s R$:

$$R(s') \approx R(s) + W(s' - s) = Ws' + b$$

where the *bias* $b := R(s) - Ws$. This approach provides a theoretical justification for the use of a linear relation function, yet we note that the Jacobian $W := J_s R$ is a function of the embedded s.

The extent to which W depends on the embedding is a core investigation of this work. This leads to an important methodological change: unlike Hernandez et al.(2024), we do not average the resulting W matrix over multiple prompts, but instead consider a W for each prompt. We measure distances of W between all pairs of prompts, yeilding a matrix W_{dist}, that quantifies pairwise generalization. Following Hernandez et al., we also measure*faithfulness* as the accuracy of retrieving object o^j from s^j when replacing W^j with W^i. Details of these measures are provided in the Supplement.

3 Results

We estimated the linear function W_r^i for each triplet (s^i, r^i, o^i) and the distance between all pairs W_r^i and W_r^j as a measure of how well the relation function generalizes between them. The resulting distance matrix W_{dist} is shown in Figure 1. It is clear that W_{dist} is not uniform or at ceiling and thus generalization is not perfect. We next compared several potential predictors of W_{dist}.

A least-squares regression showed that the W_{dist} of a pair was significantly predicted by pair accuracy, such that pairs with higher-accuracy triplets had smaller distances in W: t(818) = -4.74, p < .001. Nonetheless, this factor explained only a small amount of total variance in W_{dist}, R^2 = 0.027.

Our core hypothesis is that the generalization of the relational operation (and thus W_{dist}) varies smoothly as a function of the semantic similarity of the entities bound by it, here s and o. We used the hidden-state representation at layer 5 of each items s (in the relational context) to measure s_{dist}, and the last-layer representation of o (given prior context s and r) to measure o_{dist}, for each pair of items. Because r is constant across all items, variation in these distance measures reflect the variation derived from s and o.

In a linear regression with predictors o_{dist}, s_{dist}, and pair accuracy, we found that W_{dist} was significantly predicted by each factor: s_{dist}, t(816) = 3.57, p < .001, o_{dist}, t(816) = 24.74, p < .001,

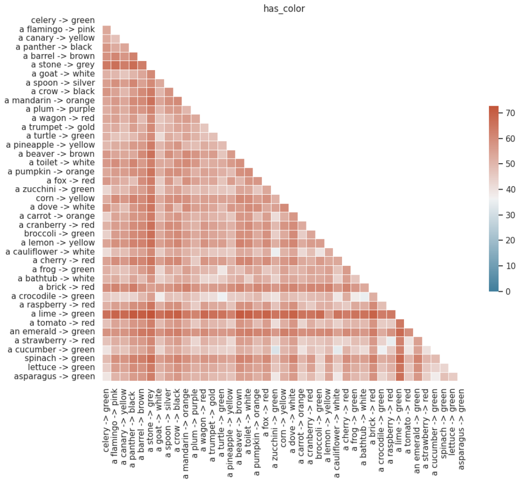


Figure 1: W_{dist}, the distance between the estimated linear relation function, of each pair of triplets under the relation 'has-color'.

and accuracy, t(816) = -3.01, p < .01. Larger distances in s and o predict larger W_{dist}, while higher accuracy predicts lower W_{dist}. Altogether, these three factors accounted for a large portion, indeed about half, of all variance in relational generalization, R^2 = 0.53.

As a convergent test of generalization in practical terms, we followed Hernandez et al. (2024) in applying the function W_r across distinct triplets to estimate the accuracy of retrieving the output o, in terms of its token representation. First, a logistic regression confirmed that W_{dist} significantly predicted faithfulness, t(818) = -5.70, p < .001, with a negative coefficient suggesting that smaller distances in W_{dist} were associated with higher faithfulness accuracy, consistent with our assumption that smaller relational distances between entities s^i and s^j are associated with higher retrieval accuracy of o^j. Accordingly, faithfulness behaved similarly to W_{dist} in our predictive model: it was significantly predicted in a logistic regression by model accuracy, z(816) = 10.66, and p < .001, by s, z(816)= -2.92, p < .01, and by o, s(816) = -6.40, p < .001, altogether accounting for R^2 = 0.208.

4 Discussion

Building on prior work showing linearly compositional representations of triplet attributive relations in GPT-J (Hernandez et al 2014), we found that imperfections in linear generalization across triplets were explainable by at least two systematic factors: model accuracy in retrieving the individual triplets and the semantic similarity of the entities bound by the relation across triplets. This demonstrates that this sort of relational binding has systematic non-linearities: it depends on the semantics of the entities being bound, rather than being purely invariant.

