

CLEVR-CONCEPTS

HIERARCHICAL COMPOSITIONAL GENERALISATION BENCHMARKING

Adam Dahlgren Lindström



UMEÅ UNIVERSITY

You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to get it into the living room, you will have to

Gary Marcus & Ernest Davis, 2020, GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about

*You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to get it into the living room, you will have to **remove the door. You have a table saw, so you cut the door in half and remove the top half.***

Gary Marcus & Ernest Davis, 2020, GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about



*“Teddy bears shopping for groceries in the style of ukiyo-e”
by DALL-E 2.*



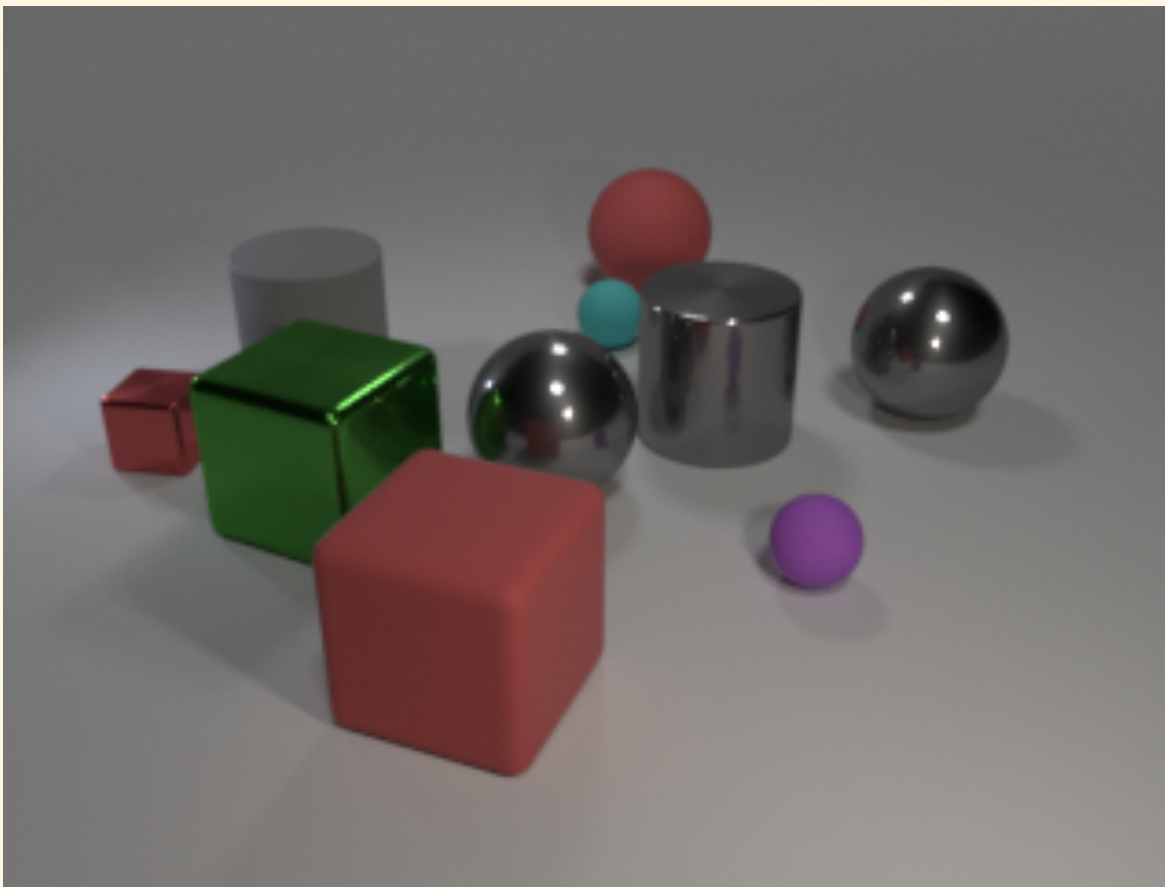
(a) some plants
surrounding a lightbulb



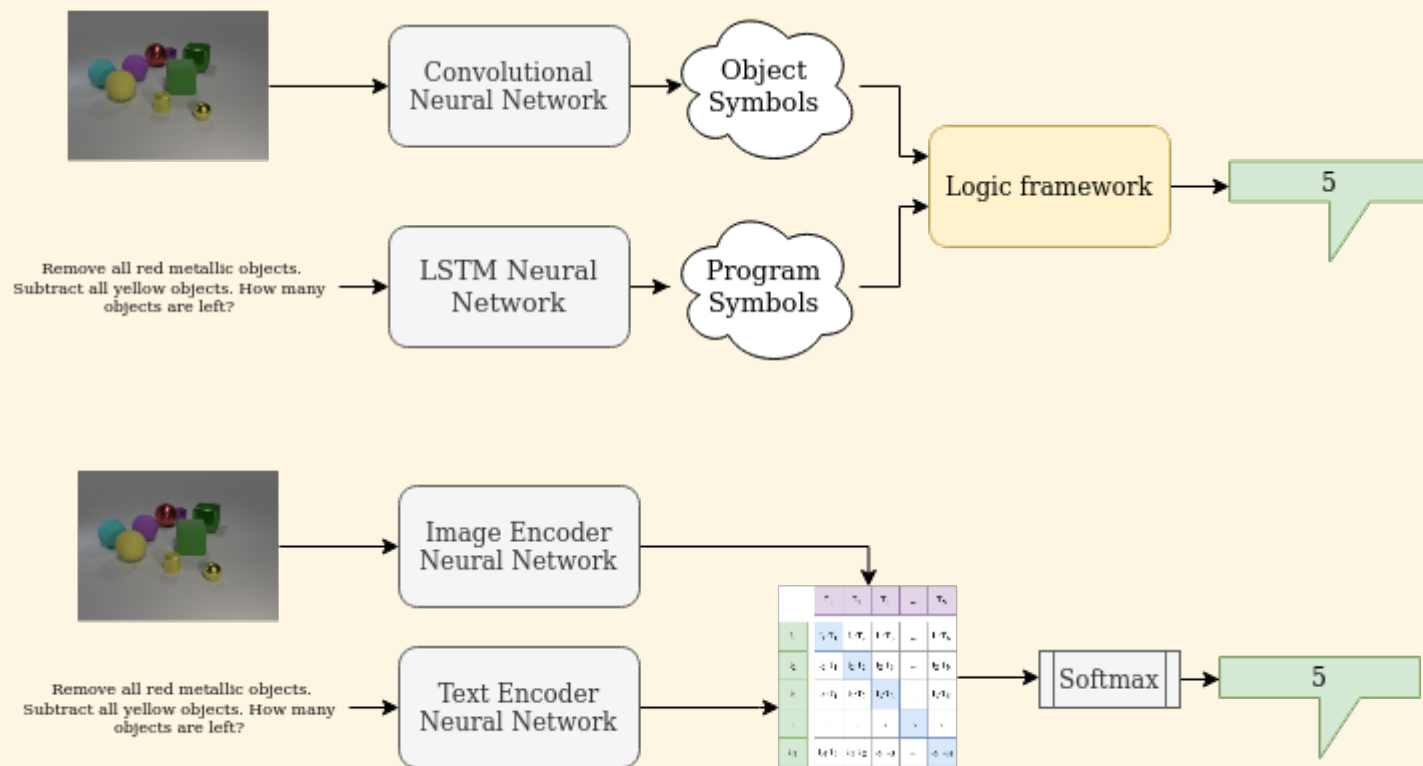
(b) a lightbulb surrounding
some plants

