Experience Grounds Language

Yonatan Bisk Ari Holtzman Jesse Thomason

Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, Joseph Turian

EMNLP 2020

A Few Open Questions

What is the limit of text-based meaning representations?

Should we be learning this way? Is it data-efficient/effective?

Why should models learn this way, but not humans?

A Few Open Questions

What is the limit of text-based meaning representations?

Should we be learning this way? Is it data-efficient/effective?

Why should models learn this way, but not humans?

NLP’s Answer?

A Few Open Questions

What is the limit of text-based meaning representations?

Should we be learning this way? Is it data-efficient/effective?

Why should models learn this way, but not humans?

NLP's Answer?

Intelligence and NLU
don't require seeing,
hearing, or doing, …

A Few Open Questions

What is the limit of text-based meaning representations?

Should we be learning this way? Is it data-efficient/effective?

Why should models learn this way, but not humans?

NLP's Answer?

Intelligence and NLU don't require seeing, hearing, or doing, …

🙈 🙉

A Few Open Questions

What is the limit of text-based meaning representations?

Should we be learning this way? Is it data-efficient/effective?

Why should models learn this way, but not humans?

NLP's Answer?

Intelligence and NLU
don't require seeing,
hearing, or doing, …
but might need speaking.

🙈 🙉

A Few Open Questions

What is the limit of text-based meaning representations?

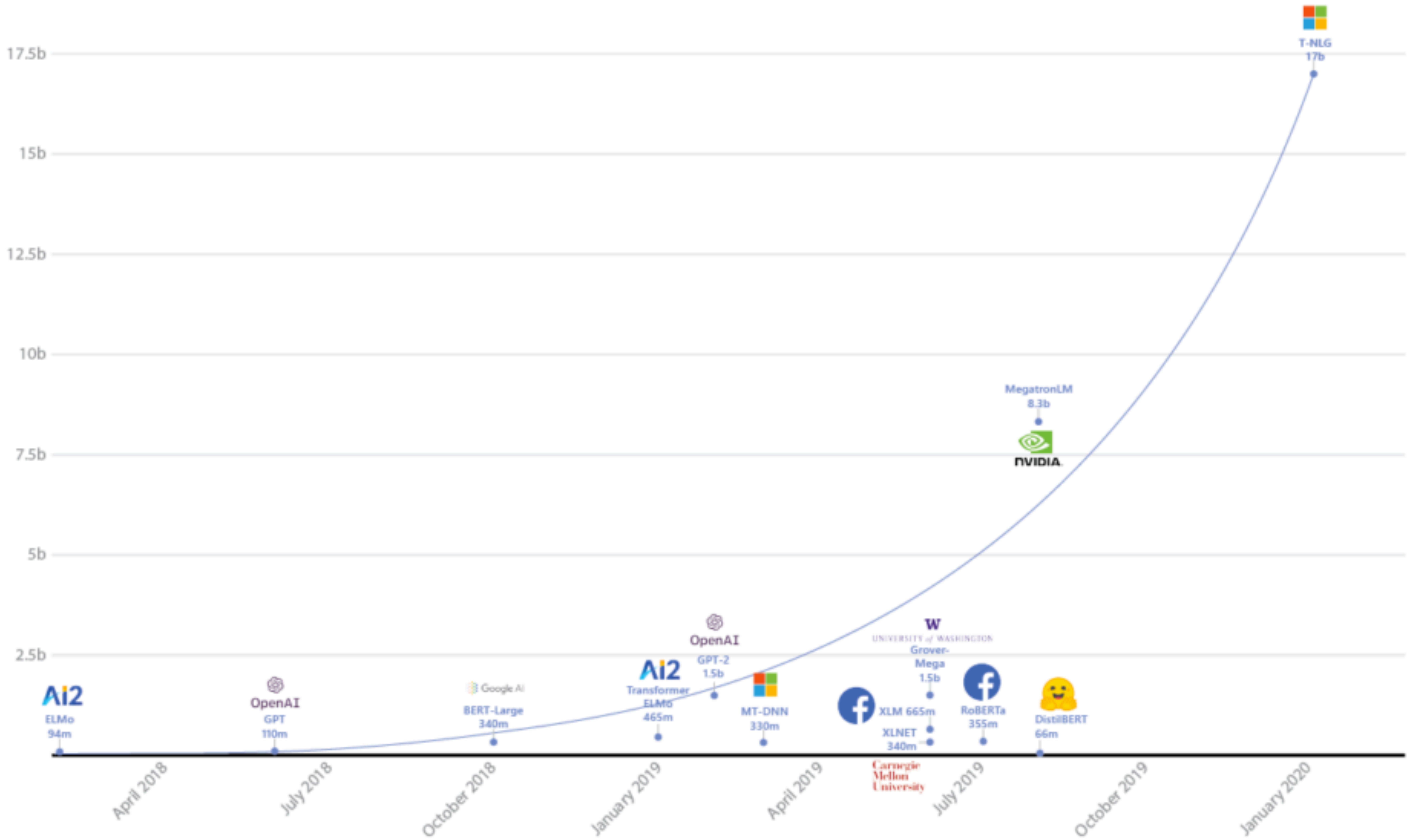
Should we be learning this way? Is it data-efficient/effective?

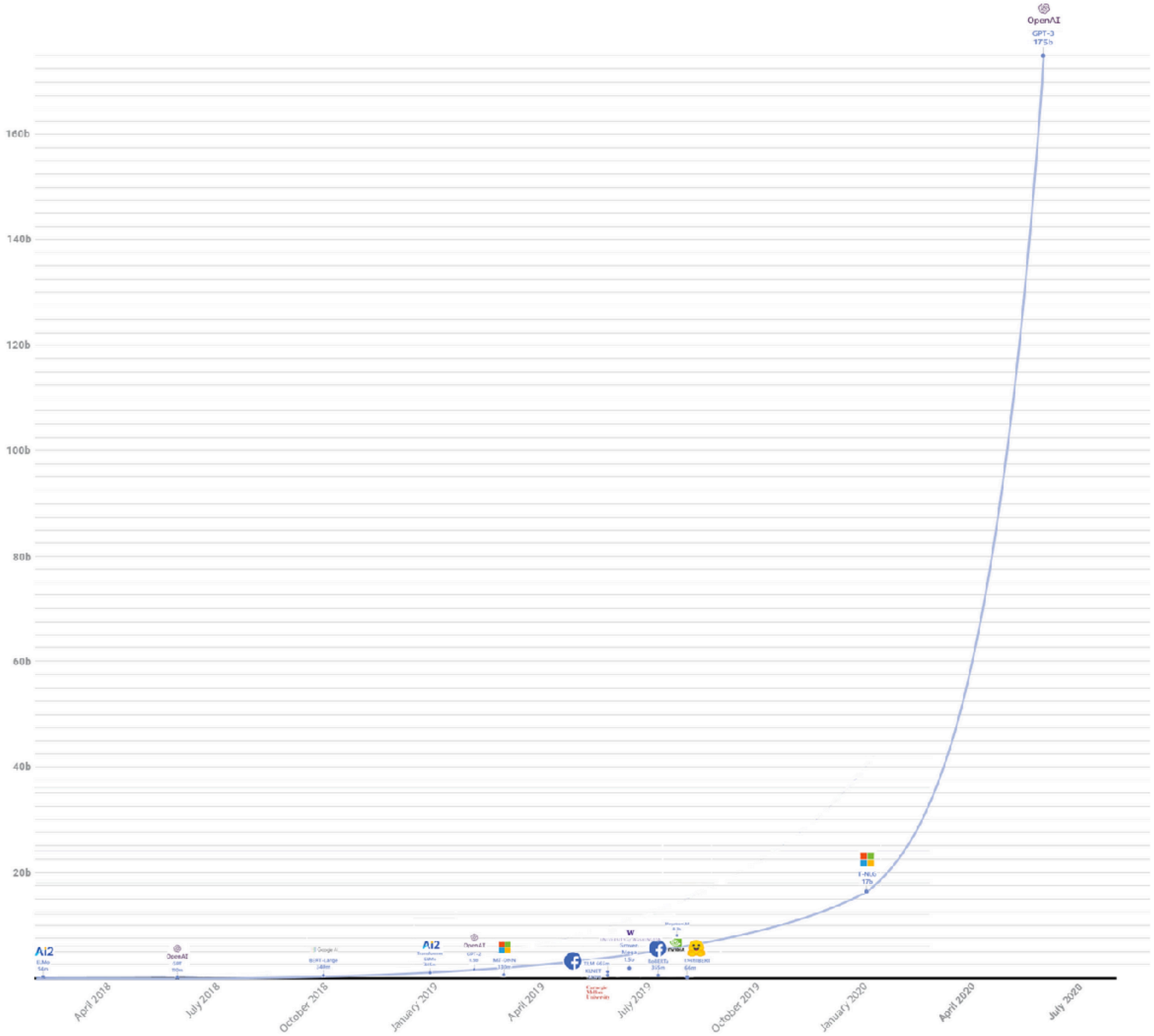
Why should models learn this way, but not humans?

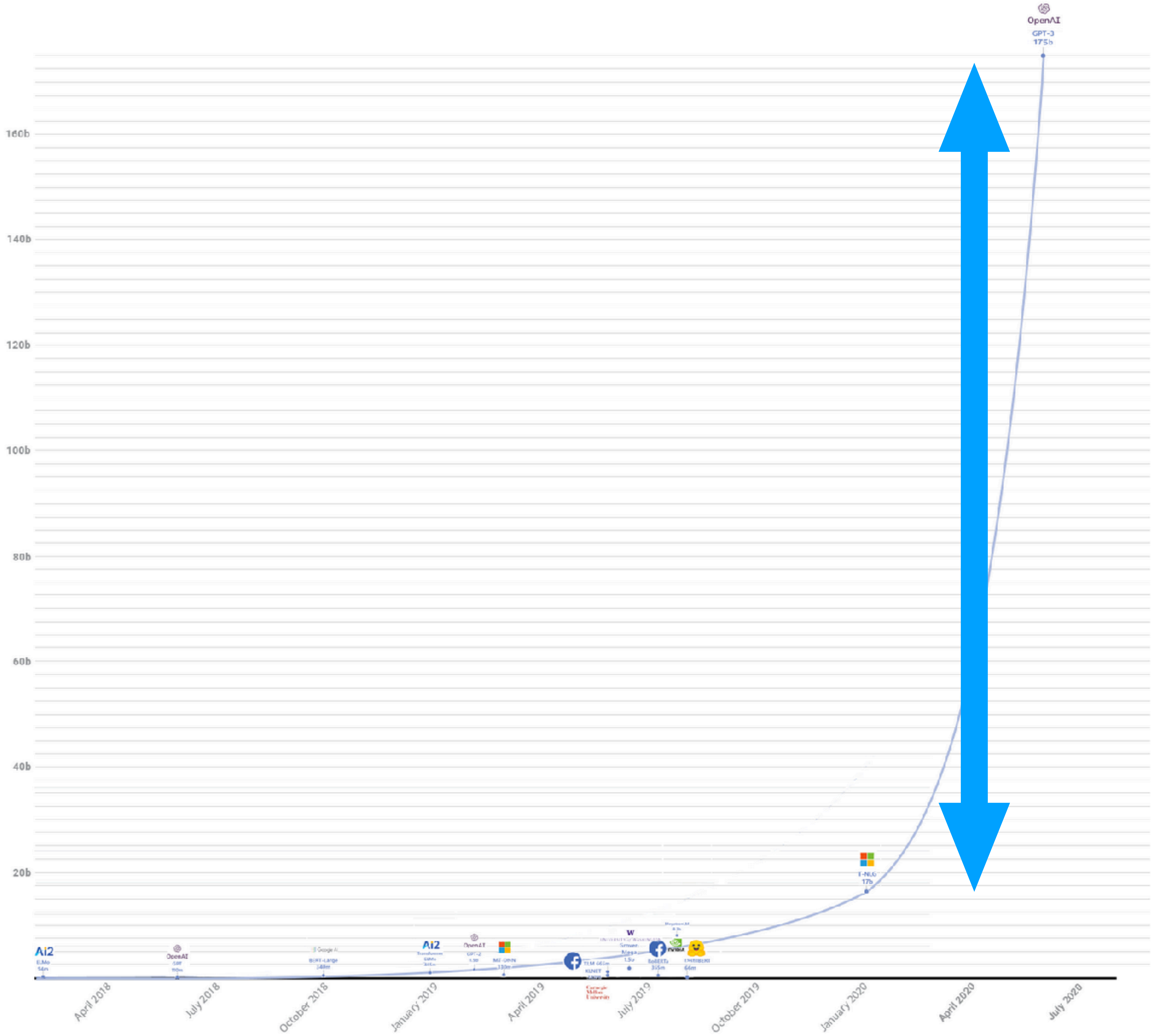
NLP’s Answer?

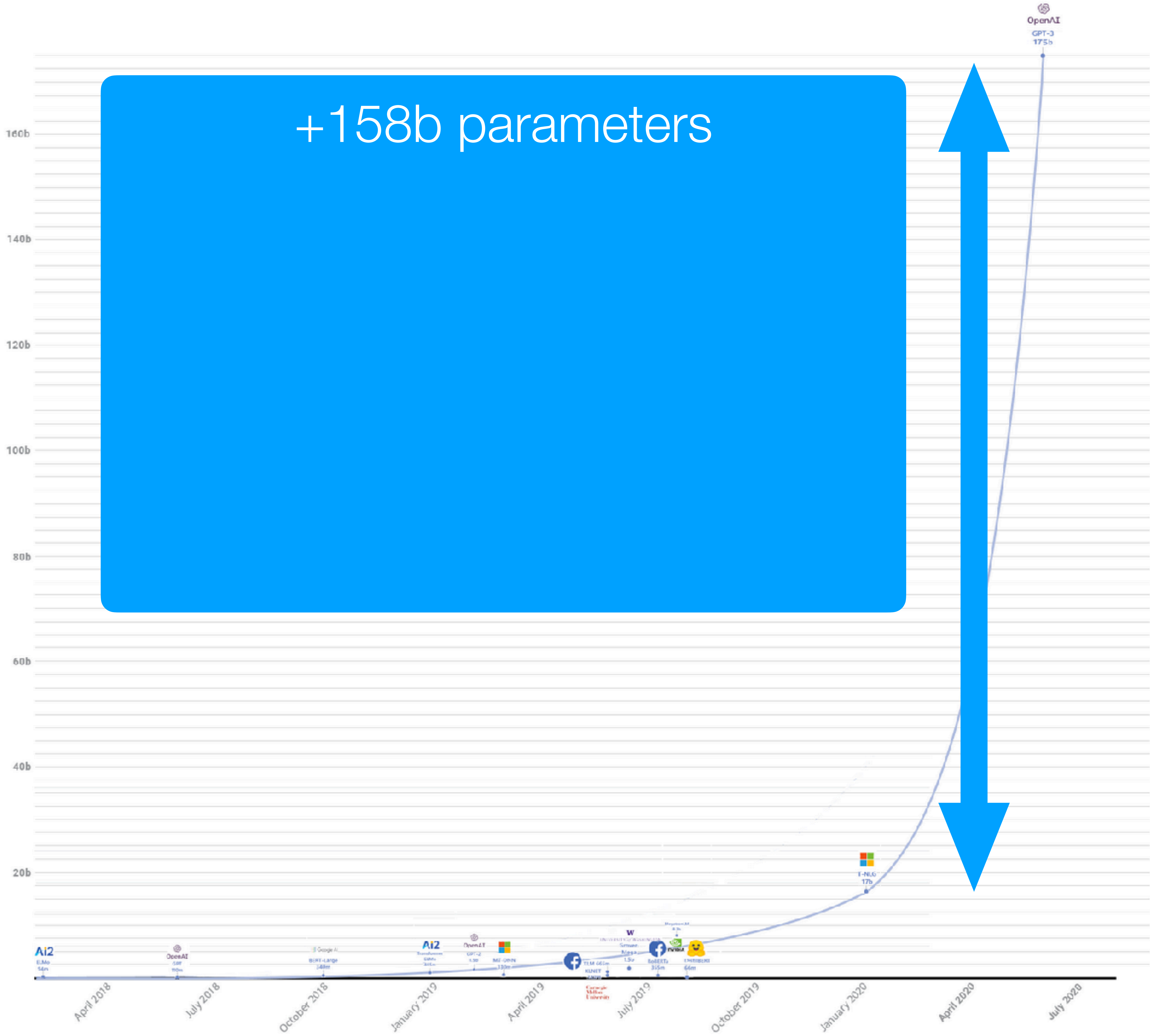
Intelligence and NLU
don’t require seeing,
hearing, or doing, …
but might need speaking.

🙈 🙉 🙊

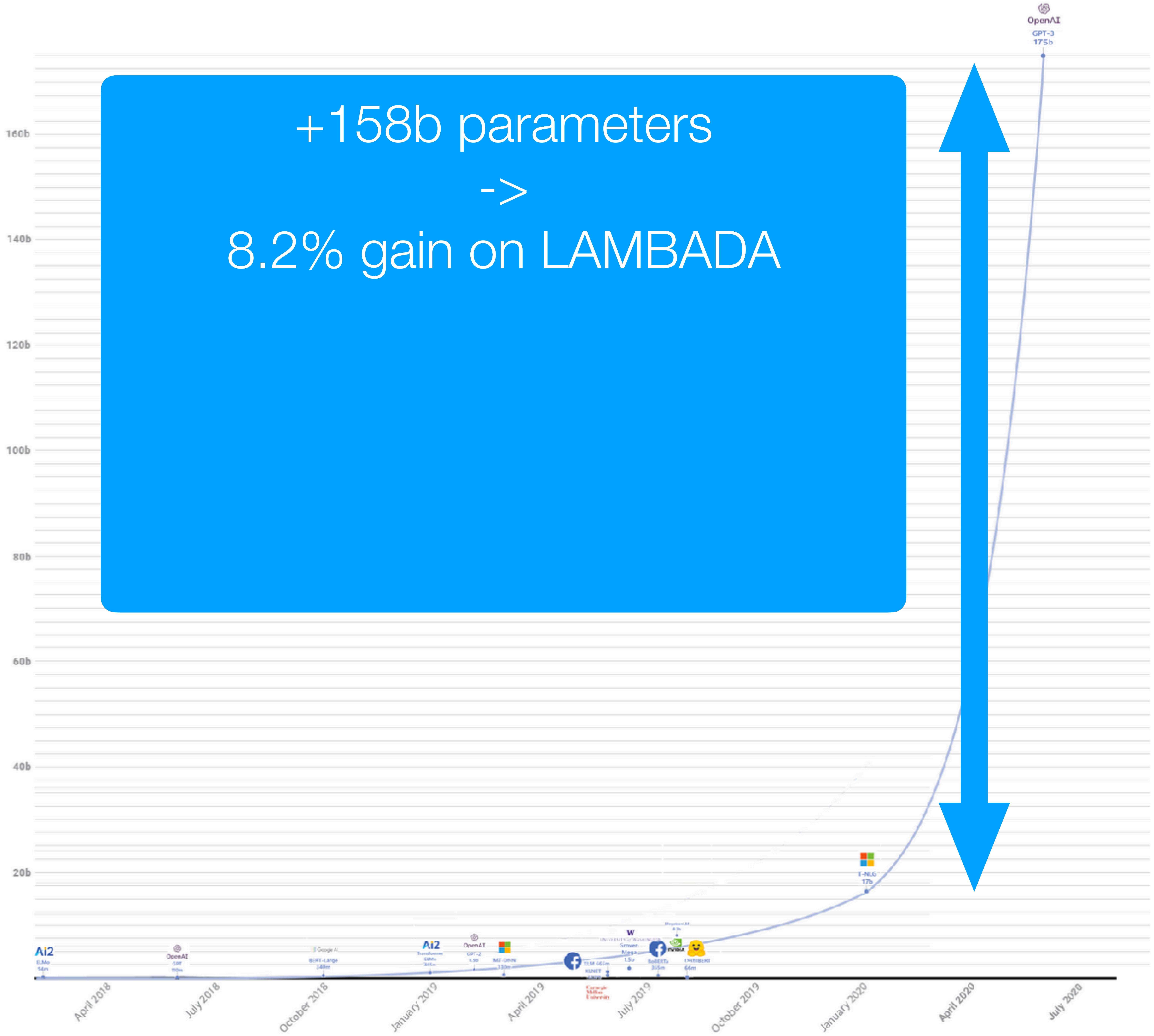

17.5b
15b
12.5b
10b
7.5b
5b
2.5b
Ai2
ELMo
94m
OpenAI
GPT
110m
Google AI
BERT-Large
340m
Ai2
Transformer ELMo
465m
OpenAI
GPT-2
1.5b
MT-DNN
330m
XLM 665m
XLNET
340m
Carnegie Mellon University
W
UNIVERSITY of WASHINGTON
Grover-Mega
1.5b
RoBERTa
355m
DistilBERT
66m
MegatronLM
8.3b
NVIDIA
T-NLG
17b
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019
October 2019
January 2020