This result aligns with longstanding findings that human judgments of adjectival combinations are likewise non-linearly interactive (Medin & Shoben, 1988). This suggests that we may not expect pure linear composition in conceptual combination, whether in humans or models. Ultimately, studying how LLMs achieve combinatorial relational binding is a promising avenue for illuminating long-standing puzzles regarding the neural network implementation of combinatorial thought, which can in turn shed light on how this is achieved in the human brain.

Limitations The scope of our findings is limited to the particular form of relational combination investigated: knowledge retrieval of attributive relations, as probed with a particular methodology. Results may vary across different approaches for testing relational operations. We also expect that relational composition might be more linear in other task domains, such as in-context binding (Feng & Stainhardt 2024), or for stimuli with more linearly separable features.

Safety implications We believe that our research can have a positive impact on safety. Understanding the limitations of relational generalization can promote responsible use of AI models.

5 References

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2), 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2), 163–201. https://doi.org/10.1037/0096-3445.132.2.163

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., & Goodman, N. D. (2018). Evaluating compositionality in sentence embeddings (arXiv:1802.04302). *arXiv*. http://arxiv.org/abs/1802.04302

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2), 195–225. https://doi.org/10.1023/A:1022699029236

Feng, J., & Steinhardt, J. (2024). How do language models bind entities in context? (arXiv:2310.17191). *arXiv*. http://arxiv.org/abs/2310.17191

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Clarendon Press; Oxford University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3–71.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., & Bau, D. (2024). Linearity of relation decoding in transformer language models (arXiv:2308.09124). *arXiv*. http://arxiv.org/abs/2308.09124

Holyoak, K. J., & Hummel, J. E. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3), 427–466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110(2), 220–264. https://doi.org/10.1037/0033-295X.110.2.220

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences of the United States of America*, 110(41), 16390–16395. https://doi.org/10.1073/pnas.1303547110

Lepori, M. A., Serre, T., & Pavlick, E. (2023). Break it down: Evidence for structural compositionality in neural networks (arXiv:2301.10884). *arXiv*. http://arxiv.org/abs/2301.10884

Marcus, G. F. (2001). *The Algebraic Mind*. MIT Press.

McClelland, J. L. & Rogers. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews*. Neuroscience, 4(4), 310–322. https://doi.org/10.1038/nrn1076

McCoy, R. T. (2022). *Implicit compositional structure in the vector representations of artificial neural networks* [Doctoral dissertation]. Johns Hopkins University.

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, 20(2), 158–190. https://doi.org/10.1016/0010-0285(88)90018-7

Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *arXiv*, 2102.10717, 1–29.

Mitchell, M., Palmarini, A. B., & Moskvichev, A. (2023). Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks (arXiv:2311.09247). *arXiv*. http://arxiv.org/abs/2311.09247

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Smith, E. E., & Medin, D. L. (1981). *Categories and Concepts*. Harvard University Press.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science*, 12(4), 485–527. https://doi.org/10.1207/s15516709cog1204_1

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1–2), 159–216. https://doi.org/10.1016/0004-3702(90)90007-M

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1201–1211.

Wang, B., & Komatsuzaki, A. (2021, May). GPT-J-6b: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax

Webb, T., Frankland, S. M., Altabaa, A., Segert, S., Krishnamurthy, K., Campbell, D., Russin, J., Giallanza, T., Dulberg, Z., O'Reilly, R., Lafferty, J., & Cohen, J. D. (2024). The relational bottleneck as an inductive bias for efficient abstraction (arXiv:2309.06629). *arXiv*. http://arxiv.org/abs/2309.06629

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

6 Supplemental material

Table 1: Stimuli used in the experiments: subjects and objects bound by the has-color relation. Accuracy denotes percent correct during retrieval in GPT-J. Production frequency is reported from human data from McRae Norms (Cree & McRae 2003).