Thrush, Tristan, et al. "Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality." Proceedings of CVPR. 2022.

EXISTING MODELS TO **INFORM BETTER** **DESIGNS**



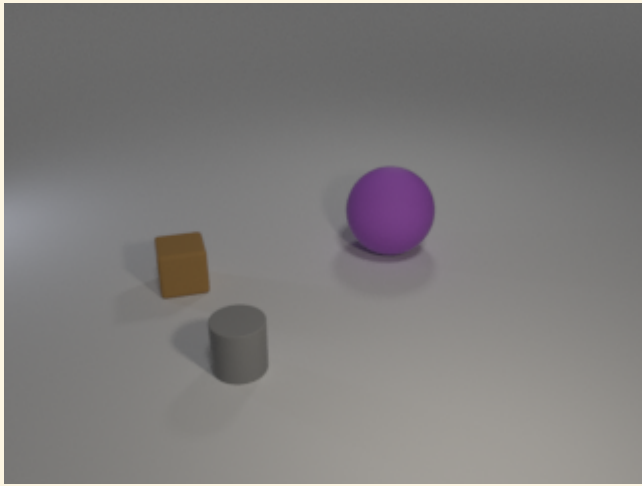
(i) Remove all gray spheres. How many spheres are there? (3), (ii) Take away 3 cubes. How many objects are there? (7), (iii) How many blocks must be removed to get 1 block? (2)



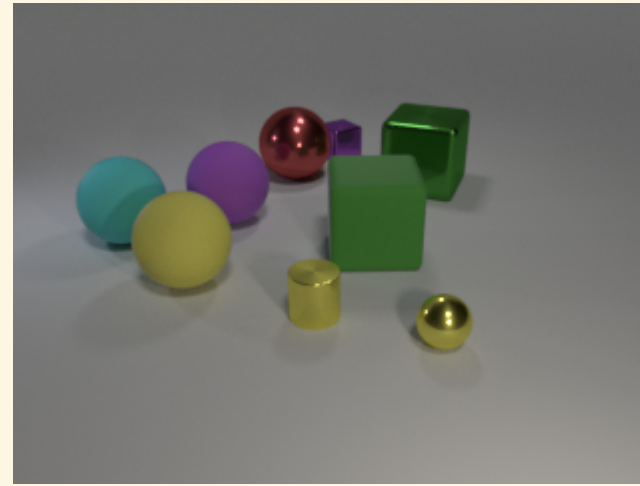
*Yi, K., ... & Tenenbaum, J. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. NeurIPS, 31.

**Radford, A., et al. (2021, July). Learning transferable visual models from natural language supervision. In ICML (pp. 8748-8763). PMLR.

Model	All	Add.	Sub.	Adversarial	Multihop	Multihop (0-shot)
NS-VQA	0.8840	0.9781	0.9948	0.9957	0.286	0.267
CLIP	0.3464	0.5699	0.3019	0.2848	0.272	0.238



Subtract all small purple matte blocks. Subtract all blocks. How many objects are left? was answered by CLIP with 3 instead of 2.