OpenAI
GPT-3
175b
160b
140b
120b
100b
80b
60b
40b
20b
T-NLG
17b
Ai2
ELMo
94m
OpenAI
GPT
110m
Google AI
BERT-Large
340m
Ai2
Transformer
ELMo
465m
OpenAI
GPT-2
1.5b
MT-DNN
330m
XLM 665m
XLNET
340m
Carnegie
Mellon
University
W
UNIVERSITY of WASHINGTON
Grover-
Mega
1.5b
RoBERTa
355m
Megatron-LM
8.3b
NVIDIA
DistilBERT
66m
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019
October 2019
January 2020
April 2020
July 2020


OpenAI
GPT-3
175b
160b
140b
120b
100b
80b
60b
40b
20b
T-NLG
17b
Ai2
ELMo
54m
OpenAI
GPT
110m
Google AI
BERT-Large
340m
Ai2
Transformer ELMo
465m
OpenAI
GPT-2
1.5b
MT-DNN
330m
XLM 665m
XLNET
340m
Carnegie Mellon University
UNIVERSITY of WASHINGTON
Grover-Mega
1.5b
RoBERTa
315m
MegatronLM
8.3b
NVIDIA
DistilBERT
66m
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019
October 2019
January 2020
April 2020
July 2020



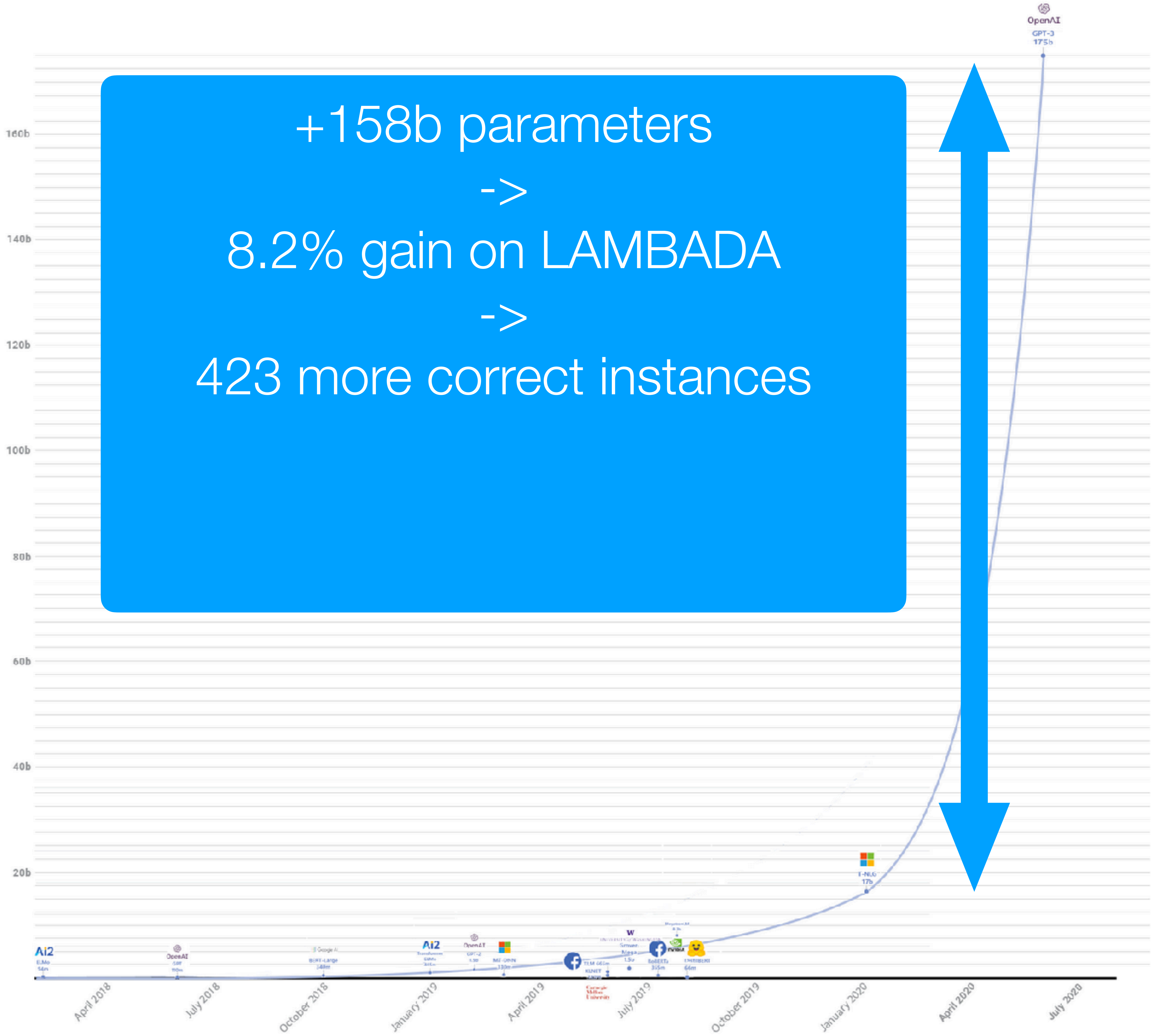
+158b parameters



+158b parameters

->

8.2% gain on LAMBADA



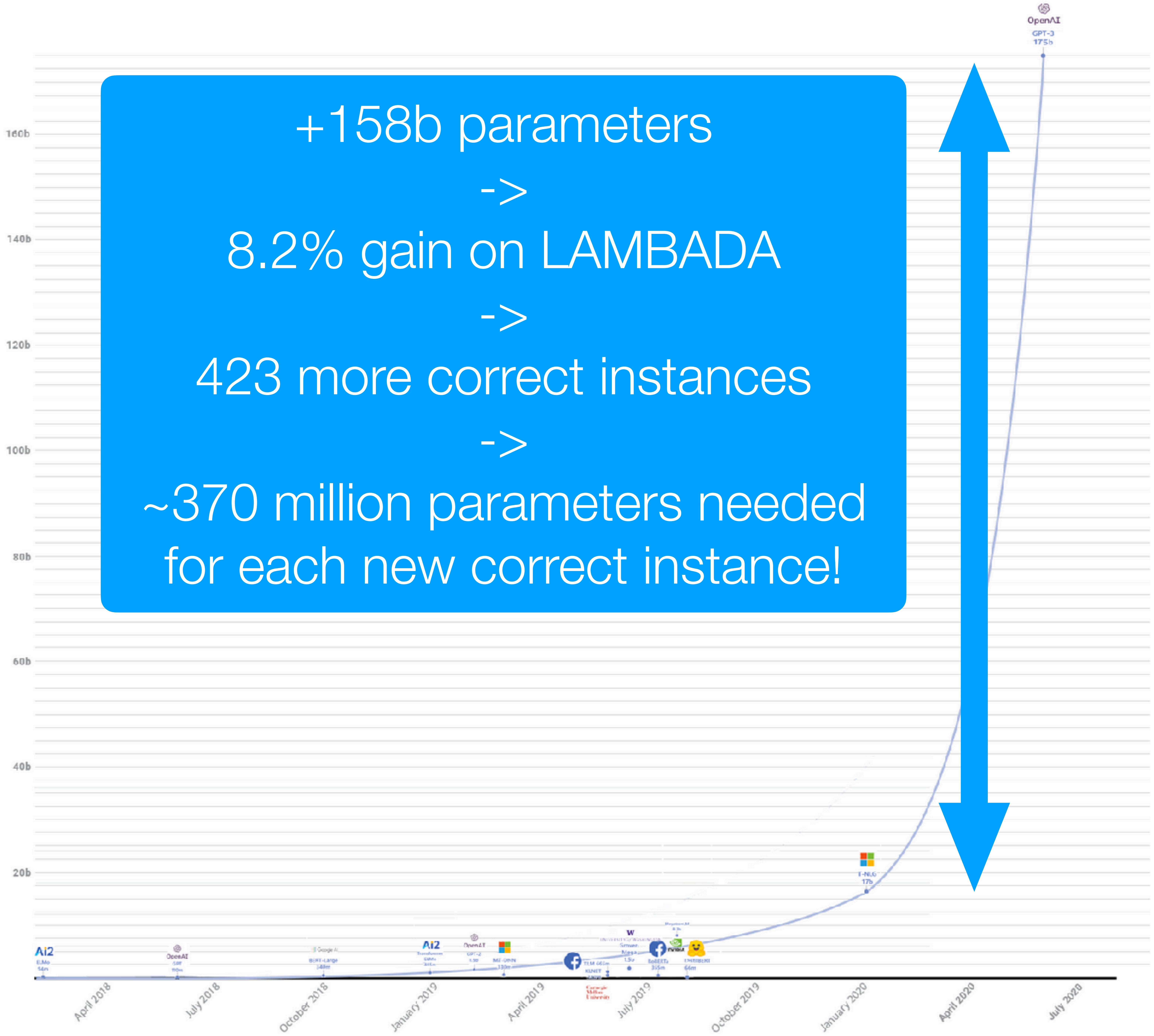
+158b parameters

->

8.2% gain on LAMBADA

->

423 more correct instances



+158b parameters

->

8.2% gain on LAMBADA

->

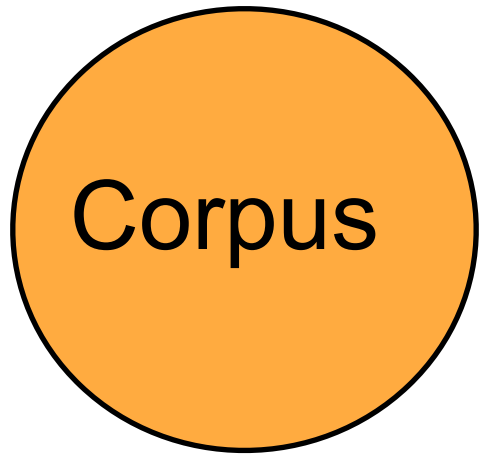
423 more correct instances

->

~370 million parameters needed
for each new correct instance!