Subject (s)	Attribute object (o)	Domain	Model Accuracy	Production frequency (out of 30)
a barn	red	object	0	12
a barrel	brown	object	0.25	8
a bathtub	white	object	1	9
a brick	red	object	0.25	14
an emerald	green	object	1	26
a spoon	silver	object	0.25	10
a stone	grey	object	0	13
a toilet	white	object	1	11
a trumpet	gold	object	0	5
a tuba	gold	object	0	5
a wagon	red	object	0	17
a cauliflower	white	vegetable	0.5	26
a carrot	orange	vegetable	0.75	27
celery	green	vegetable	1	27
corn	yellow	vegetable	0.5	23
a cucumber	green	vegetable	1	26
lettuce	green	vegetable	1	28
a pumpkin	orange	vegetable	0.75	27
spinach	green	vegetable	1	30
a zucchini	green	vegetable	1	21
asparagus	green	vegetable	1	25
broccoli	green	vegetable	1	29
a cherry	red	fruit	1	27
a cranberry	red	fruit	1	24
an eggplant	purple	fruit	0.75	22
a lemon	yellow	fruit	1	28
a lime	green	fruit	0.75	29
a mandarin	orange	fruit	0.25	22
a pineapple	yellow	fruit	0.25	15
a plum	purple	fruit	0.25	23
a raspberry	red	fruit	1	19
a strawberry	red	fruit	1	25
a tomato	red	fruit	0.75	28
a turtle	green	reptile	0	14
a crocodile	green	reptile	0.75	10
a frog	green	amphibian	0.75	26
a dove	white	bird	0.75	25
a canary	yellow	bird	0.75	28
a crow	black	bird	0.75	25
a flamingo	pink	bird	1	23
a beaver	brown	mammal	0.75	19
a goat	white	mammal	0.75	6
a fox	red	mammal	0	15

6.1 Additional methodological details

6.1.1 Code availability

The code is provided under https://github.com/relcoglab/hernandez-relations-fork and was run on Google Colab with an A100 GPU. Results can be reproduced with a minimum of compute time (under one hour). Our repository is a fork of material from the original project released by Hernandez et al (2024) at https://github.com/evandez/relations/ under an MIT licence allowing re-use.

6.1.2 Distance computation

For each pair of relations in each class (s^i, r^i, o^i) and (s^j, r^j, o^j) , we compute the distance between each s embeddings, the W matrices and the o outputs, as the usual Euclidean norm:

$$\begin{aligned} \mathrm{dist}(s^i, s^j)^2 \quad &= \textstyle\sum_k |s^i_k - s^j_k|^2, \\ \mathrm{dist}(o^i, o^j)^2 \quad &= \textstyle\sum_k |o^i_k - o^j_k|^2, \text{ and} \\ \mathrm{dist}(W^i, W^j)^2 \quad &= \textstyle\sum_{k,\ell} |W^i_{k\ell} - W^j_{k\ell}|^2. \end{aligned}$$

In the case of matrices W, this corresponds to the Frobenius norm.

6.1.3 In-context learning

In order to generate the vector embedding s, we create an in-context learning prompt with three example relations. For example, we embed the (s,r,o) relation '(celery, color, green)' via the prompt

```
On the outside, the color of an eggplant is purple
On the outside, the color of a barn is red
On the outside, the color of a tuba is gold
On the outside, the color of celery is
```

We use three exemplars in our prompt as a form of in-context learning (ICL), which increases the likelihood that the LLM will generate an appropriate output. The three ICL prompts were chosen at random and fixed for the experiment, and were not otherwise included in the analysis.

6.1.4 Faithfulness

We use the notion of *faithfulness* defined in (Hernandez et al., 2024) to determine how well the operator W^i for relation for subject s^i generalizes to subject s^j. In particular, we say a relation is *faithful* if the output vector $o^{i,j} := W^i(s^j)$ correctly predicts the target o^j. To allow for some ambiguity in the correct answer, we say it "correctly predicts" when the expected answer is in the top three predicted tokens. When using W_{dist} as the predictor for faithfulness between s^i and s^j, we only test $i < j$ to avoid duplicate W_{dist}).

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We clearly define what we test (two predictions about relational generalization), which bear out with statistical analysis. We note that our results are limited to the kind of composition tested here and may not generalize to every kind of relational composition.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We provide a limitations section.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

Answer: [NA]

Justification: Our findings are not theoretical results.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: We provide a code repository and ample description as well as references to past work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: We have provided code that we believe is easy to reproduce in a Colab notebook environment.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: We described how test data are separated. There is no model training in our work.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: While we do not plot any error bars, we report details information about the results of statistical tests.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: We explain that the analyses can be run easily in Google Colan using an L4 GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the code of ethics and have provided a statement on safety implications. Human data was not used in this research.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: There was limited space to comment on societal impacts but we mention implications for longstanding questions in cognitive science and safety.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release models or datasets that have safety implications as such.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit the prior work whose code we build on throughout the manuscript and refer to it in detail in section 6.1.1.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We have not released new assets at this moment. We provide a notebook with the code to generate the results of this paper for reviewers.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not use human subjects data.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not use human subjects data.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.