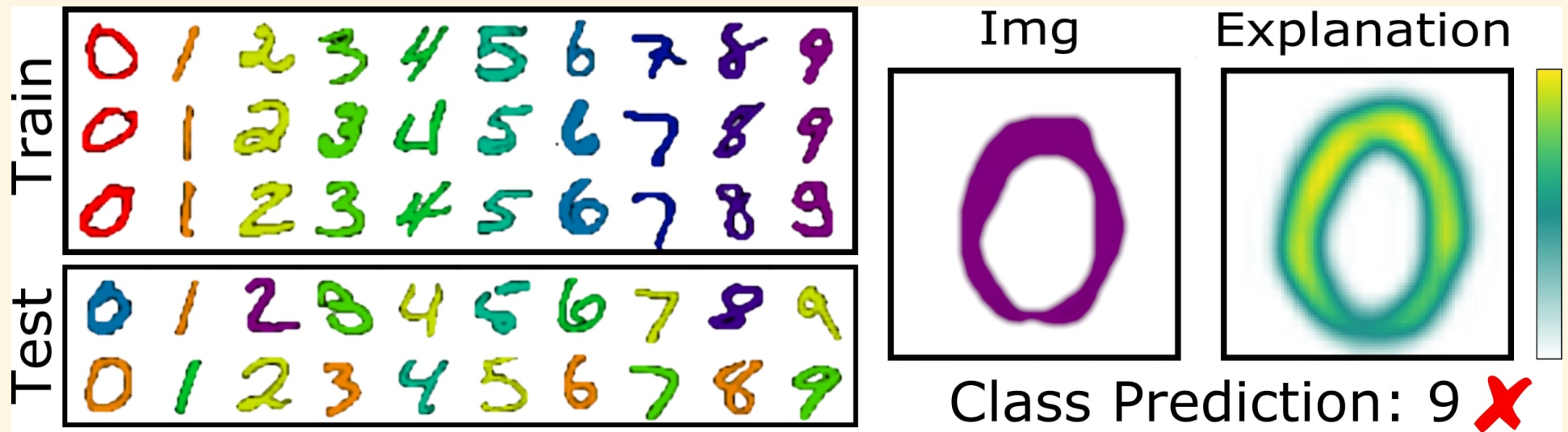
Subtract all orange metallic objects. Subtract all yellow objects. How many objects are left? was answered with 9 instead of 5 by NS-VQA.

Some level of **compositionality** in these models.

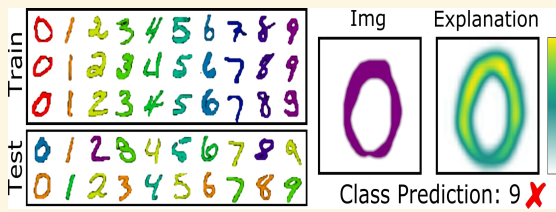
Models learn **concepts**, but not necessarily **structure**.

Difficult to disentangle problems with **compositionality and reasoning**, especially in multimodal settings.

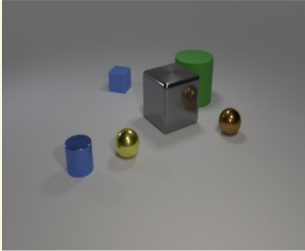
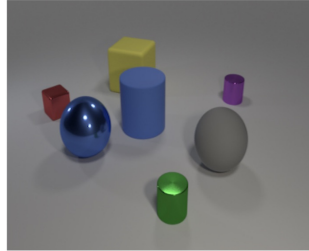
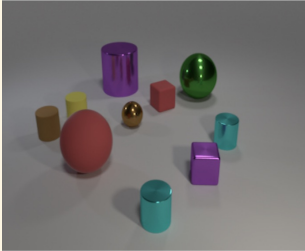
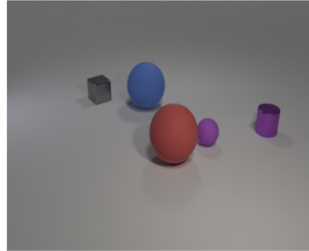
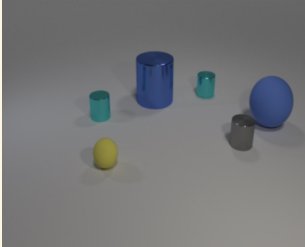
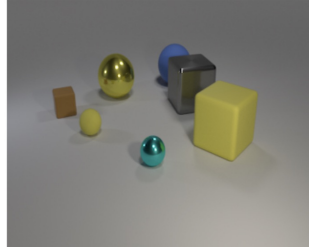
**INVESTIGATE
COMPOSITIONALITY
FURTHER?**