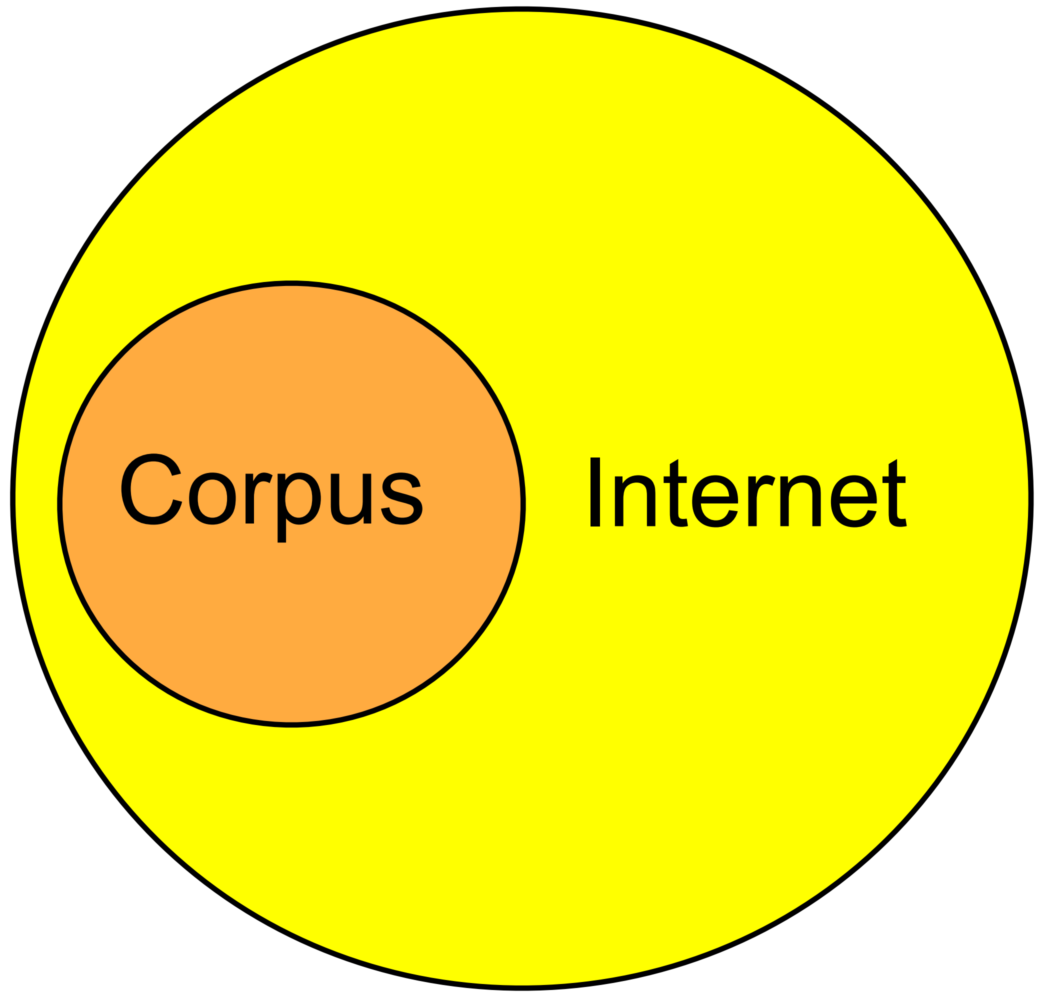
World Scopes

World Scopes

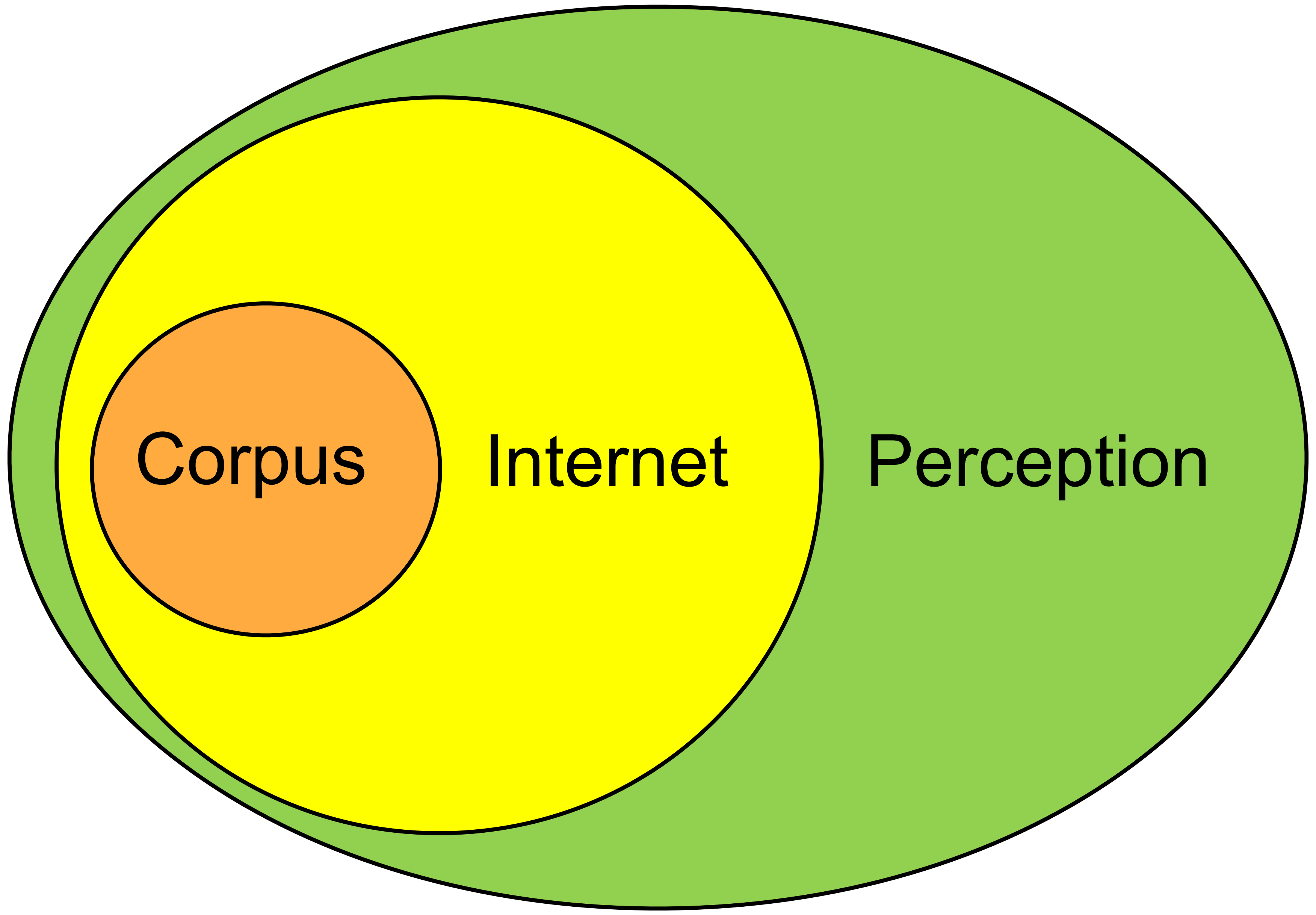


Corpus

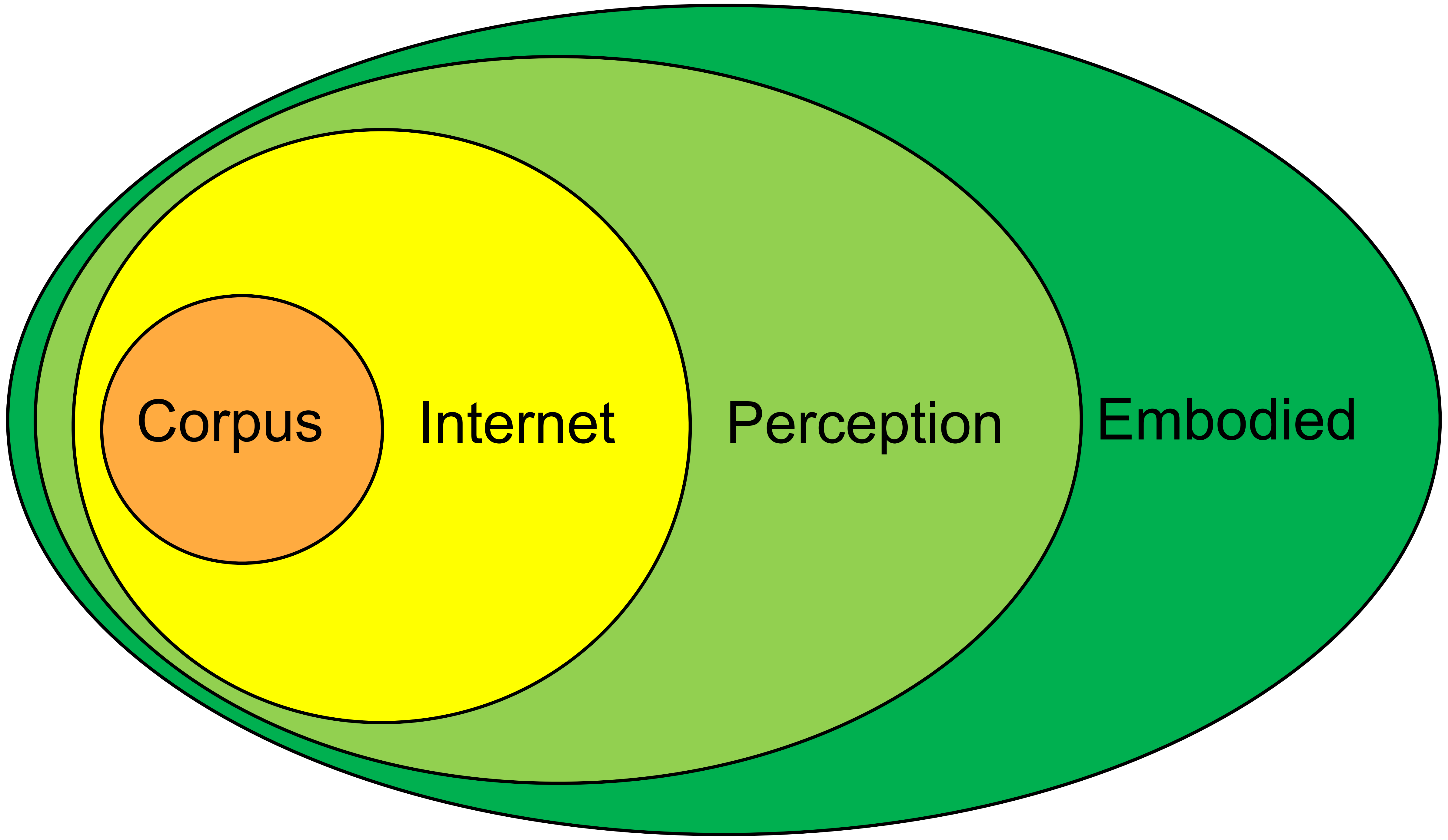
World Scopes



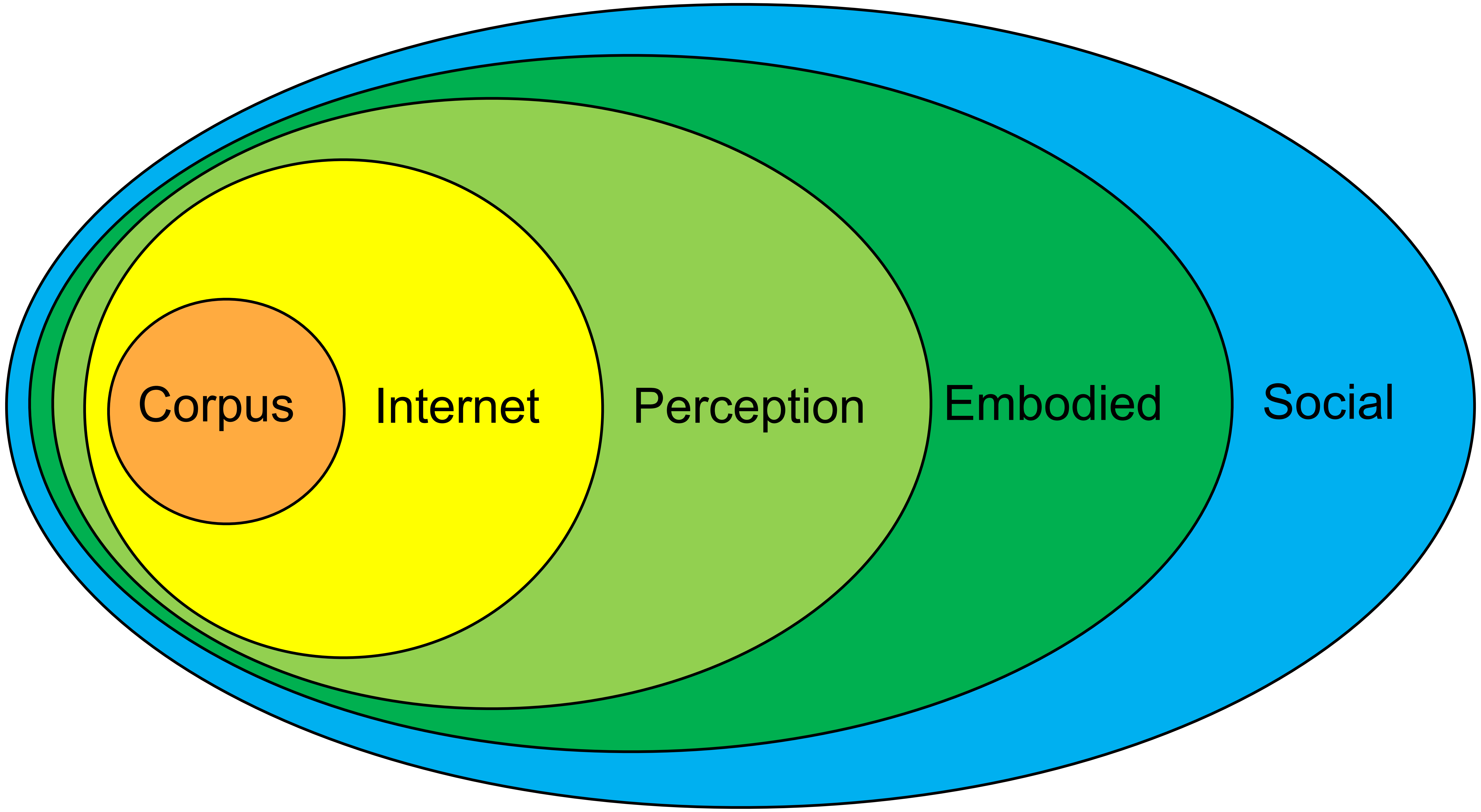
World Scopes



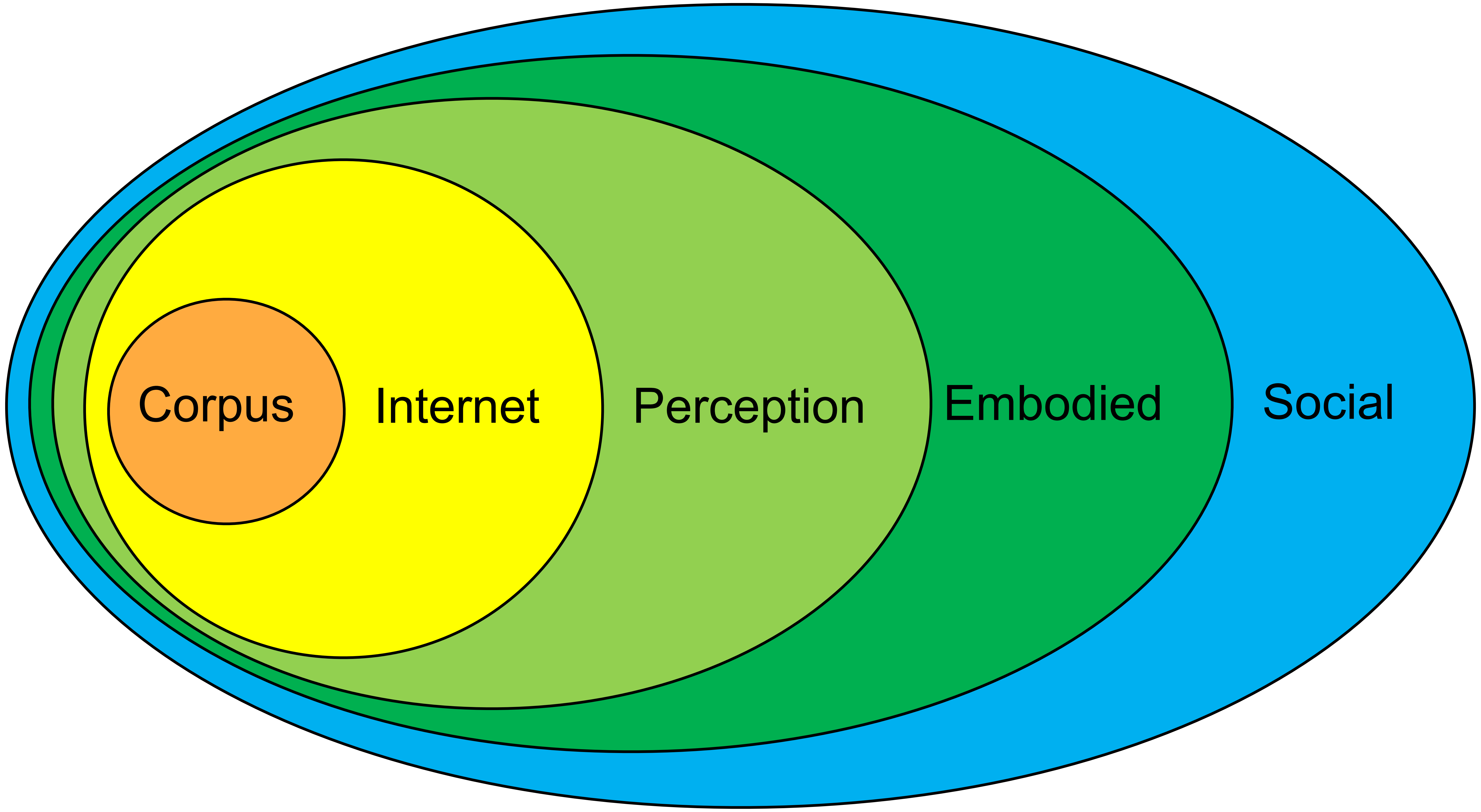
World Scopes



World Scopes

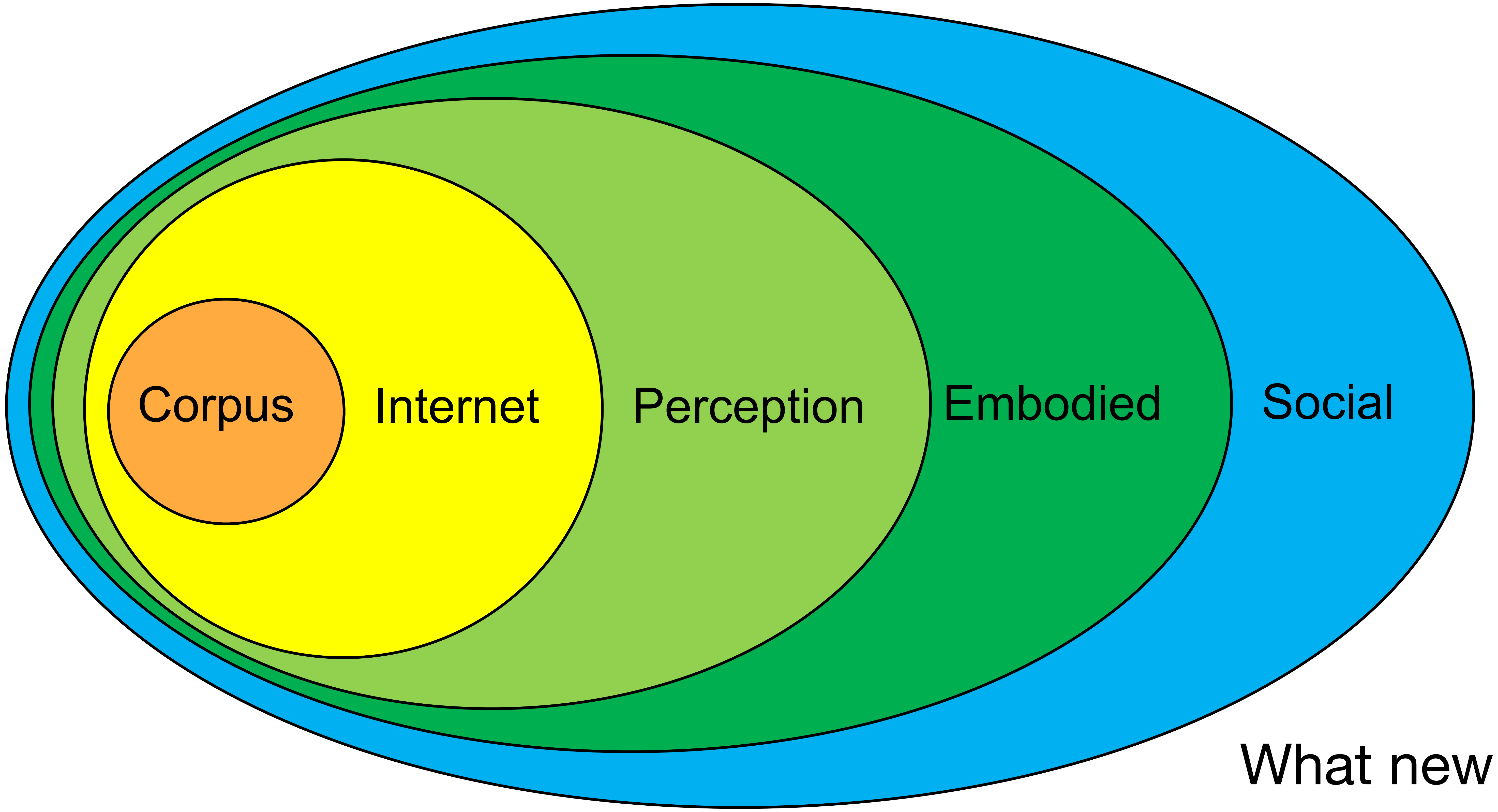


World Scopes



How does reporting bias change?

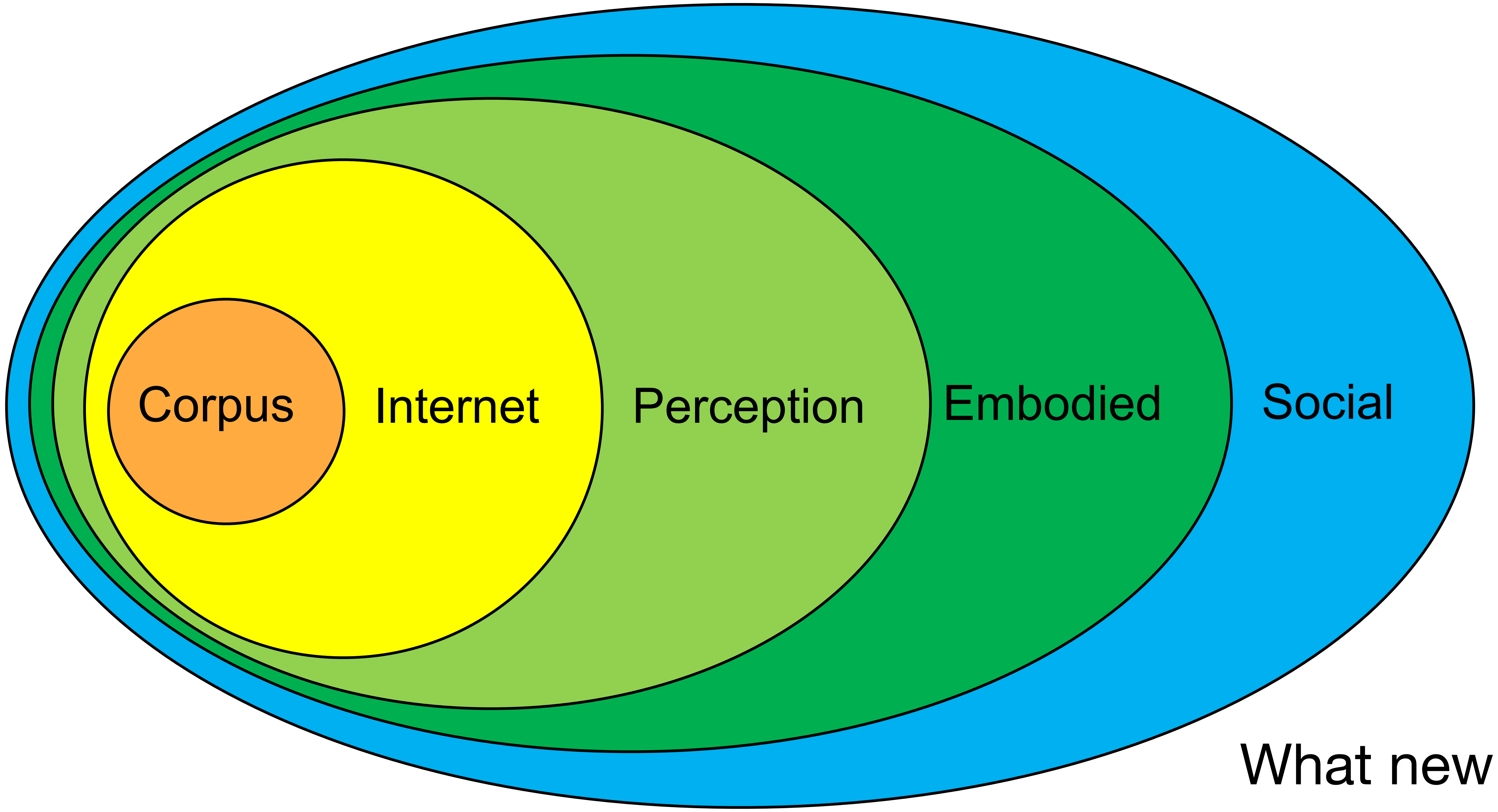
World Scopes



How does reporting bias change?

What new knowledge do models have access to?

World Scopes



How does reporting bias change?

What new knowledge do models have access to?

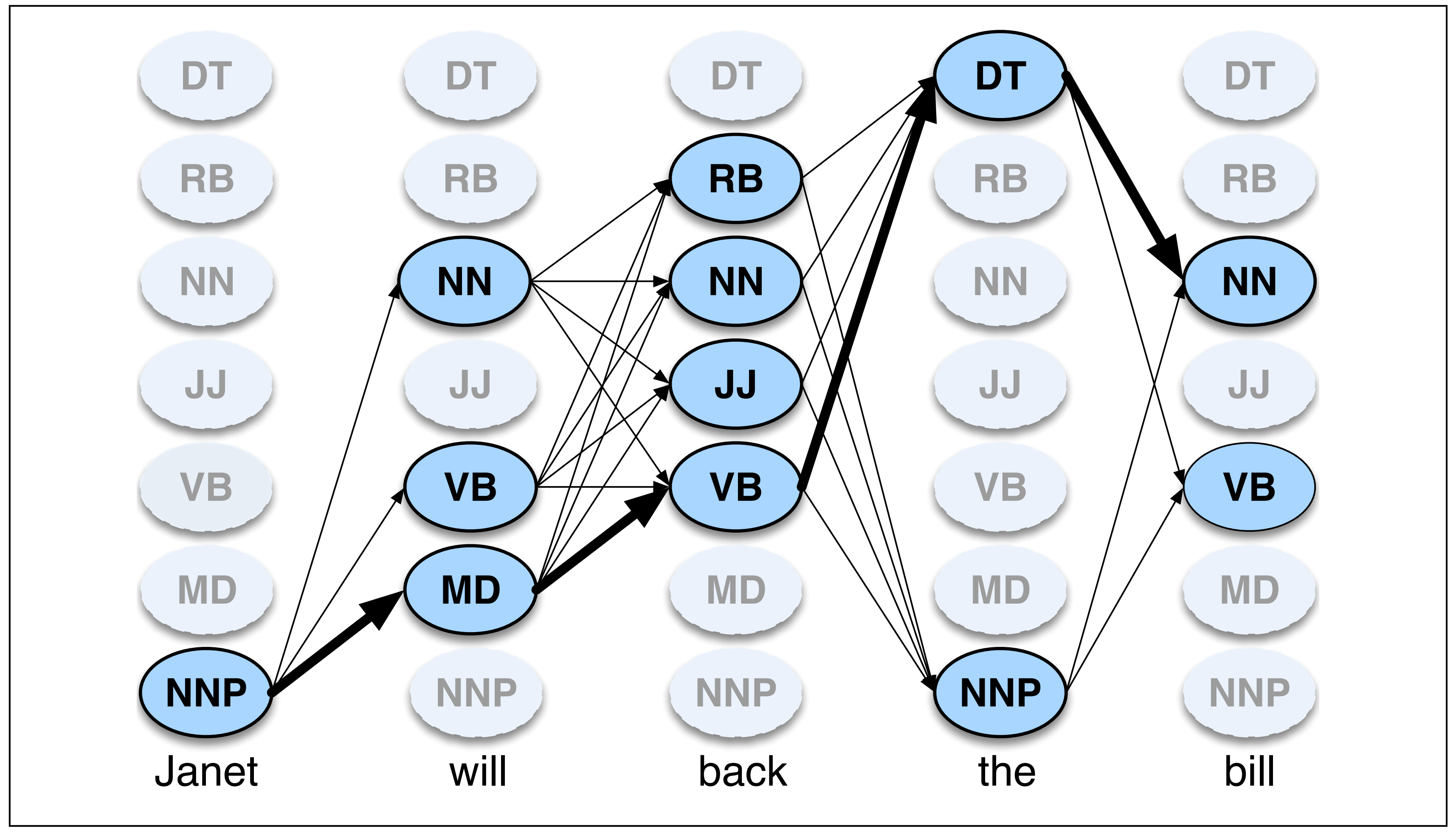
What advances in modeling and fusion are necessary?