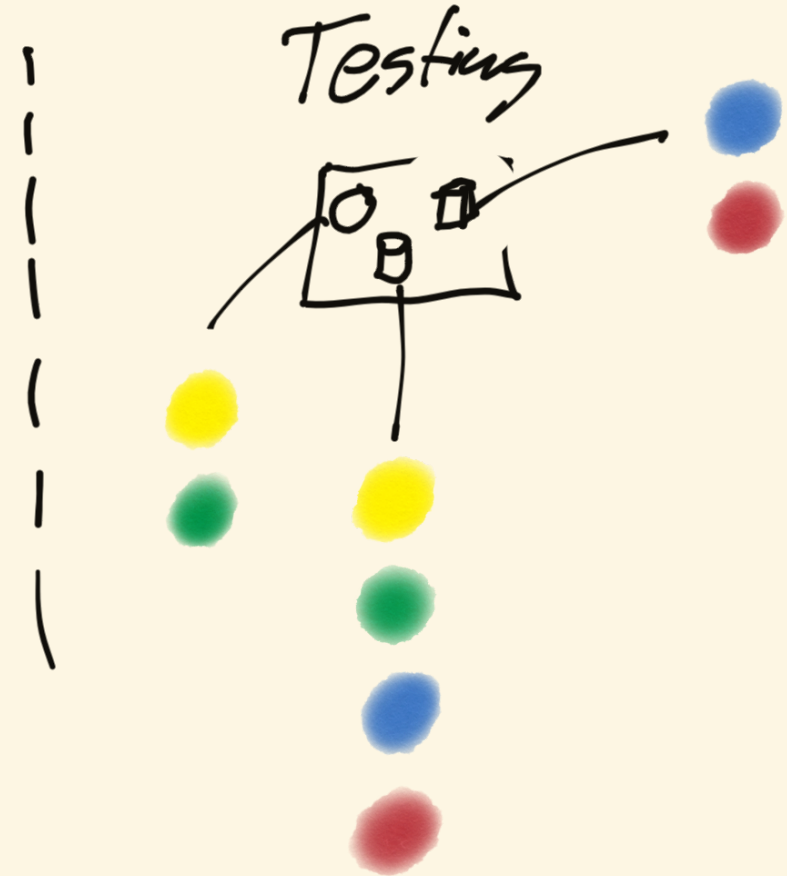
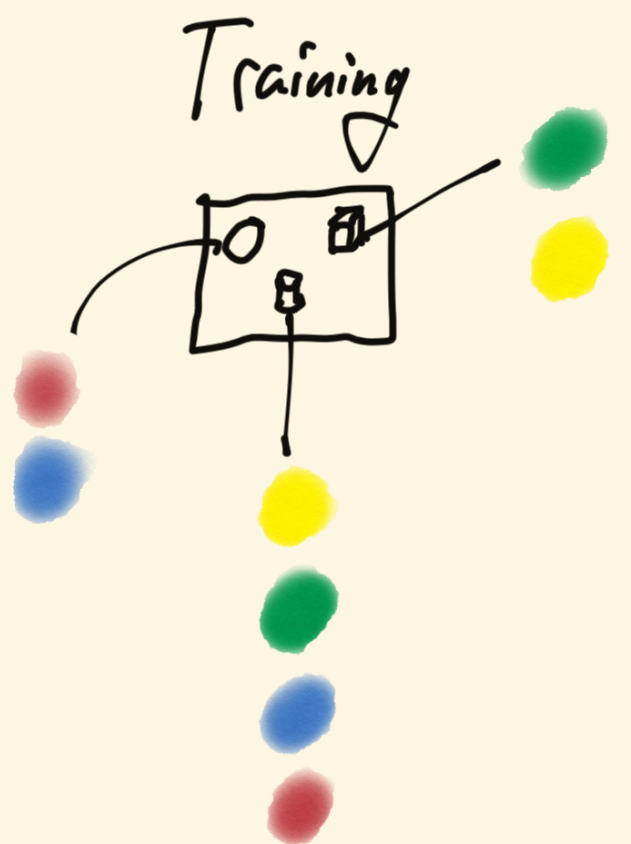


Stammer, W., Schramowski, P., & Kersting, K. (2021). Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In Proceedings of CVPR (pp. 3619-3629).

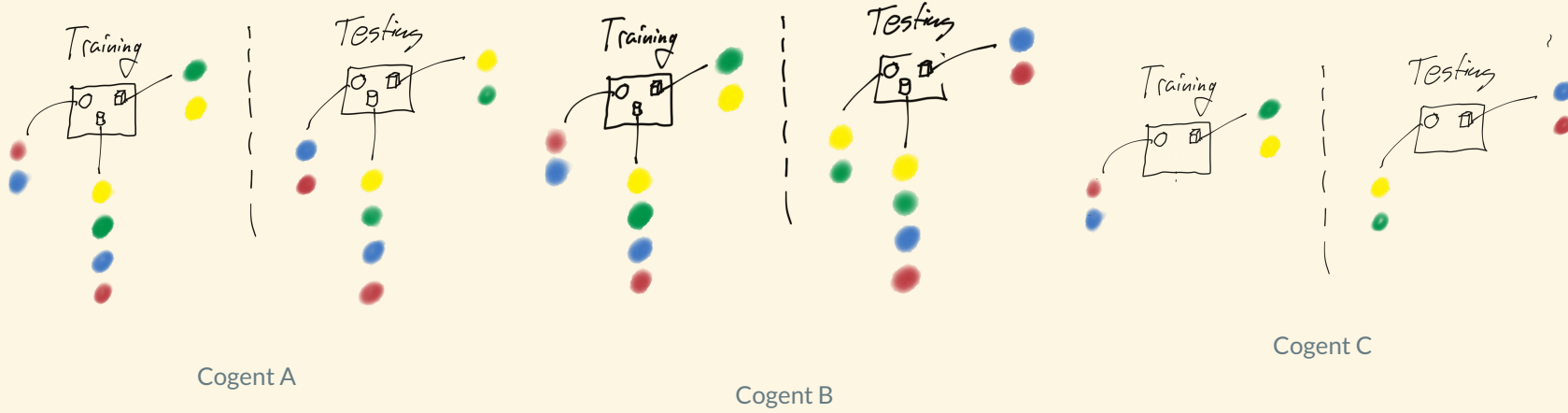


Stammer, W., Schramowski, P., & Kersting, K. (2021). Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In Proceedings of CVPR (pp. 3619-3629).

Validation (confounded)	Test (non-confounded)	Class Rule
		Large (gray) cube and Large cylinder
		Small metal cube and Small (metal) sphere
		Large blue sphere and Small yellow sphere



Model	All	CoGent A	CoGent B	CoGent C
NS-VQA	0.8840	-	0.2130	-
CLIP	0.5823	0.5410	0.4894	0.3798/0.3685



**CLIP HAS SEEN COLORS, SHAPES ET C.
DURING TRAINING, LIKELY ALSO ORIGINAL
CLEVR**

**NS-VQA RELIES TOO HEAVILY ON THE
INDUCTIVE BIAS FROM THE DATA. DOES
NOT SEEM TO LEARN CATEGORIES.**

**COMPOSITIONALITY IN MULTIMODAL DATA
IS ONLY TESTED FOR IN ONE LAYER, NO
HIERARCHICAL LEVELS OF ABSTRACTION**

Meta-training episodes

Possible inputs: dax, wif, lug, zup

Possible outputs: ●, ●, ●, ●

Support set

dax ●
wif ●
lug ●

Support set

dax ●
lug ●
zup ●

Test episode

Support set

wif ●
lug ●
zup ●

Query set

wif zup dax ● ● ●
lug dax lug zup lug ● ● ● ● ●
dax wif lug ● ● ●
...

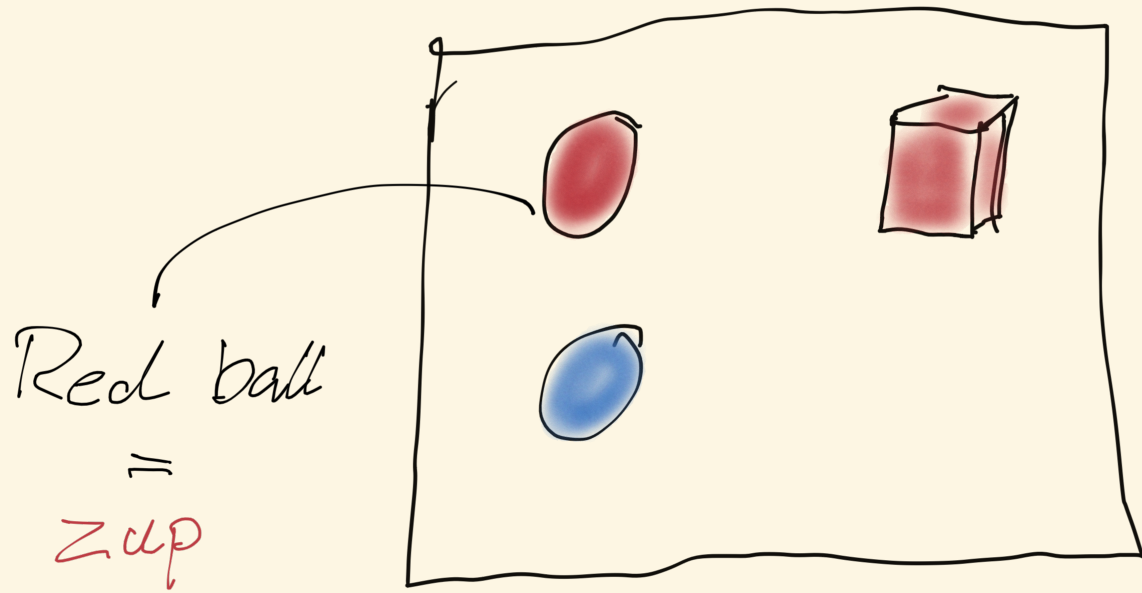
Query set

dax dax ● ●
wif dax lug zup lug wif ● ● ● ● ● ● ●
wif lug lug ● ● ●
...

Query set

zup dax wif ● ● ●
lug zup lug wif dax zup ● ● ● ● ● ● ●
lug dax dax wif lug ● ● ● ● ●
...

* Lake, B. M. (2019). Compositional generalization through meta sequence-to-sequence learning. Advances in neural information processing systems, 32.



Red ball
=
zup

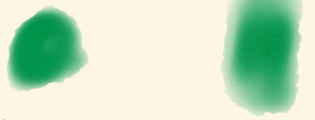
Remove all zups.
How many objects are
left?