WS1 The Corpus

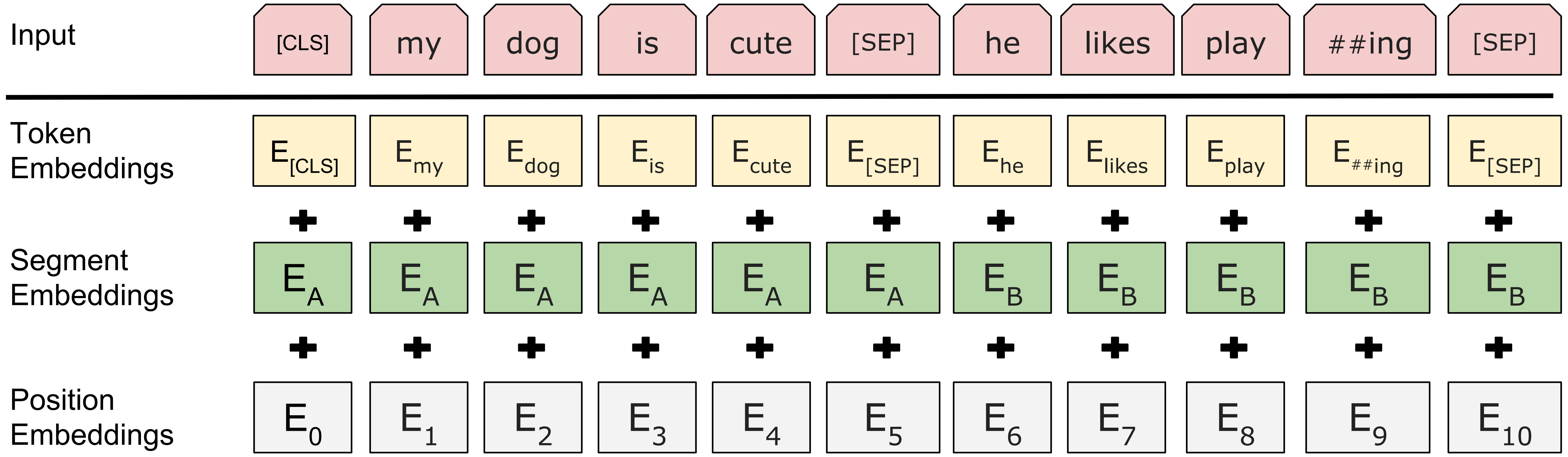
WS2 The (linguistic) Internet

Local context and
Parameters that saturate on a small corpus

Multi-sentence context and
Parameters that saturate eventually?

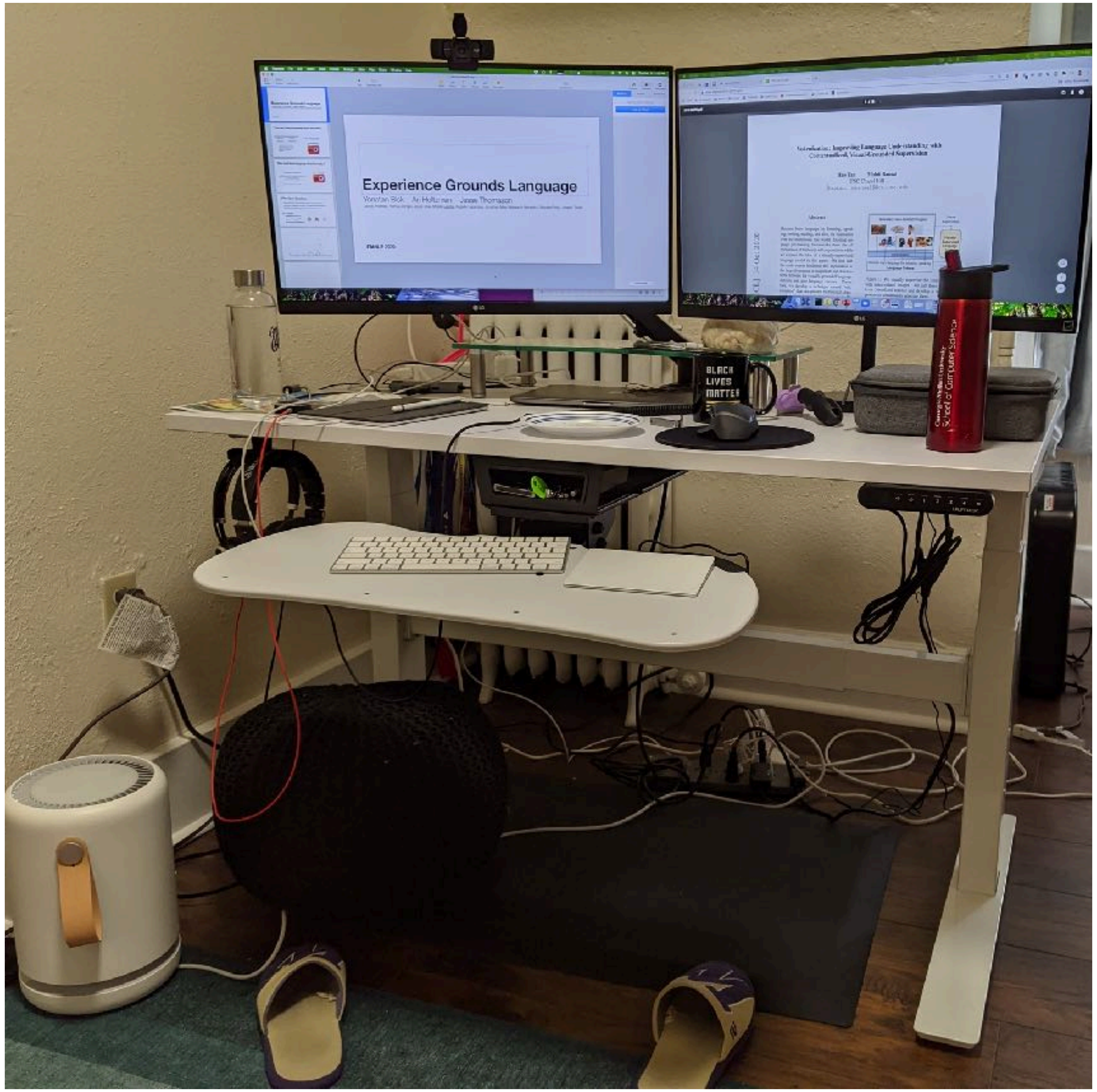


Jurafsky & Martin

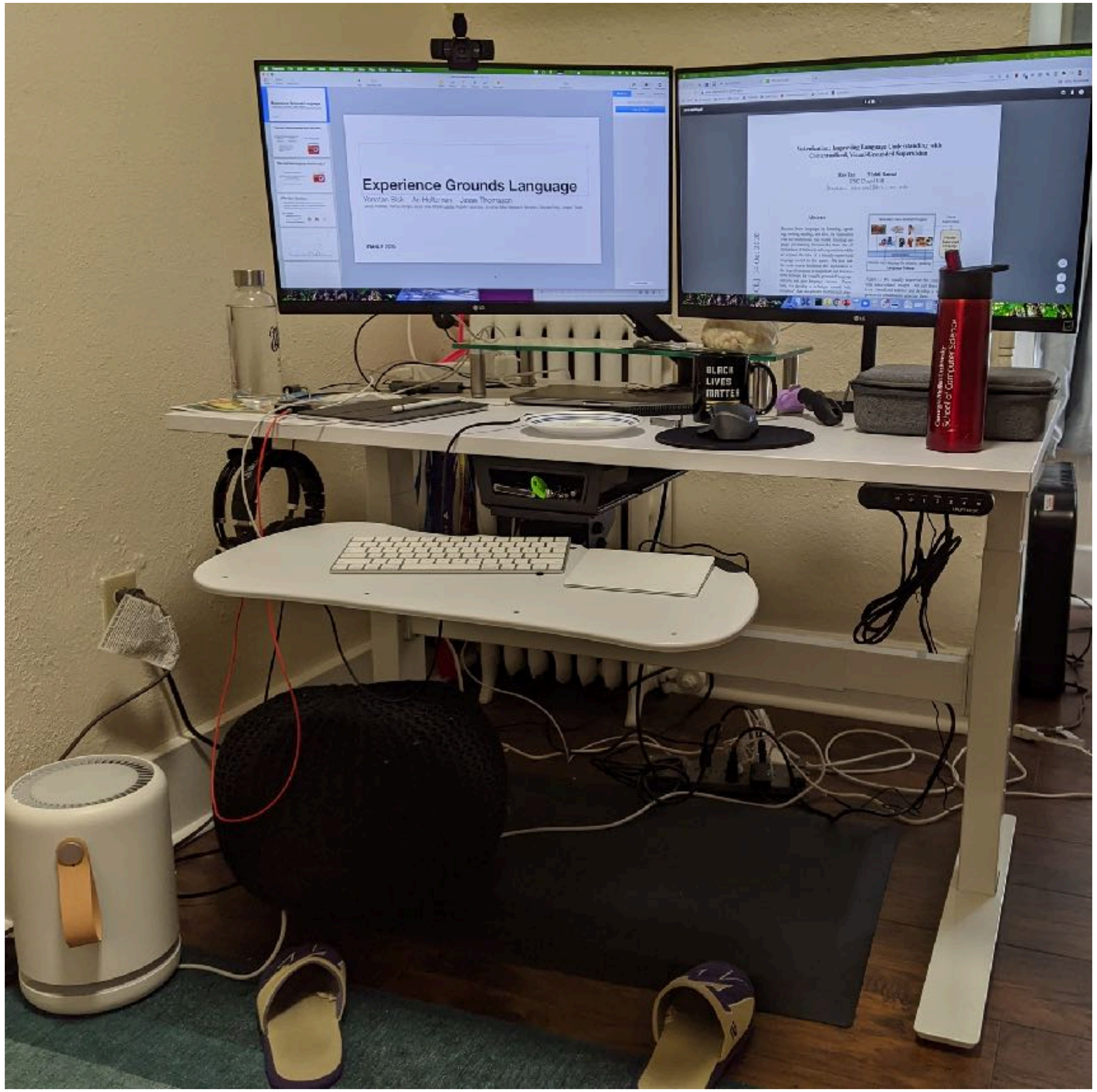

Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{##ing}$ $E_{[SEP]}$
+
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
+
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Devlin et al. 2018

WS3 — How much “knowledge” is in an image?

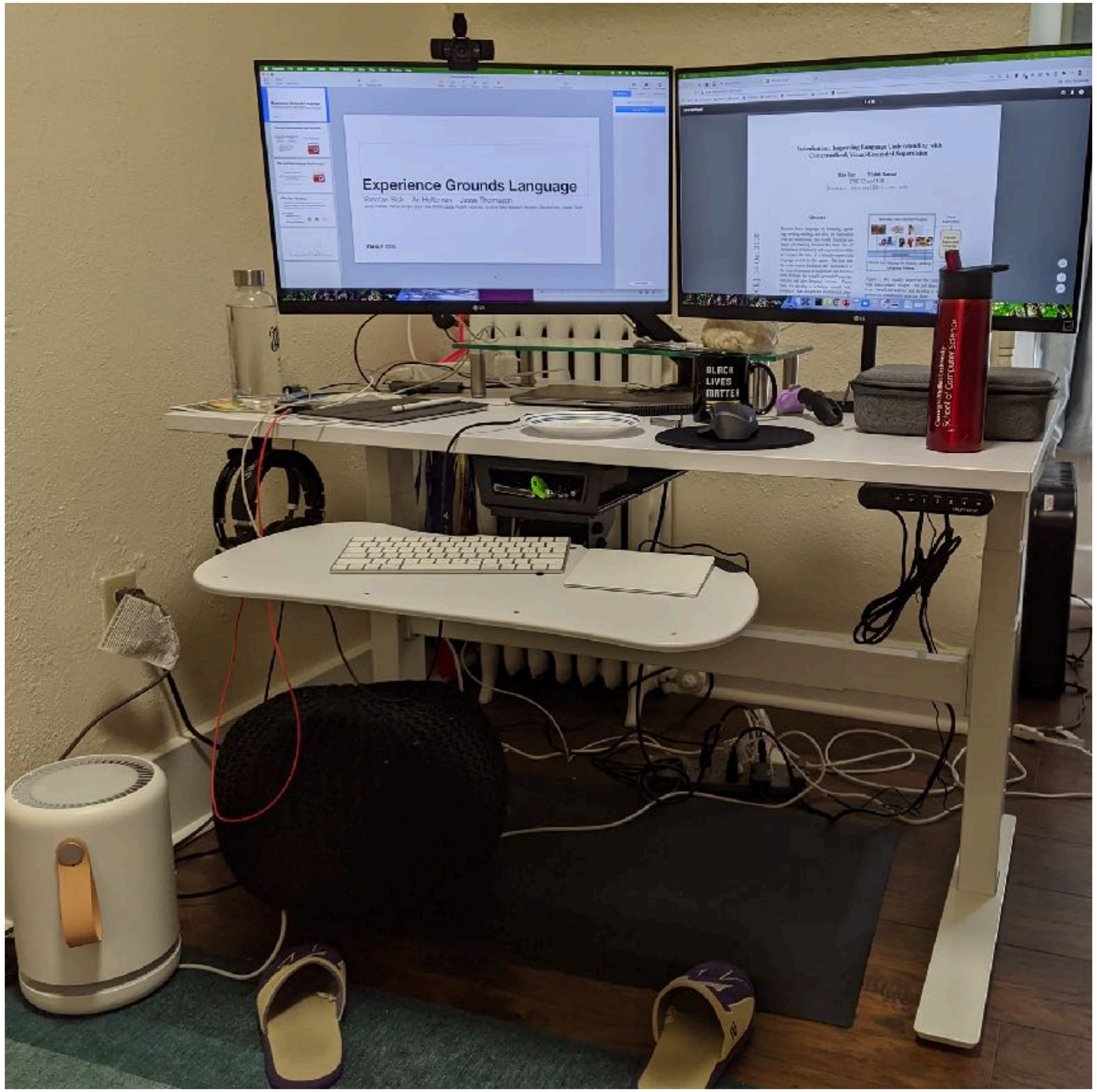


WS3 — How much “knowledge” is in an image?



Is it every unary relationship?

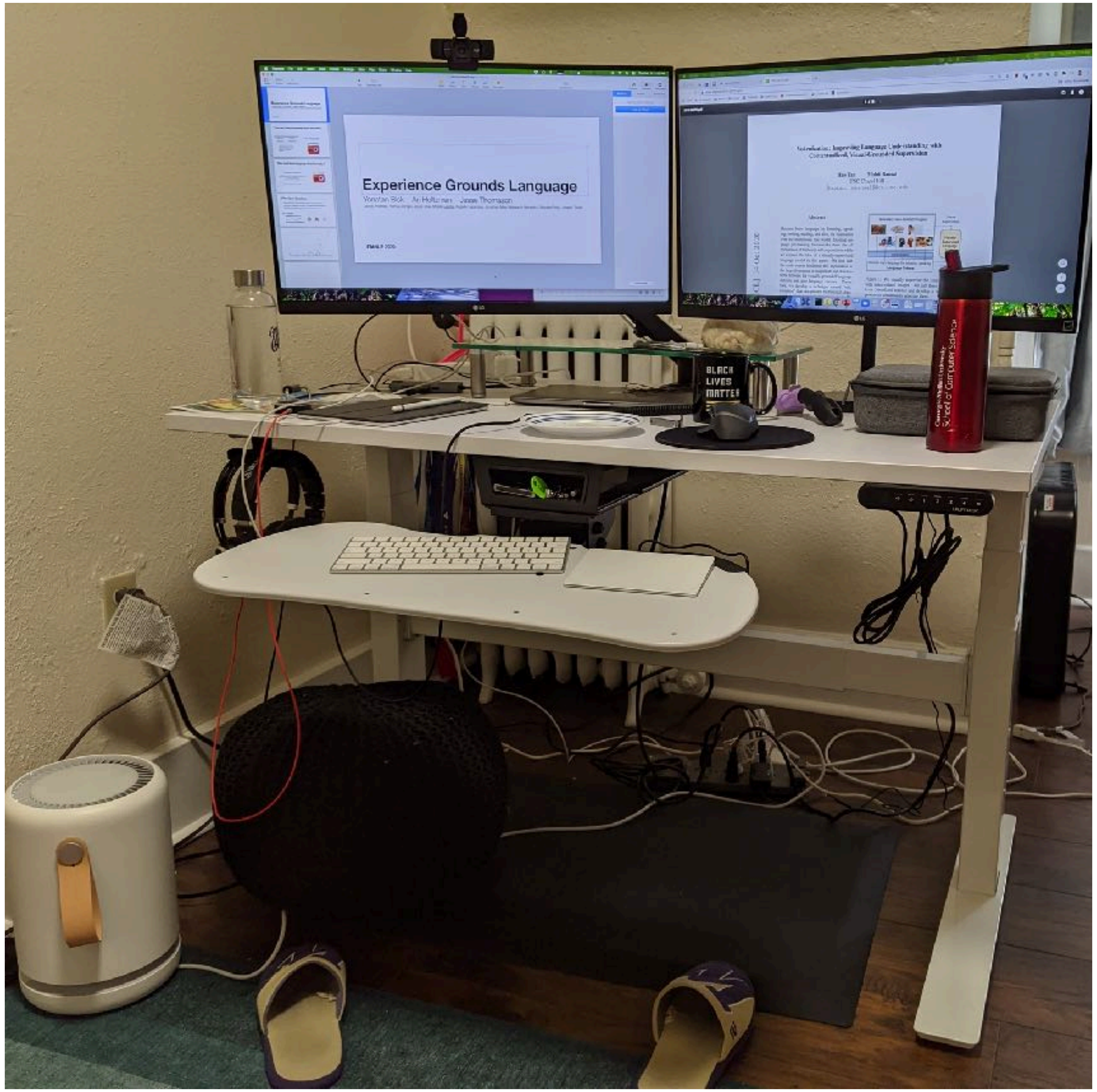
WS3 — How much “knowledge” is in an image?



Is it every unary relationship?

Water bottles are clear
Water bottles are metallic
Mugs have words
Water bottles have words
…

WS3 — How much “knowledge” is in an image?