Zax ^{trax} : blargh
a wif



a lug ^{trax}
a cylinder of
matching color

: Zax

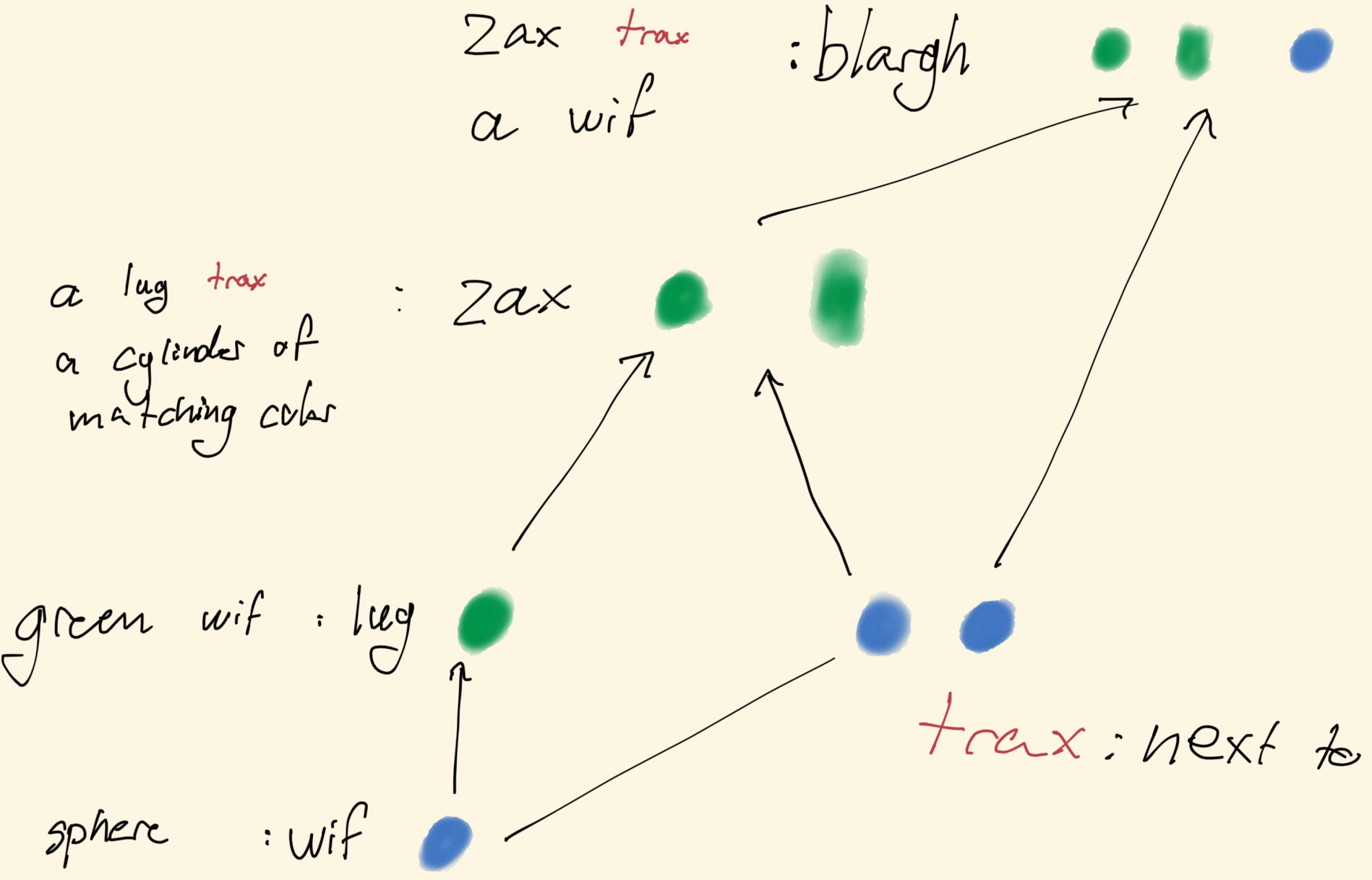
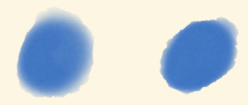


green wif : lug



^{trax} : next to

sphere : wif



- Hierarchical compositions with pseudowords
 - Learning category abstractions
- Analyse **curriculum learning** vs. random order
- Evaluate on tasks different levels of comprehension
 - **Wug Test, Chromium Test** from dev. psychology
 - CLEVR-Math
- Utilize hierarchy to construct Maximize Presupposition-tasks
 - The **cat** is big vs. The **animal** is big



Hugging Face

<https://huggingface.co/datasets/dali-does/clevr-math>



<https://github.com/dali-does/clevr-math>

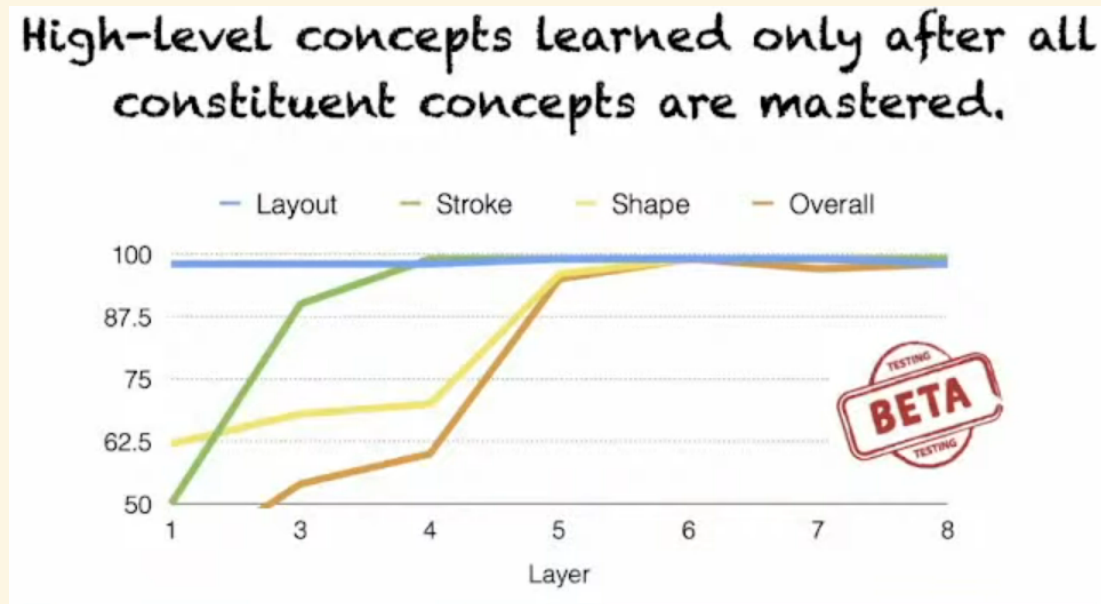
- Neuro-symbolic system with stronger linguistic capabilities
- Prompting, e.g. use chatGPT to parse program from instructions
- Add intermediate representations

Type	Template
Remove group	<i>Remove all How many ... are there?</i> <i>Take away all How many objects are there?</i>
Remove subset	<i>Remove X How many ... are there?</i>
Insertion	<i>Add X How many ... are there?</i>
Count backwards	<i>How many ... must be removed to get X ... ?</i> <i>Take away How many were removed if there are X ... left?</i>
Multihop	<i>Take away all A. Remove all B. How many objects are left?</i>
Adversarial	<i>Remove all A. How many B are left?</i>

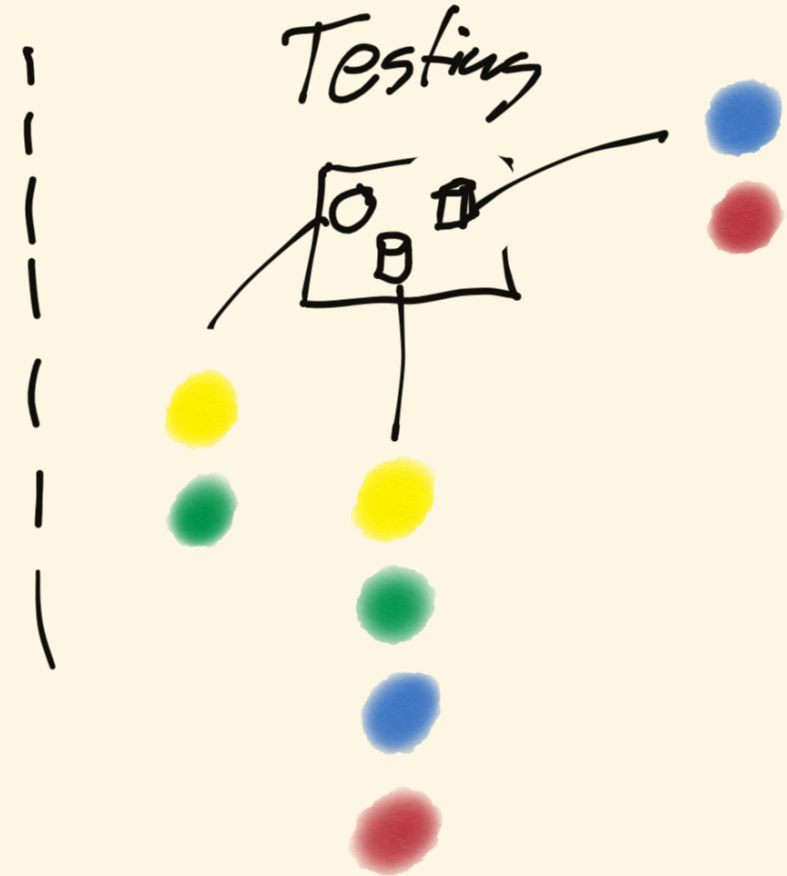
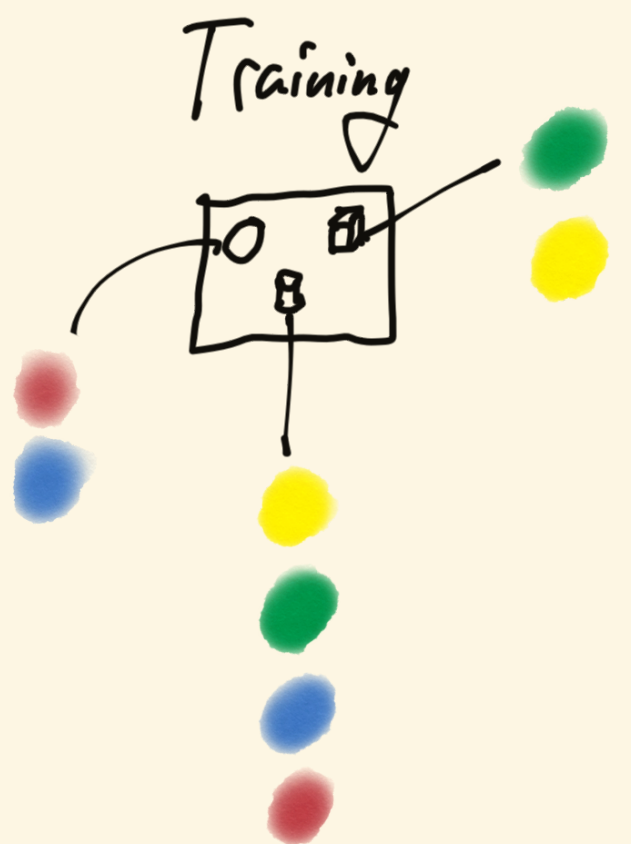
... == <size><color><material><shape>

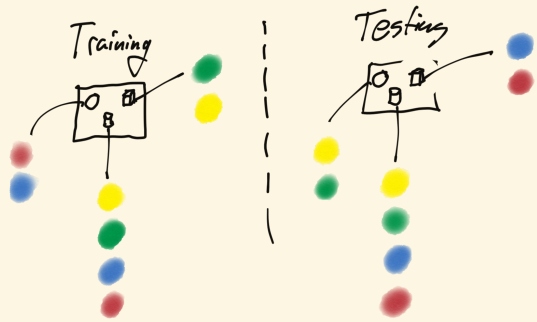
PICTURE

This work should be complemented by [probing of the internal structure of a model](#).