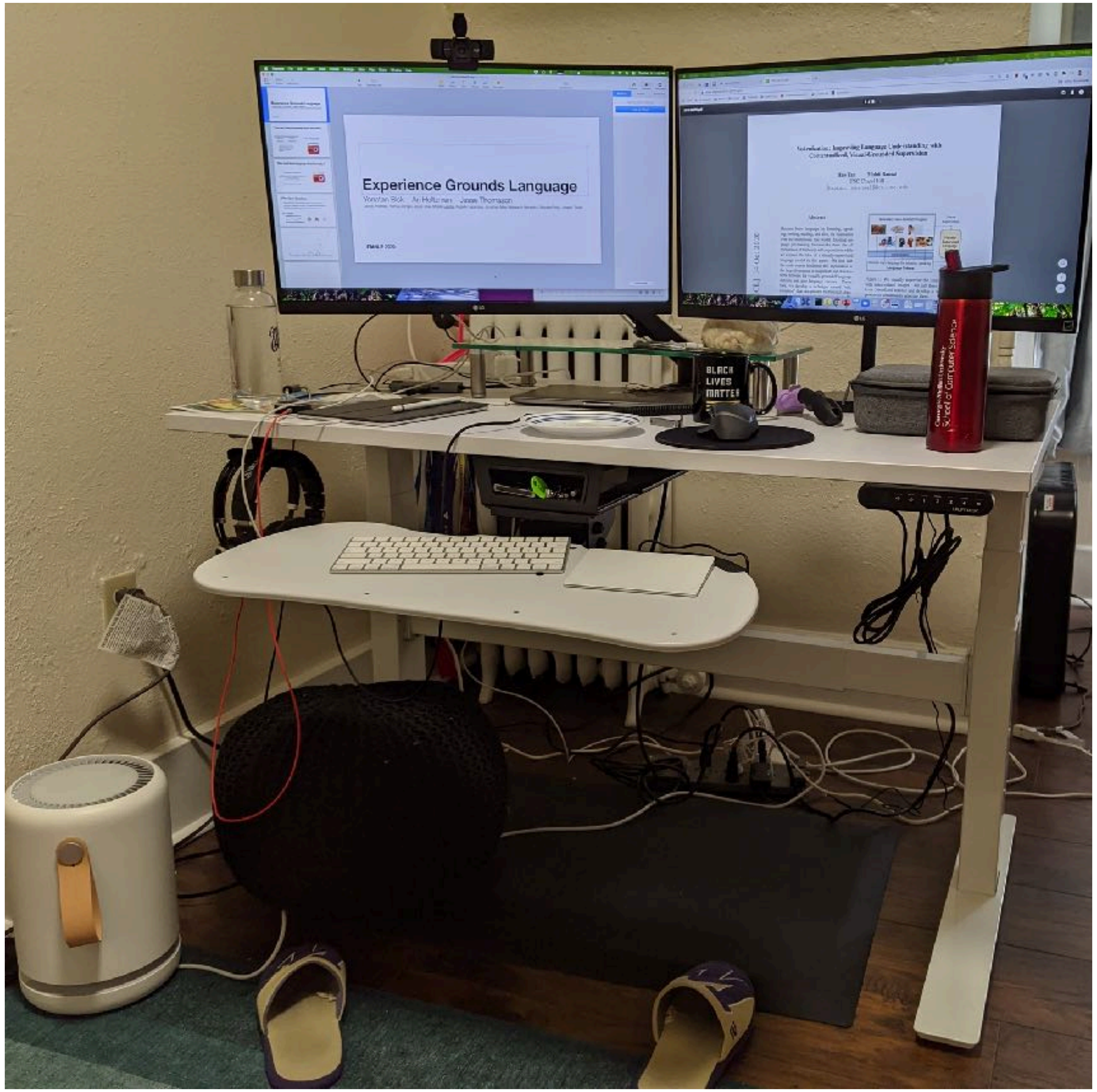


Is it every unary relationship?

Water bottles are clear
Water bottles are metallic
Mugs have words
Water bottles have words
…

Is it every pairwise relationship?

WS3 — How much “knowledge” is in an image?



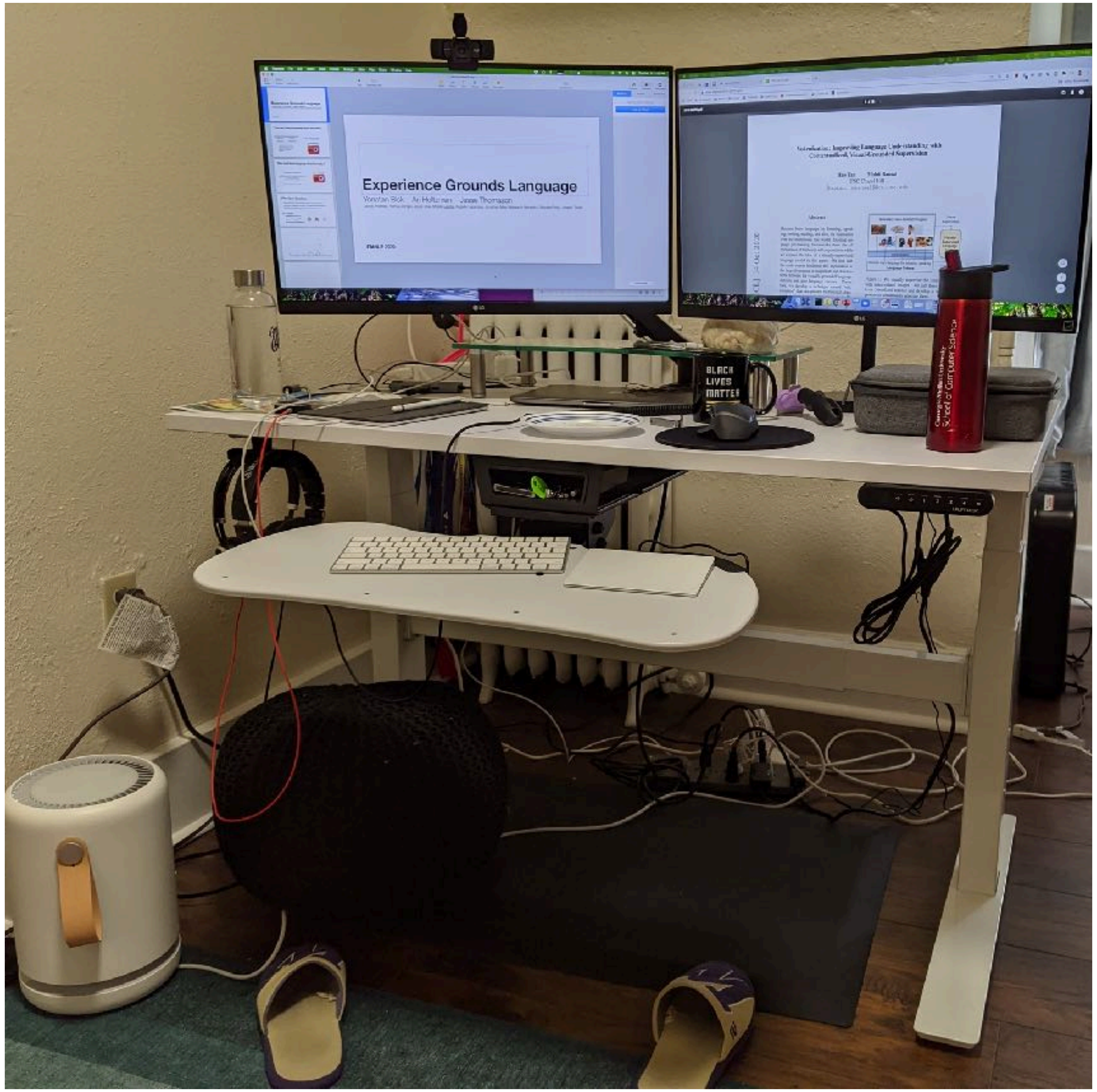
Is it every unary relationship?

Water bottles are clear
Water bottles are metallic
Mugs have words
Water bottles have words
…

Is it every pairwise relationship?

Water bottles go on desks,
Slippers go on floors,
Poofs are smaller than desks
Poofs are larger than slippers
…

WS3 — How much “knowledge” is in an image?



Is it every unary relationship?

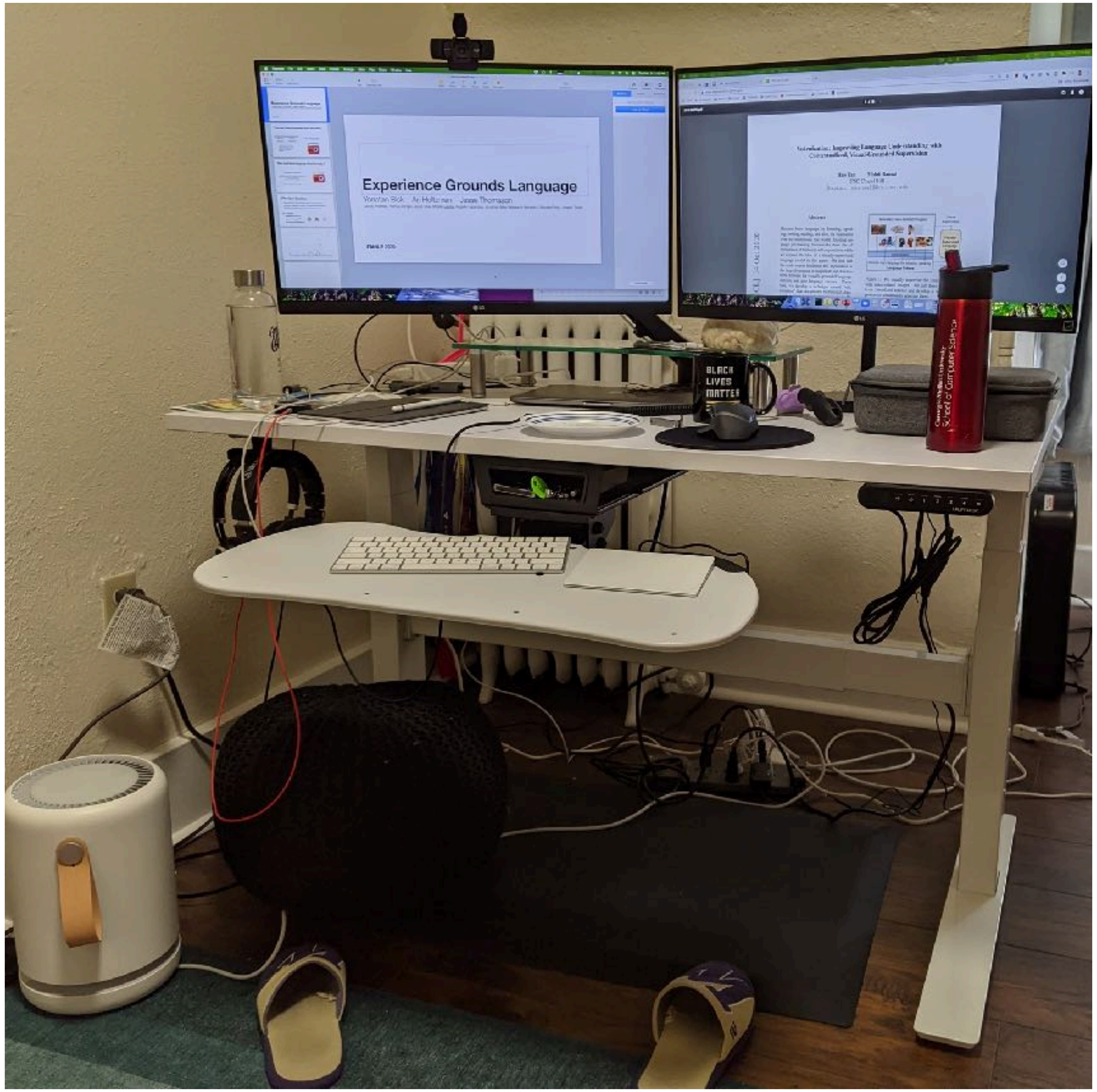
Water bottles are clear
Water bottles are metallic
Mugs have words
Water bottles have words
…

Is it every pairwise relationship?

Water bottles go on desks,
Slippers go on floors,
Poofs are smaller than desks
Poofs are larger than slippers
…

Is it that the privacy cover on the webcam is still up so I probably just got off a call indicating it’s during the work day and since the water bottle is full and there’s a small plate on the desk, can we infer that Yonatan is taking work meetings during lunch?

WS3 — How much “knowledge” is in an image?



Is it every unary relationship?

Water bottles are clear
Water bottles are metallic
Mugs have words
Water bottles have words
…

Is it every pairwise relationship?

Water bottles go on desks,
Slippers go on floors,
Poofs are smaller than desks
Poofs are larger than slippers
…

Is it that the privacy cover on the webcam is still up so I probably just got off a call indicating it’s during the work day and since the water bottle is full and there’s a small plate on the desk, can we infer that Yonatan is taking work meetings during lunch?

Yes

WS3 - Language beyond Text

Gestures

WS3 - Language beyond Text

Gestures

WS3 - Language beyond Text

Gestures

Facial expression

WS3 - Language beyond Text

Gestures

Facial expression

WS3 - Language beyond Text

Gestures

Facial expression

(Even just emojis 🙃)

WS3 - Language beyond Text

Gestures

Facial expression

(Even just emojis 🙃)

Intonation/Stress

WS3 - Language beyond Text

Gestures

Facial expression

(Even just emojis 🙃)

Intonation/Stress

I didn't take the test yesterday.
(Somebody else did.)
I ***didn't*** take the test yesterday.
(I did not take it.)
I didn't ***take*** the test yesterday.
(I did something else with it.)
I didn't take ***the*** test yesterday.
I didn't take the ***test*** yesterday.
I didn't take the test ***yesterday***.

https://en.wikipedia.org/wiki/Stress_(linguistics)

WS3 - Language beyond Text

Gestures

Facial expression

(Even just emojis 🙃)

Intonation/Stress

(eVeN JuSt fOnTs)

I didn't take the test yesterday.
(Somebody else did.)
I ***didn't*** take the test yesterday.
(I did not take it.)
I didn't ***take*** the test yesterday.
(I did something else with it.)
I didn't take ***the*** test yesterday.
I didn't take the ***test*** yesterday.
I didn't take the test ***yesterday***.

https://en.wikipedia.org/wiki/Stress_(linguistics)

WS4 — How much “knowledge” is in an environment?

https://www.youtube.com/watch?v=HgjpMtRpNrM

WS4 — How much “knowledge” is in an environment?

https://www.youtube.com/watch?v=HgjpMtRpNrM

WS4 — How much “knowledge” is in an environment?

https://www.youtube.com/watch?v=HgjpMtRpNrM

Is it every object affordance?

WS4 — How much “knowledge” is in an environment?

https://www.youtube.com/watch?v=HgjpMtRpNrM

Is it every object affordance?

Blocks can stack.
Well, some blocks can stack.
Wheels can roll.
Wheels experience friction.
…

WS4 — How much “knowledge” is in an environment?

https://www.youtube.com/watch?v=HgjpMtRpNrM

Is it every object affordance?

Blocks can stack.
Well, some blocks can stack.
Wheels can roll.
Wheels experience friction.
…

Is it every possible state or transition?

WS4 — How much “knowledge” is in an environment?

https://www.youtube.com/watch?v=HgjpMtRpNrM

Is it every object affordance?

Blocks can stack.
Well, some blocks can stack.
Wheels can roll.
Wheels experience friction.
…

Is it every possible state or transition?

Car can have 3 pink block on top.
Car can have one green on top.
Car can slide forward or backward.
Top pink block can be removed.
…

WS4 — How much “knowledge” is in an environment?

https://www.youtube.com/watch?v=HgjpMtRpNrM

Is it every object affordance?

Blocks can stack.
Well, some blocks can stack.
Wheels can roll.
Wheels experience friction.
…

Is it every possible state or transition?

Car can have 3 pink block on top.
Car can have one green on top.
Car can slide forward or backward.
Top pink block can be removed.
…

It starts to seem silly to even imagining using language to enumerate the possibilities of an environment, even a tabletop with blocks. The affordances, states, and transitions we learn for planning are strikingly mundane to express verbally, even on something like WikiHow.

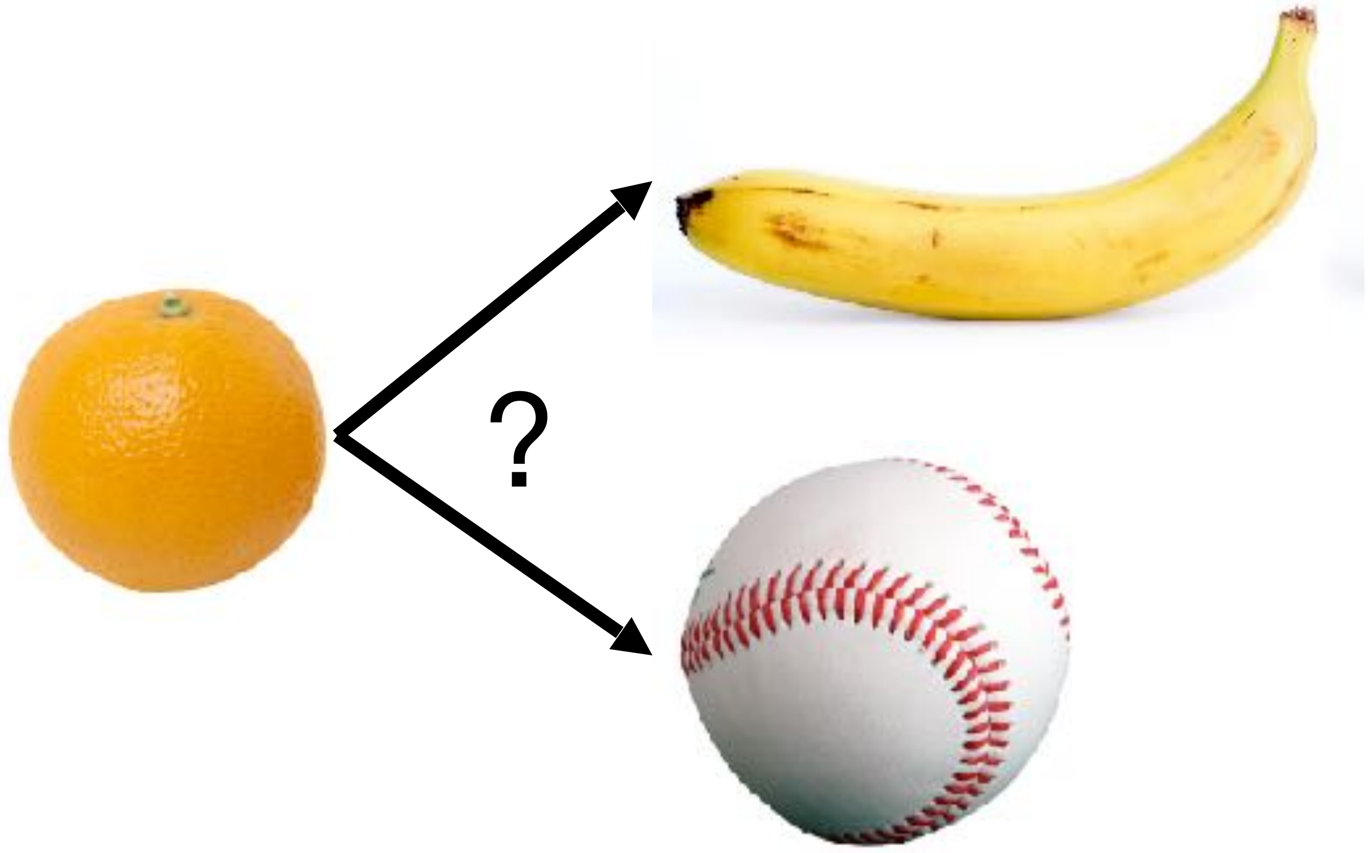
WS4 - The Language of the Physical World

WS4 - The Language of the Physical World

Object Representations

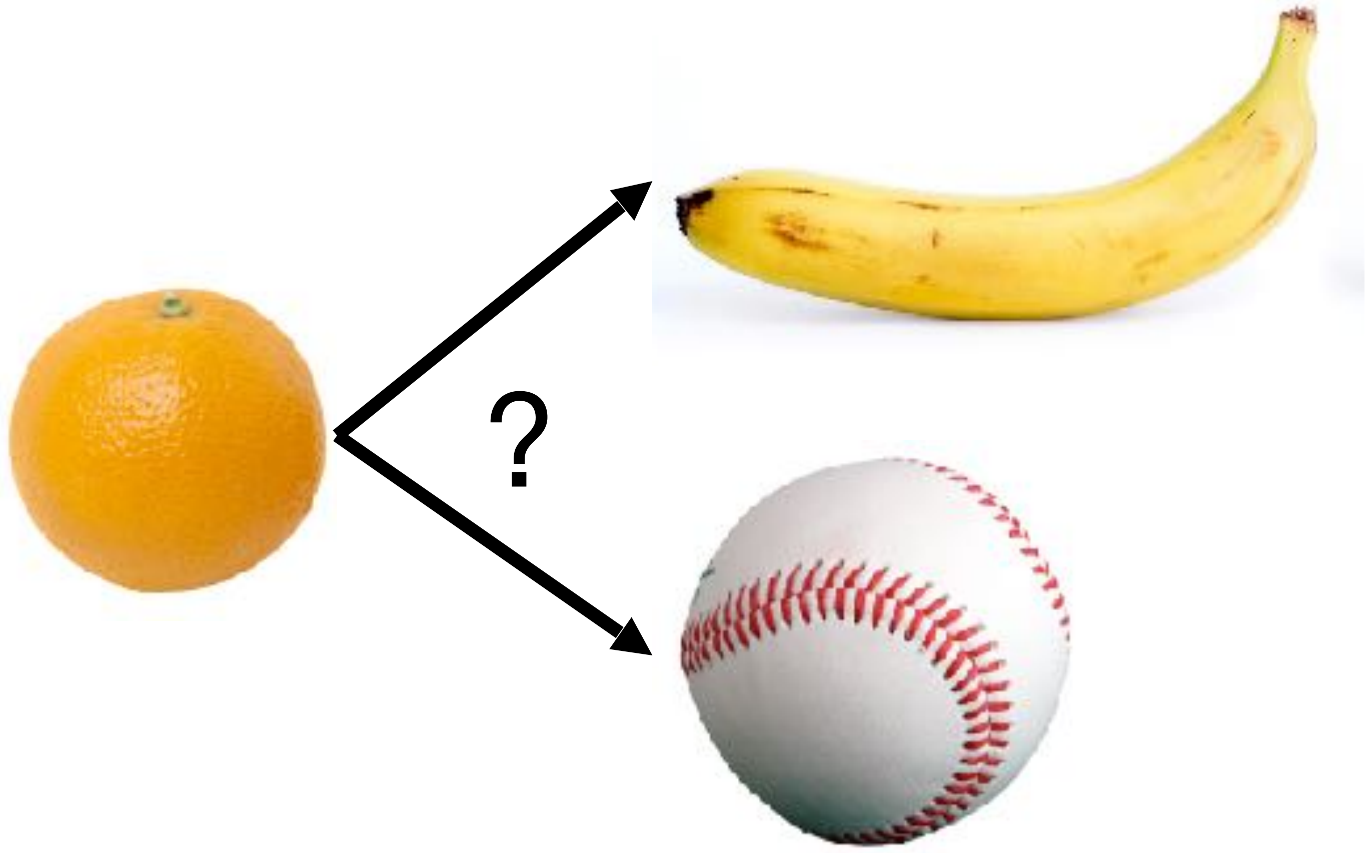
WS4 - The Language of the Physical World

Object Representations



WS4 - The Language of the Physical World

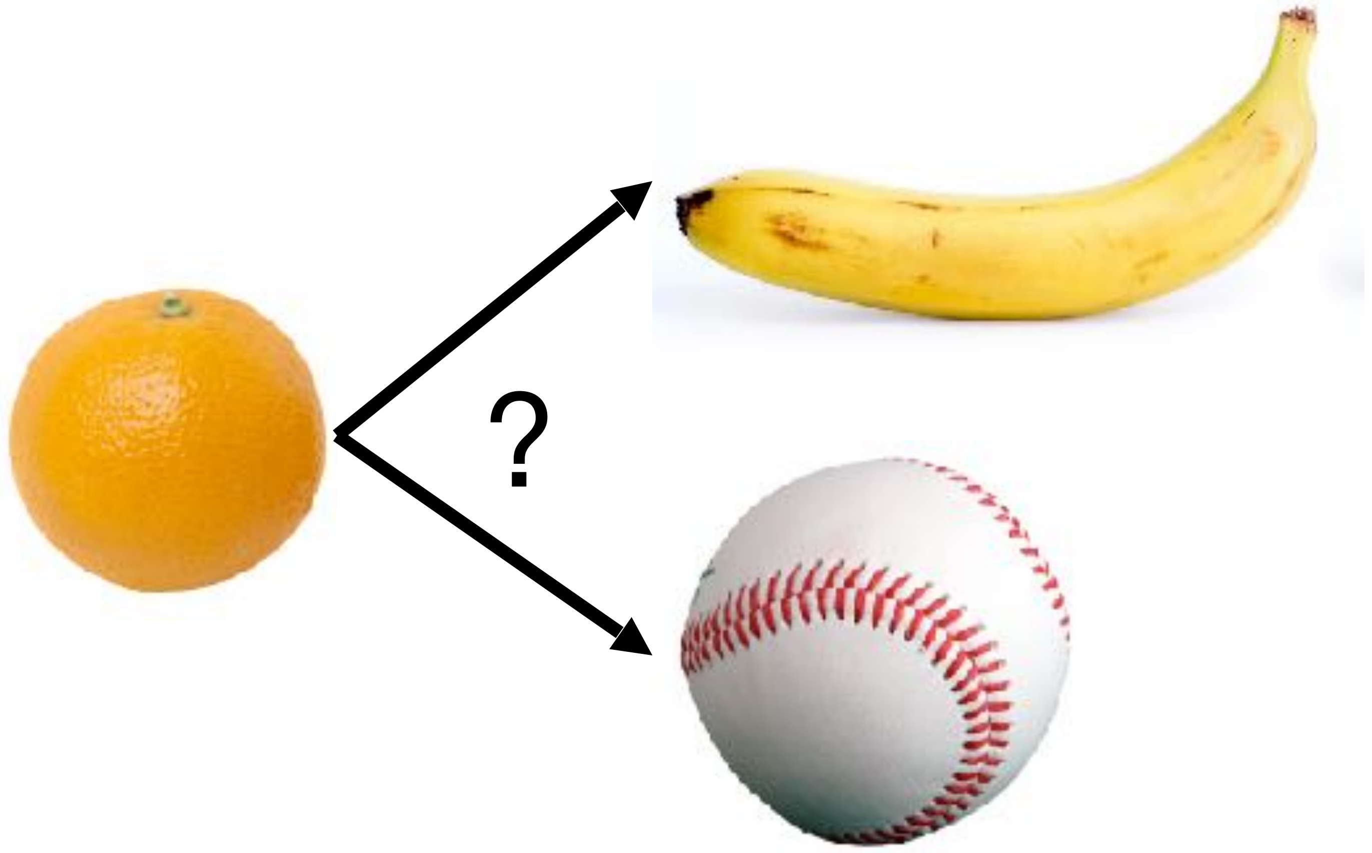
Object Representations



WS4 - The Language of the Physical World

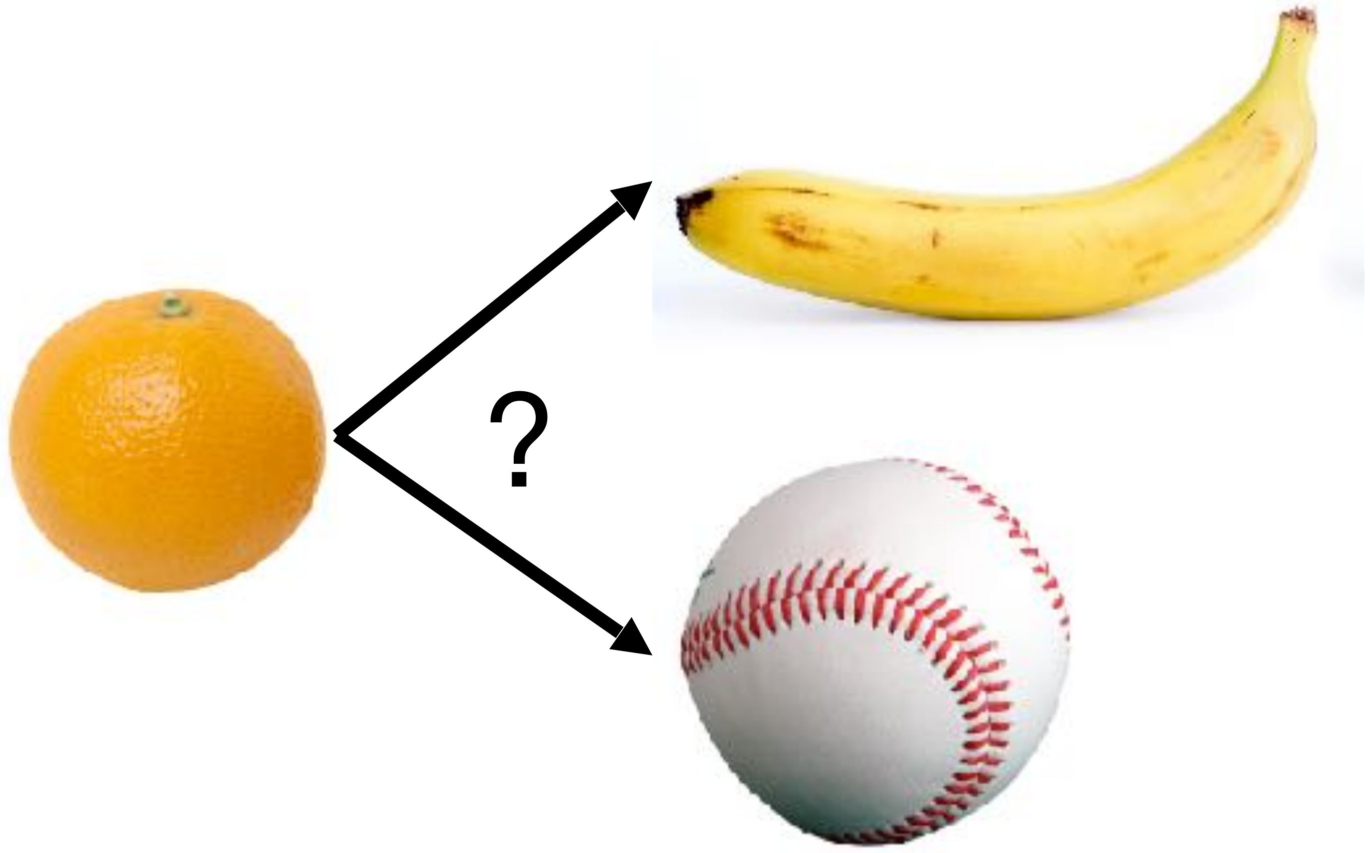
Object Representations

Physical Reasoning



WS4 - The Language of the Physical World

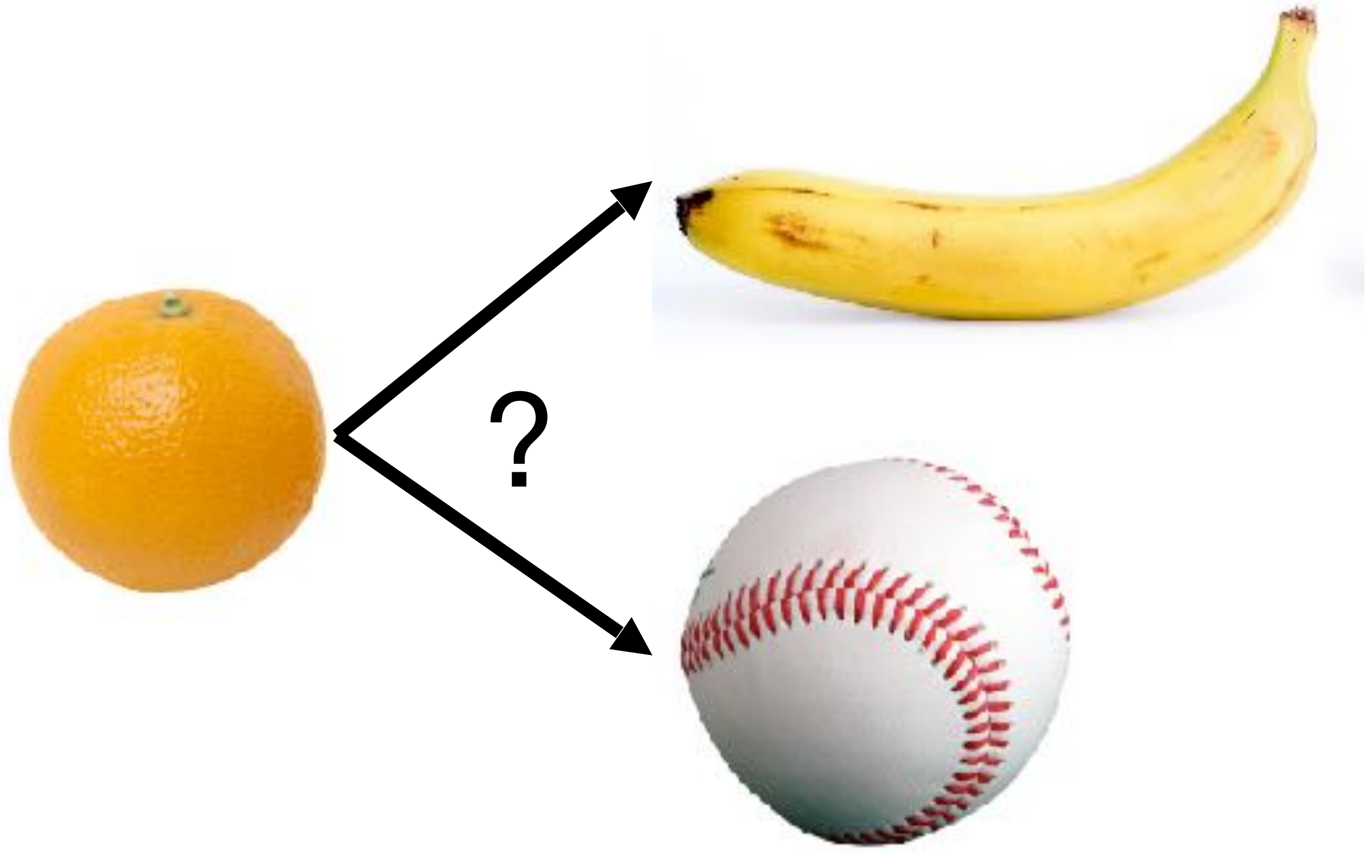
Object Representations



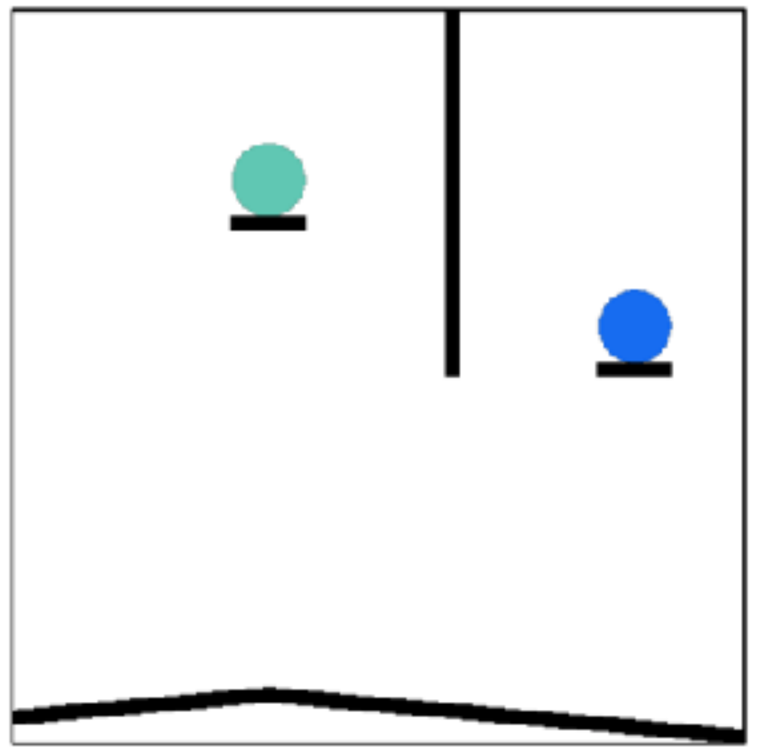
Physical Reasoning

WS4 - The Language of the Physical World

Object Representations



Physical Reasoning

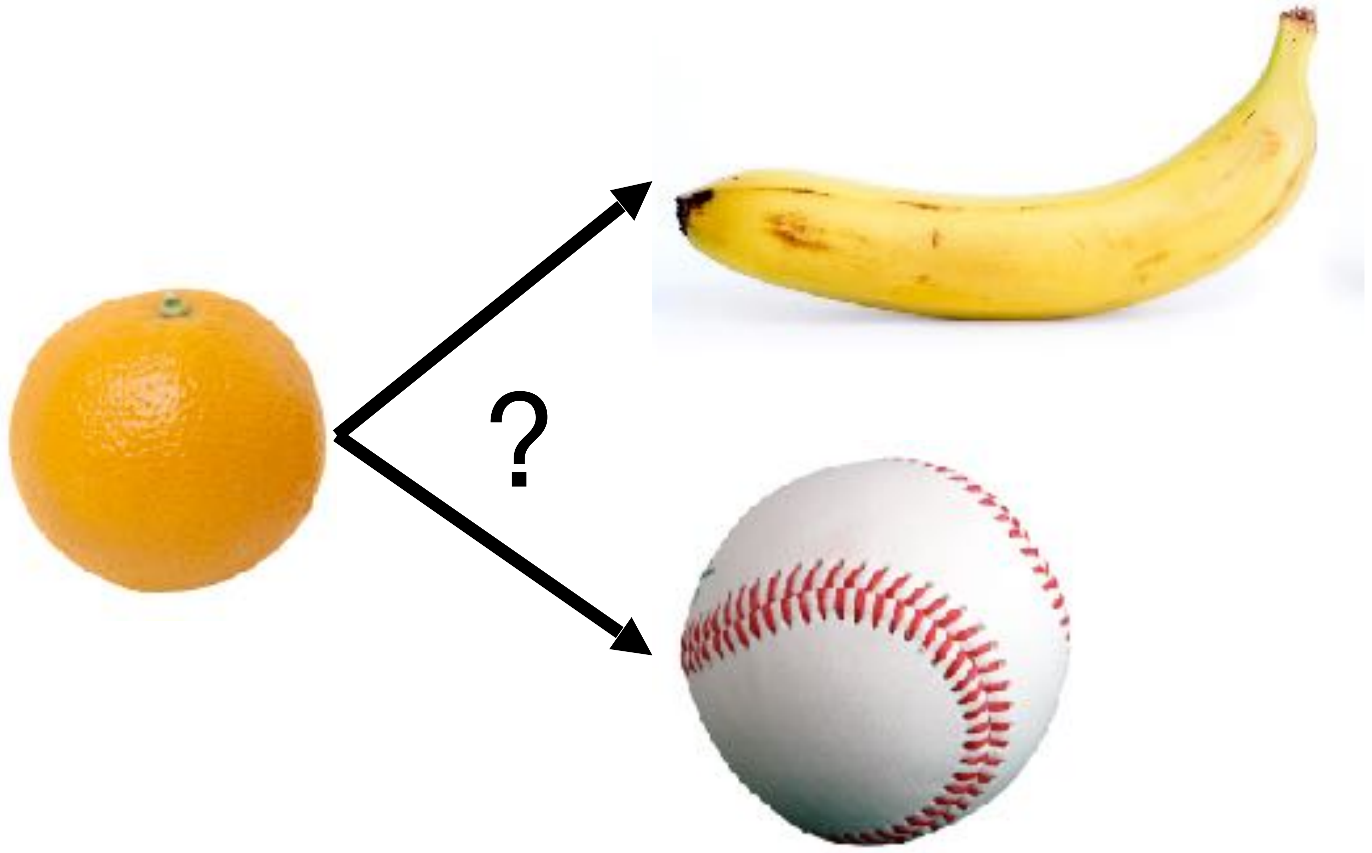


Make the green ball touch the blue ball

[Bakhtin et al., NeurIPS'19]