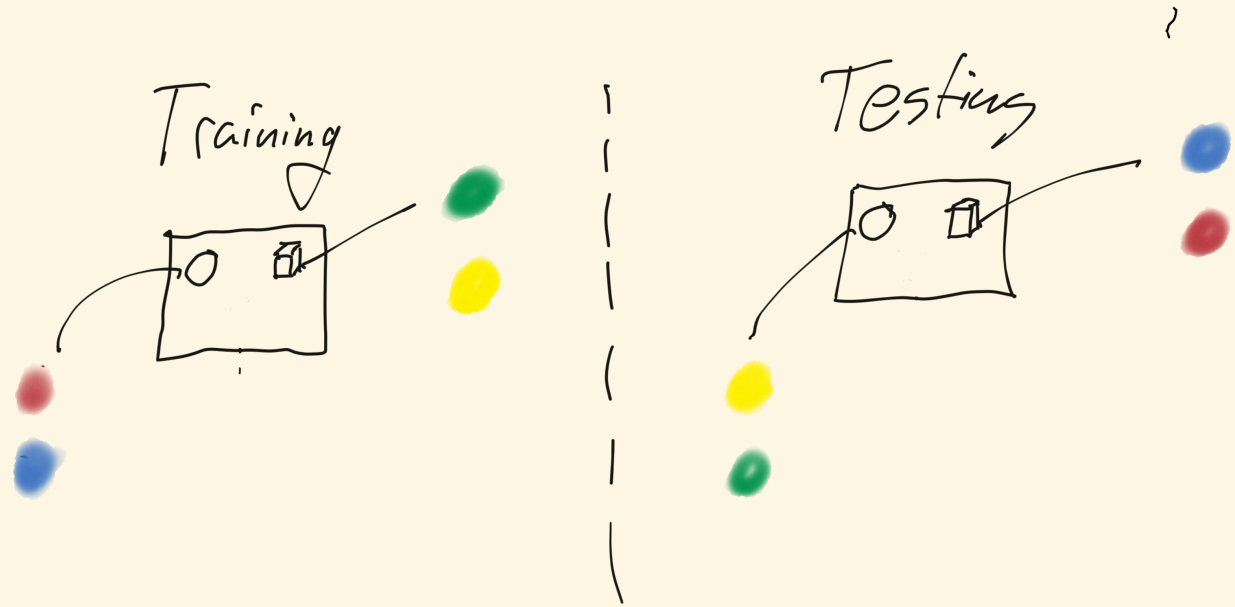


Lovering, C., & Pavlick, E. (2022). Unit testing for concepts in neural networks. *Transactions of the Association for Computational Linguistics*, 10, 1193-1208.

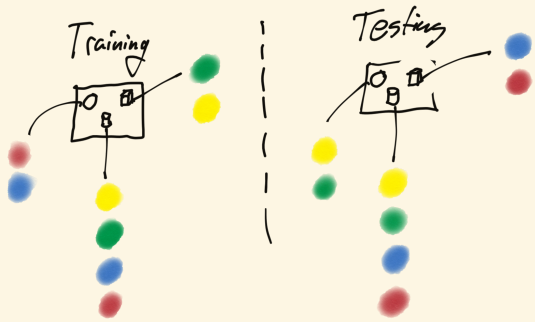




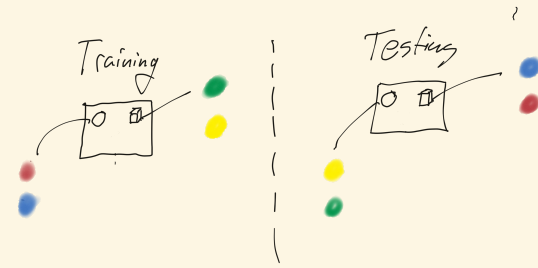
Cogent B



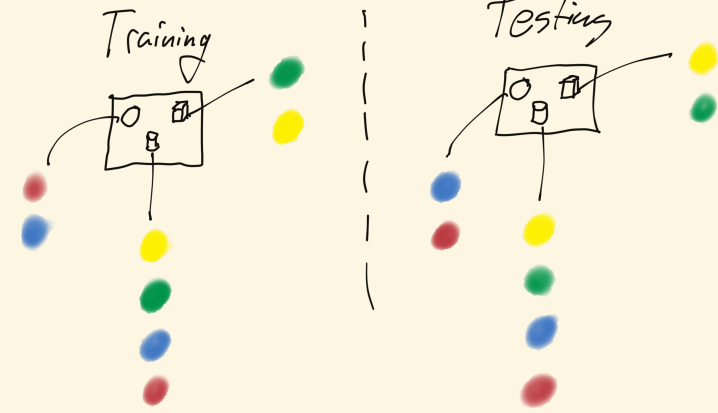
Cogent C



Cogent B



Cogent C



Cogent A