WS4 - The Language of the Physical World

Object Representations

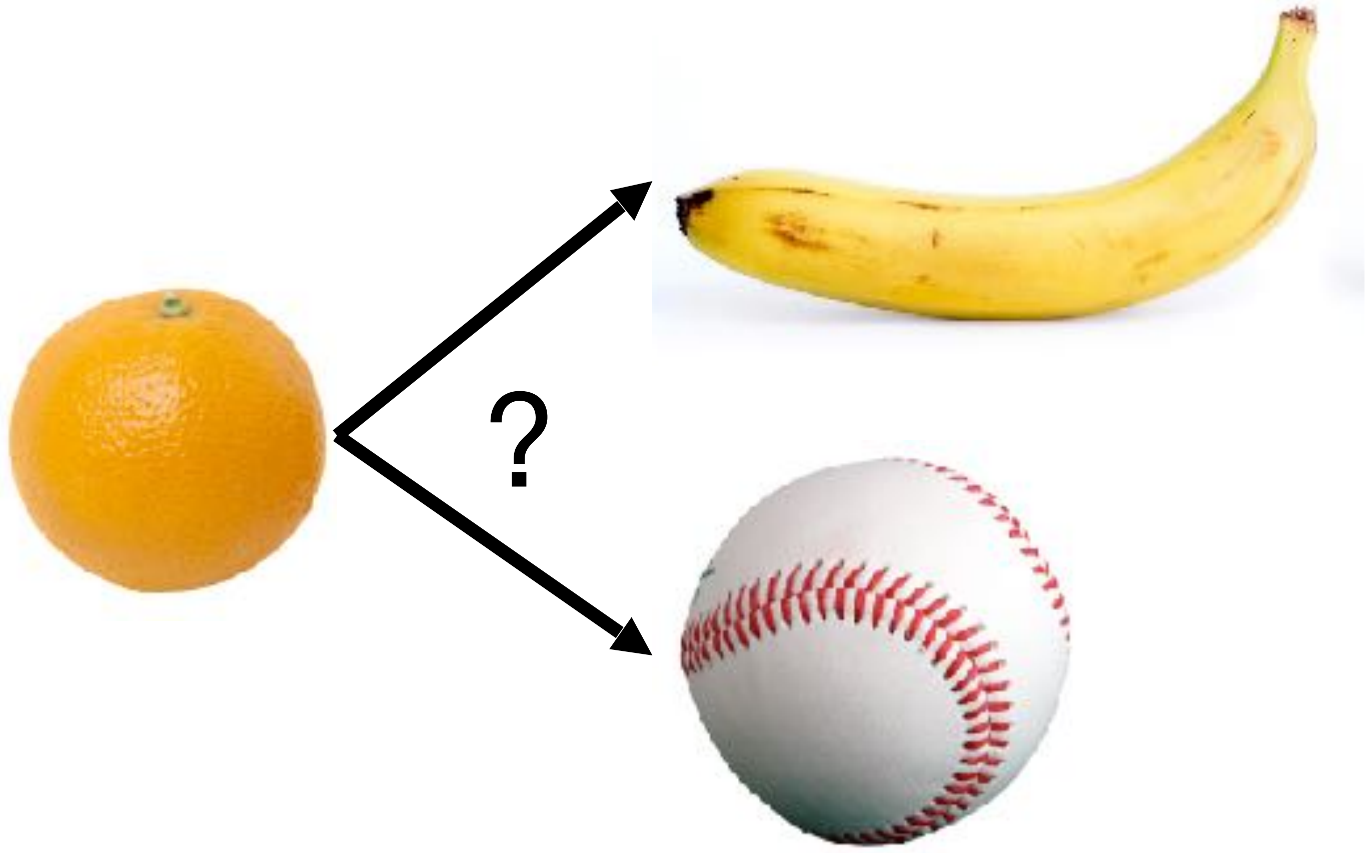


Physical Reasoning

A distant concern
A real moonshot
Thick tension
Heavy emotions

WS4 - The Language of the Physical World

Object Representations



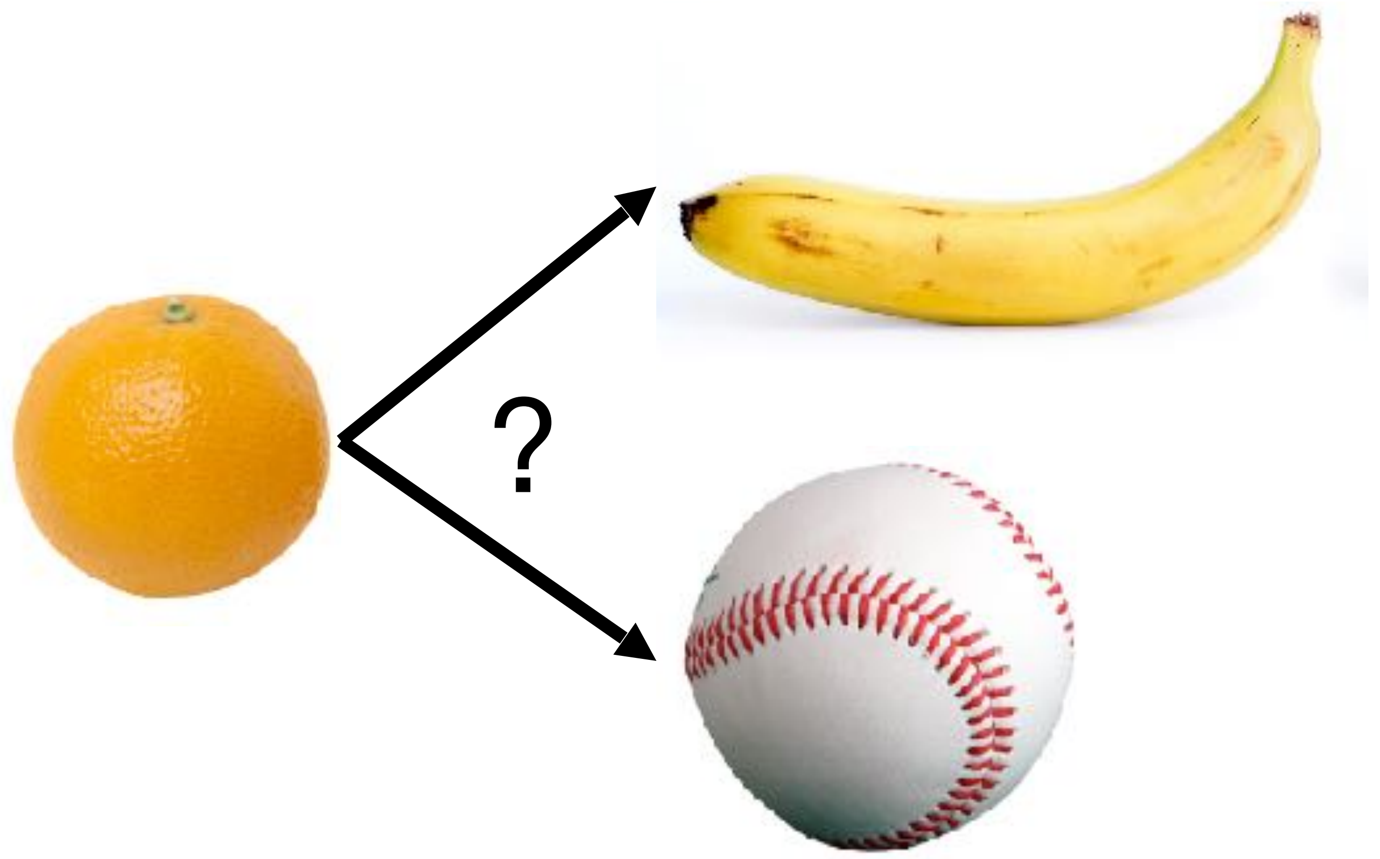
Physical Reasoning

A distant concern
A real moonshot
Thick tension
Heavy emotions

Affordances and Plans

WS4 - The Language of the Physical World

Object Representations



Physical Reasoning

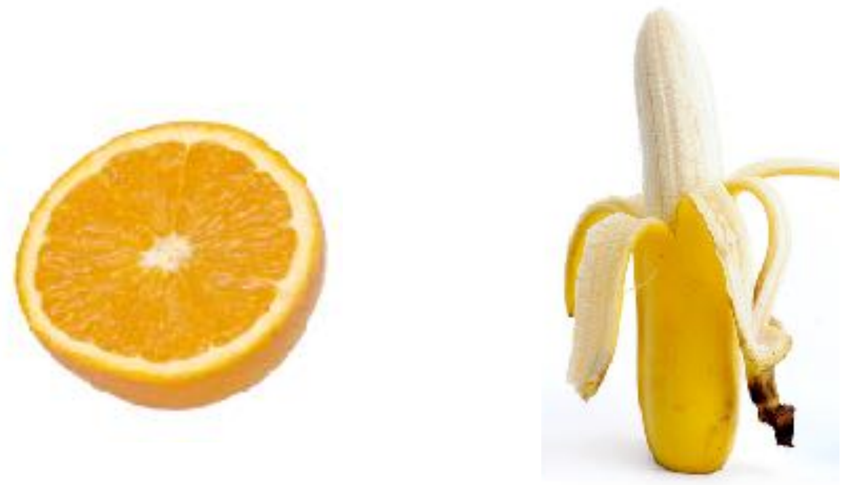
A distant concern

A real moonshot

Thick tension

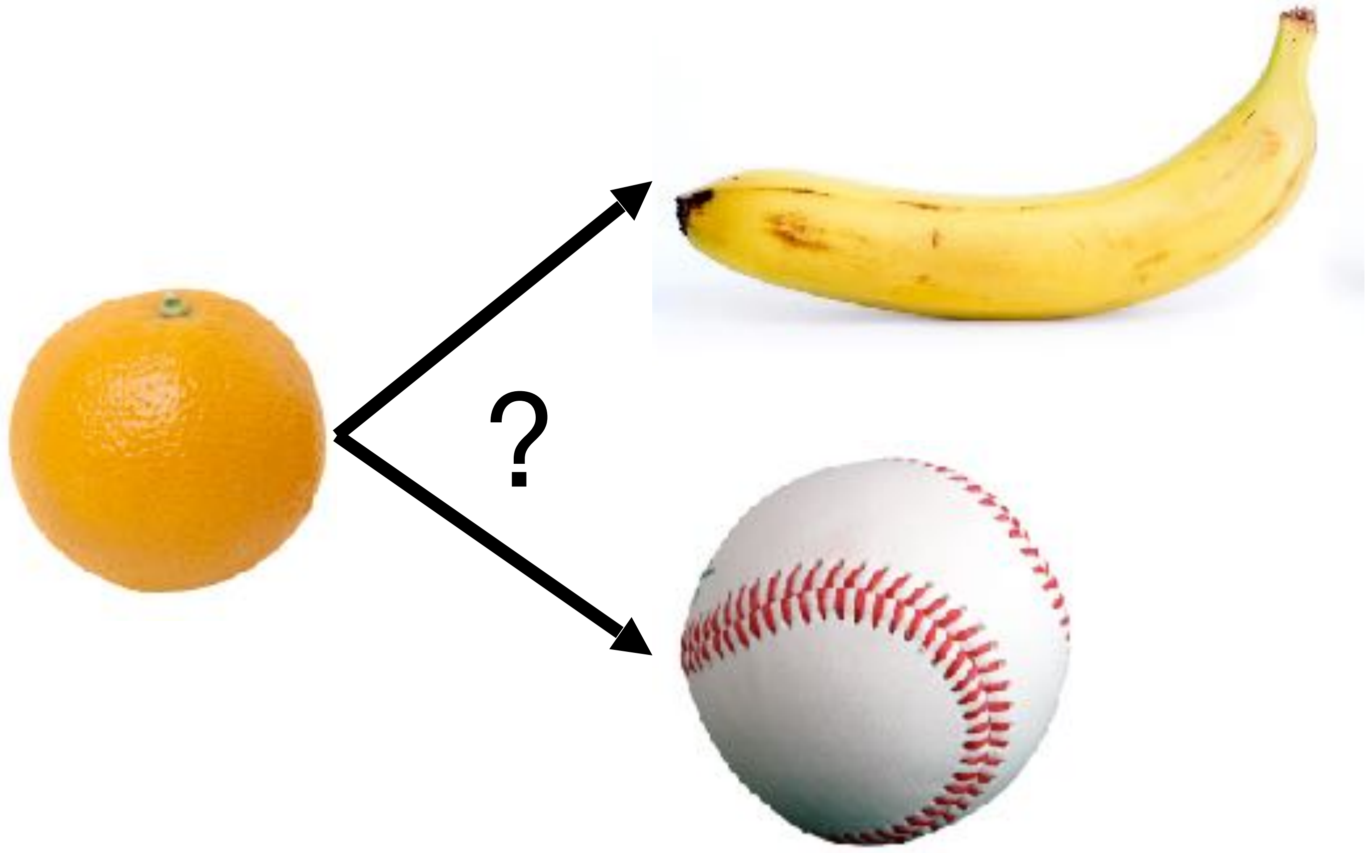
Heavy emotions

Affordances and Plans



WS4 - The Language of the Physical World

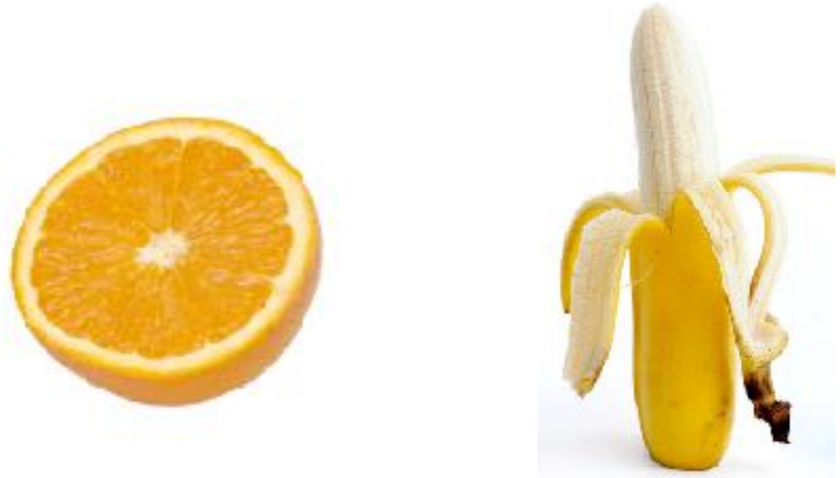
Object Representations

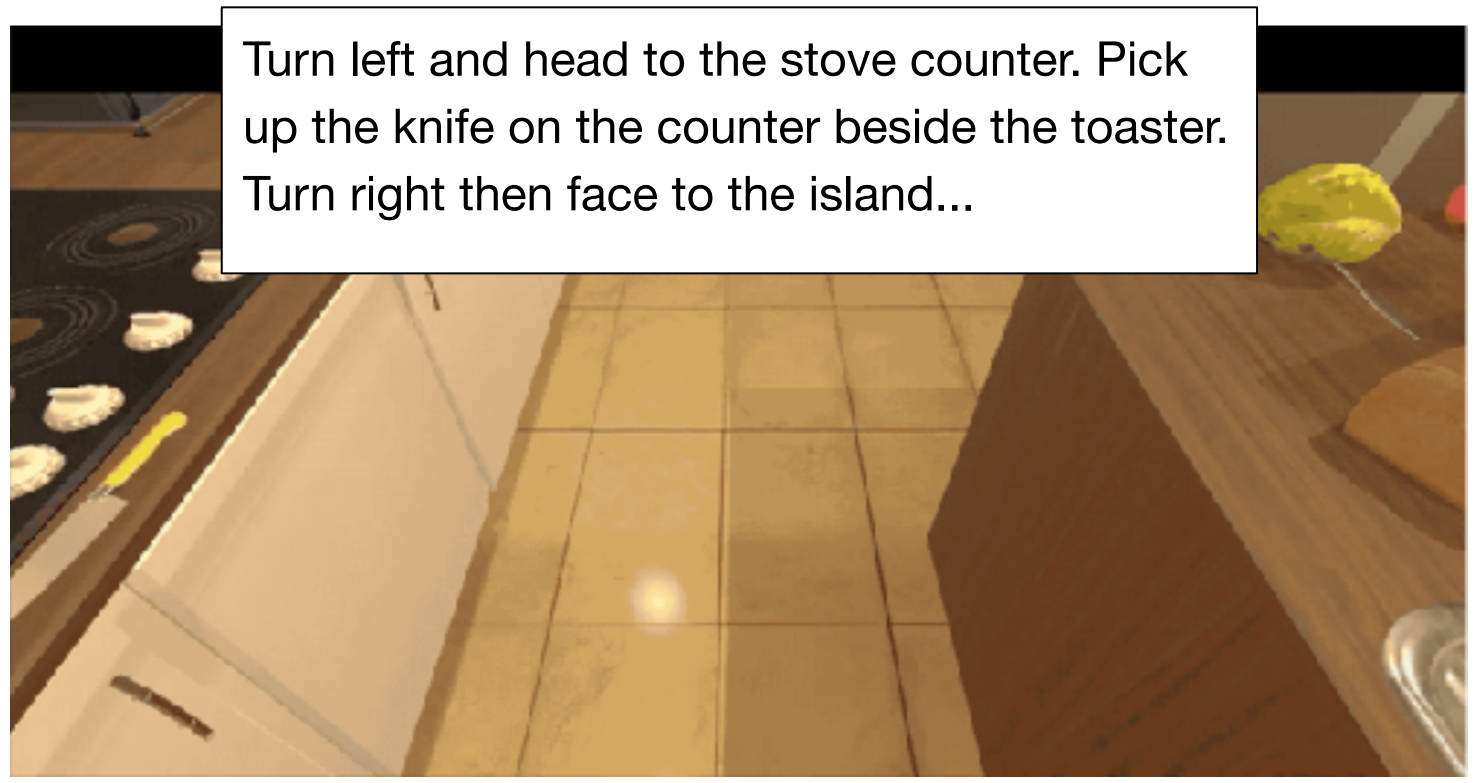


Physical Reasoning

A distant concern
A real moonshot
Thick tension
Heavy emotions

Affordances and Plans



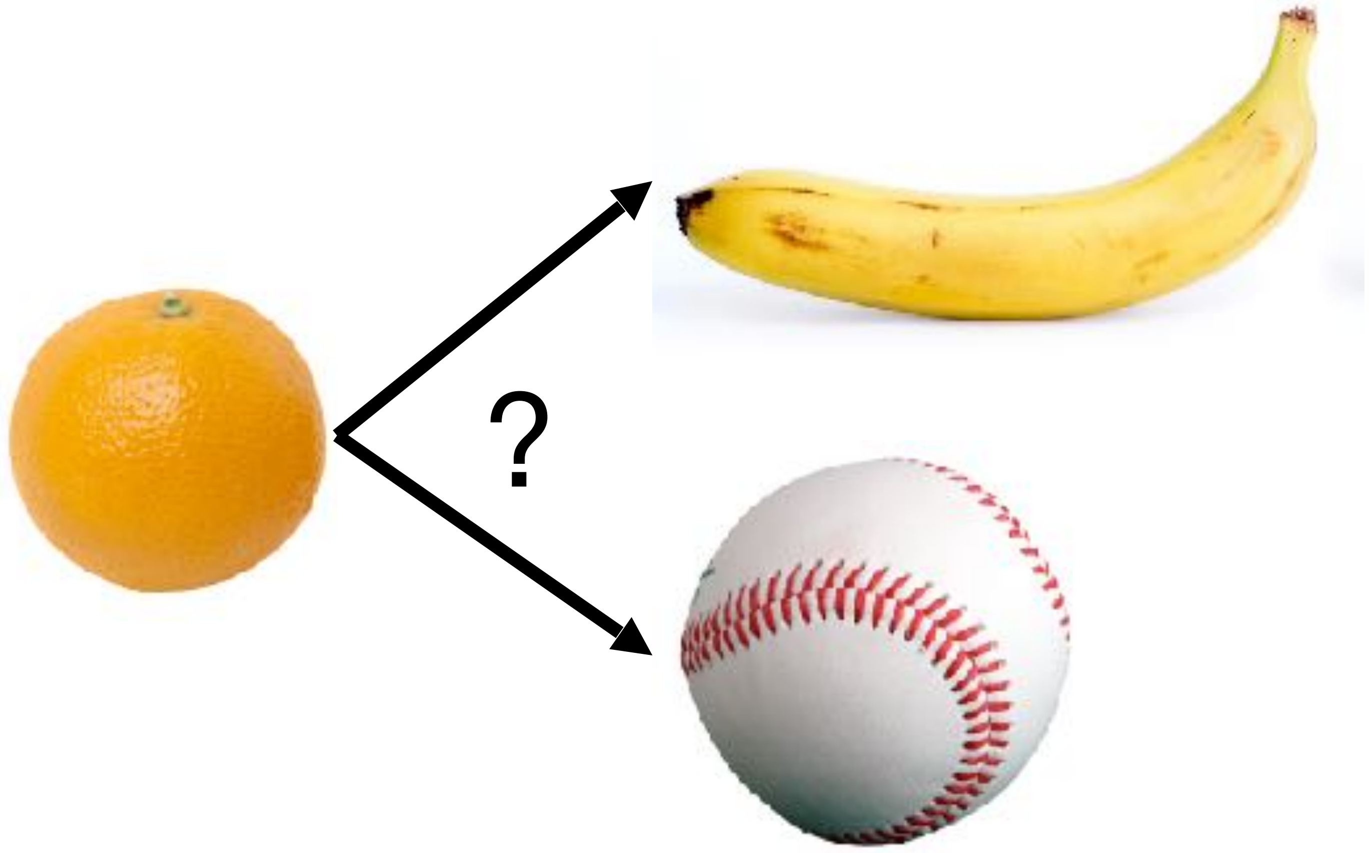


Turn left and head to the stove counter. Pick up the knife on the counter beside the toaster. Turn right then face to the island...

[Shridhar et al., CVPR'20]

WS4 - The Language of the Physical World

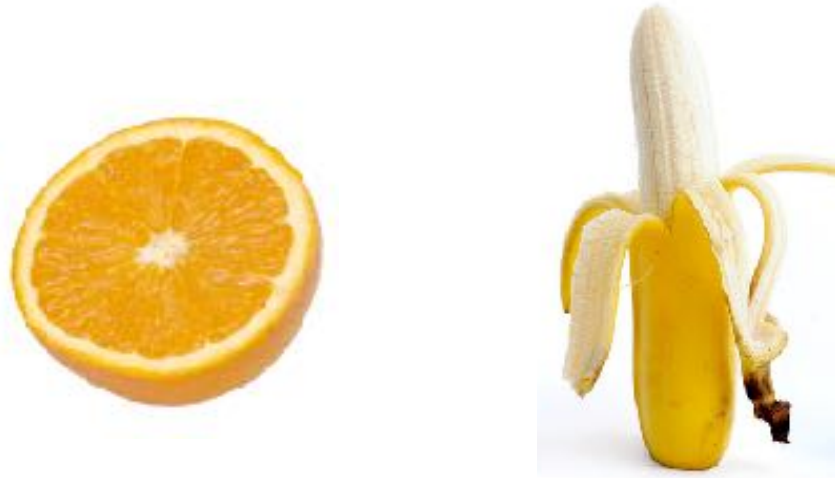
Object Representations



Physical Reasoning

A distant concern
A real moonshot
Thick tension
Heavy emotions

Affordances and Plans

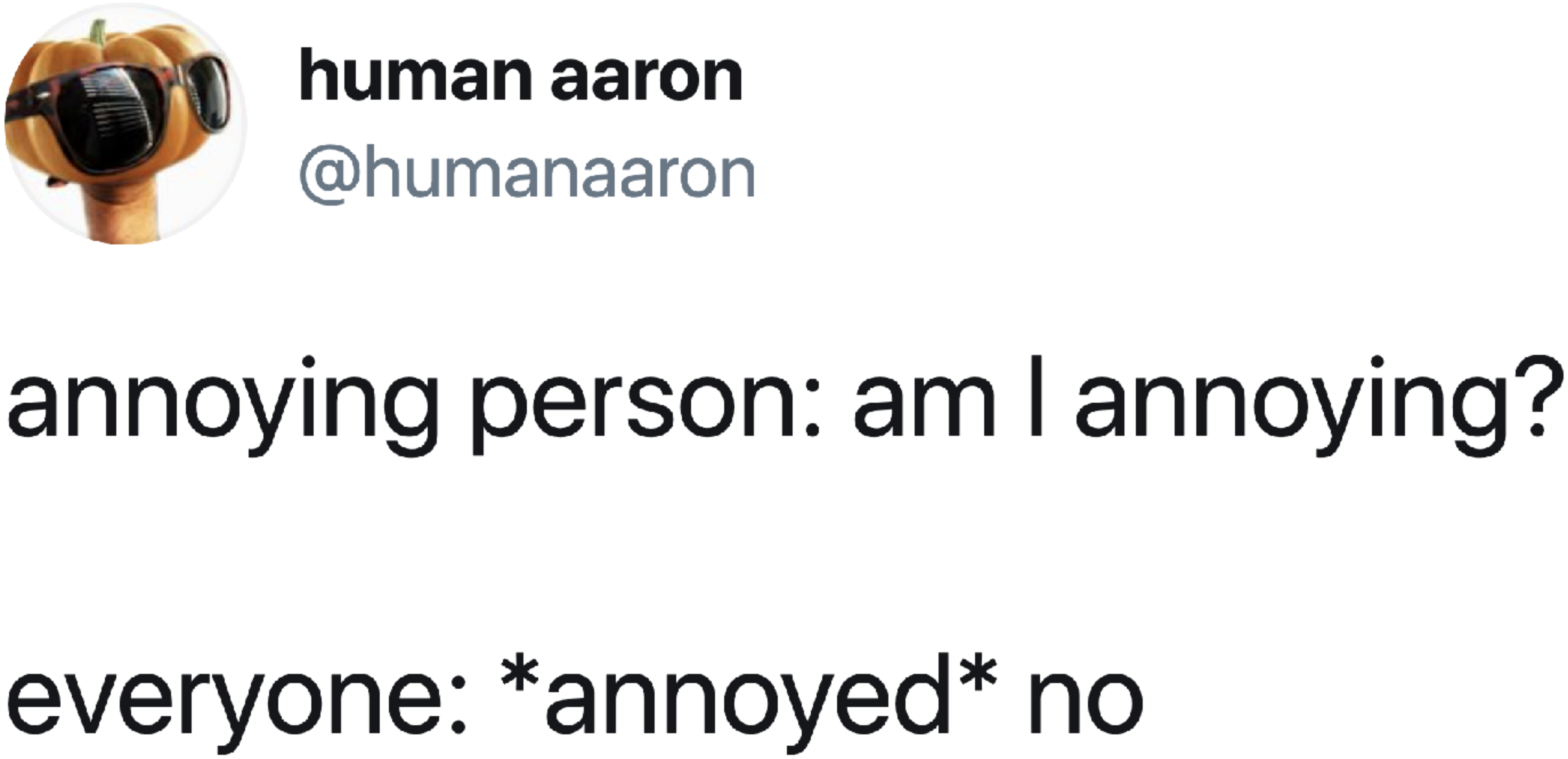


Should I hammer a nail with
1) a screwdriver or
2) a rock?

[Bisk et al., AAAI'20]

WS5 — The Social World

WS5 — The Social World

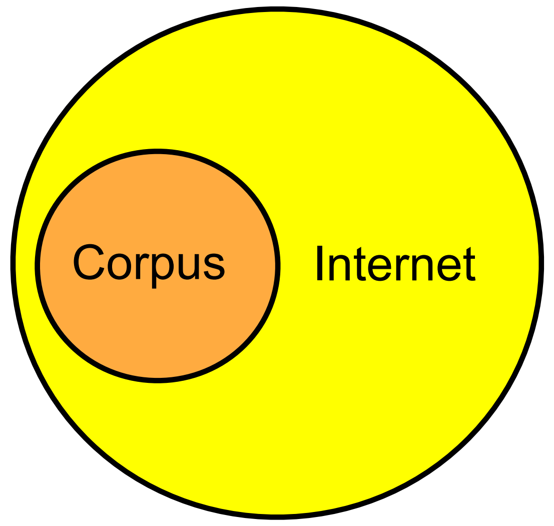


human aaron
@humanaaron

annoying person: am I annoying?

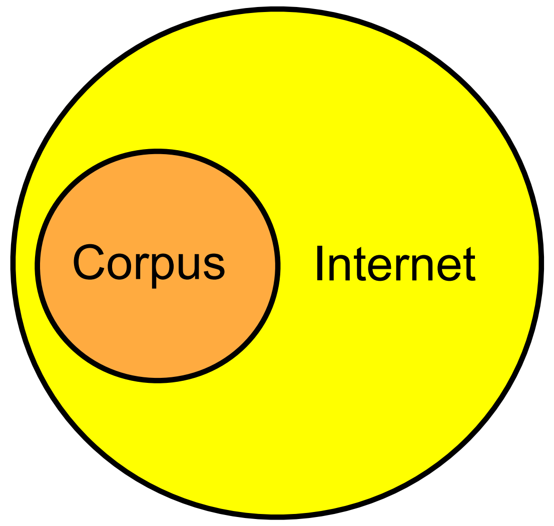
everyone: *annoyed* no

What delimits a World Scope?

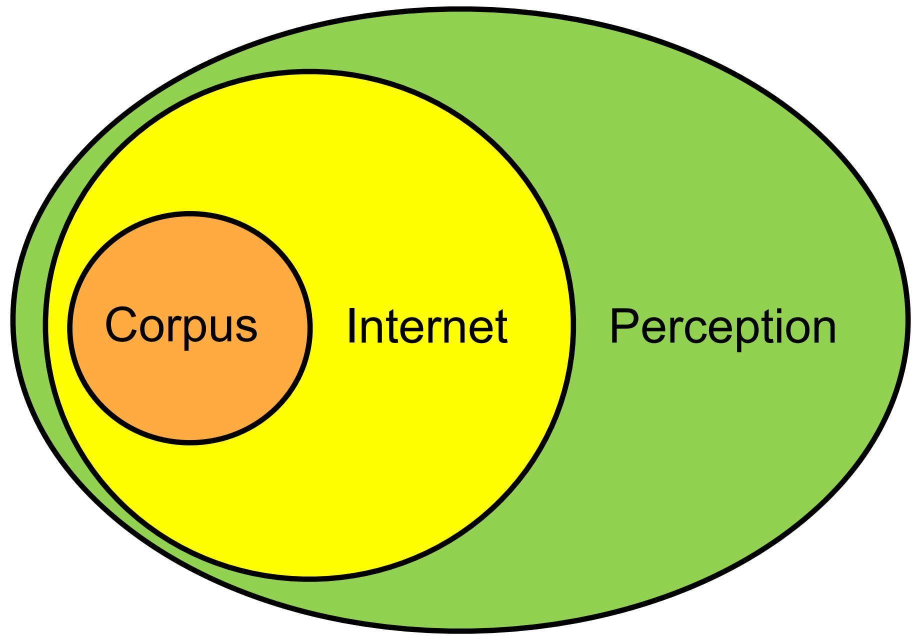


What delimits a World Scope?

You can’t learn language ...



What delimits a World Scope?

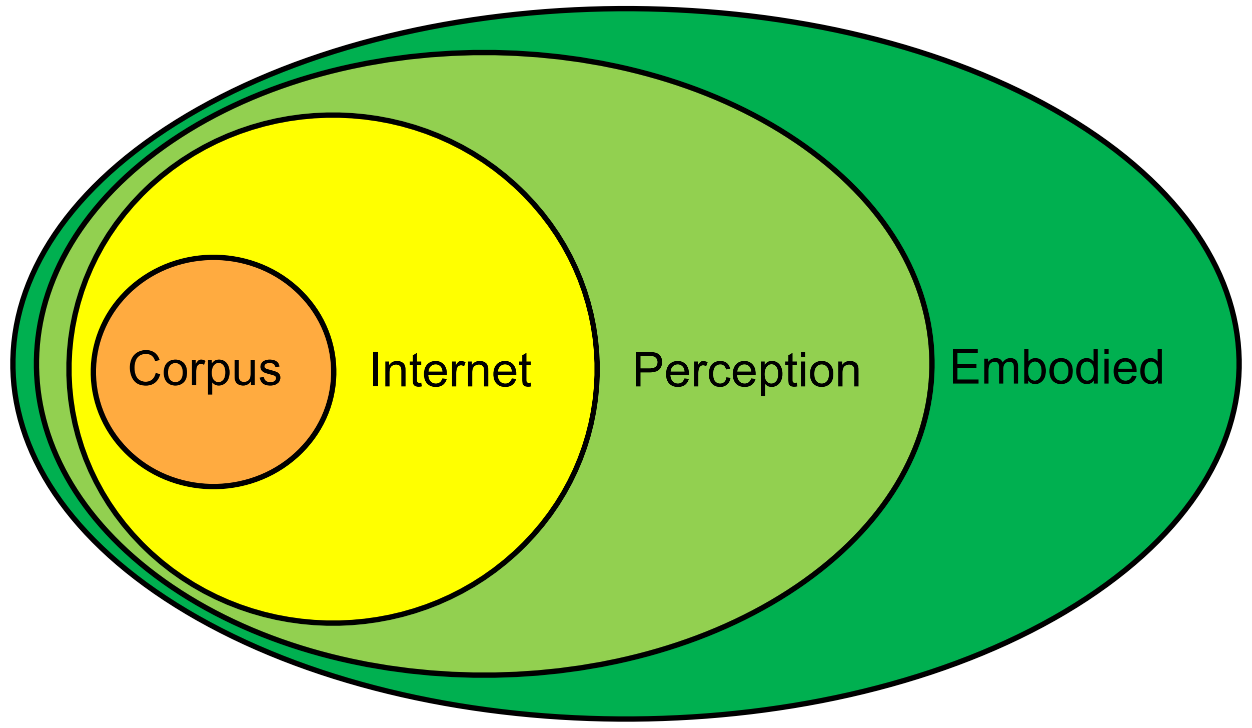


You can't learn language ...

... from the radio (Internet). WS2 ⊂ WS3

A task learner cannot be said to be in WS3 if it can succeed without perception (e.g., visual, auditory).

What delimits a World Scope?



You can’t learn language ...

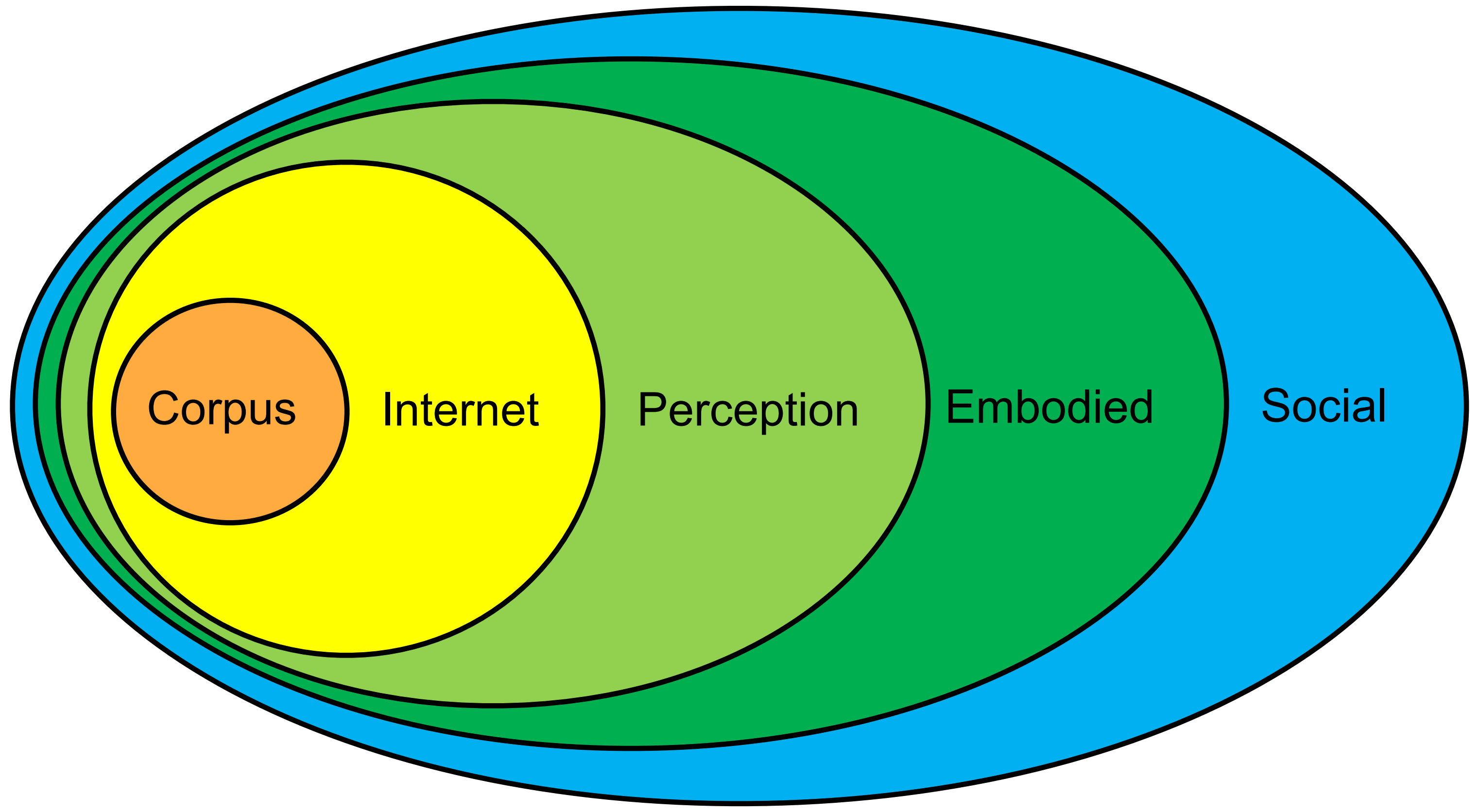
... from the radio (Internet). WS2 ⊂ WS3

A task learner cannot be said to be in WS3 if it can succeed without perception (e.g., visual, auditory).

... from a television. WS3 ⊂ WS4

A task learner cannot be said to be in WS4 if the space of its world actions and consequences can be enumerated.

What delimits a World Scope?



You can’t learn language ...

... from the radio (Internet). WS2 ⊂ WS3

A task learner cannot be said to be in WS3 if it can succeed without perception (e.g., visual, auditory).

... from a television. WS3 ⊂ WS4

A task learner cannot be said to be in WS4 if the space of its world actions and consequences can be enumerated.

... by yourself. WS4 ⊂ WS5

A task learner cannot be said to be in WS5 unless achieving its goals requires cooperating with a human in the loop.

What should we do about this?

Rethink every task

Rethink every task

Include phenomena from WS3-5 in your model’s universe …

Rethink every task

Include phenomena from WS3-5 in your model’s universe …

Will grounding make generation more controllable?   Yes.

Rethink every task

Include phenomena from WS3-5 in your model’s universe …

Will grounding make generation more controllable?   Yes.

Should model evaluation use humans-in-the-loop?   Yes.

Rethink every task

Include phenomena from WS3-5 in your model’s universe …

Will grounding make generation more controllable?   Yes.

Should model evaluation use humans-in-the-loop?   Yes.

Can MT systems use interaction for more targeted learning? Yes.

Rethink every task

Include phenomena from WS3-5 in your model's universe …

Will grounding make generation more controllable?   Yes.

Should model evaluation use humans-in-the-loop?   Yes.

Can MT systems use interaction for more targeted learning? Yes.

Can physical exploration improve knowledge of entailment?   Yes.

Rethink every task

Include phenomena from WS3-5 in your model’s universe …

Will grounding make generation more controllable?   Yes.

Should model evaluation use humans-in-the-loop?   Yes.

Can MT systems use interaction for more targeted learning? Yes.

Can physical exploration improve knowledge of entailment?   Yes.

Can perception resolve syntactic/anaphoric/discursive ambiguity?   Yes.

Rethink every task

Include phenomena from WS3-5 in your model’s universe …

Will grounding make generation more controllable?   Yes.

Should model evaluation use humans-in-the-loop?   Yes.

Can MT systems use interaction for more targeted learning? Yes.

Can physical exploration improve knowledge of entailment?   Yes.

Can perception resolve syntactic/anaphoric/discursive ambiguity?   Yes.

Can tests in a simulator capture distinctions that MC struggles with?   Yes.

Rethink every task

Include phenomena from WS3-5 in your model’s universe …

Will grounding make generation more controllable?   Yes.

Should model evaluation use humans-in-the-loop?   Yes.

Can MT systems use interaction for more targeted learning? Yes.

Can physical exploration improve knowledge of entailment?   Yes.

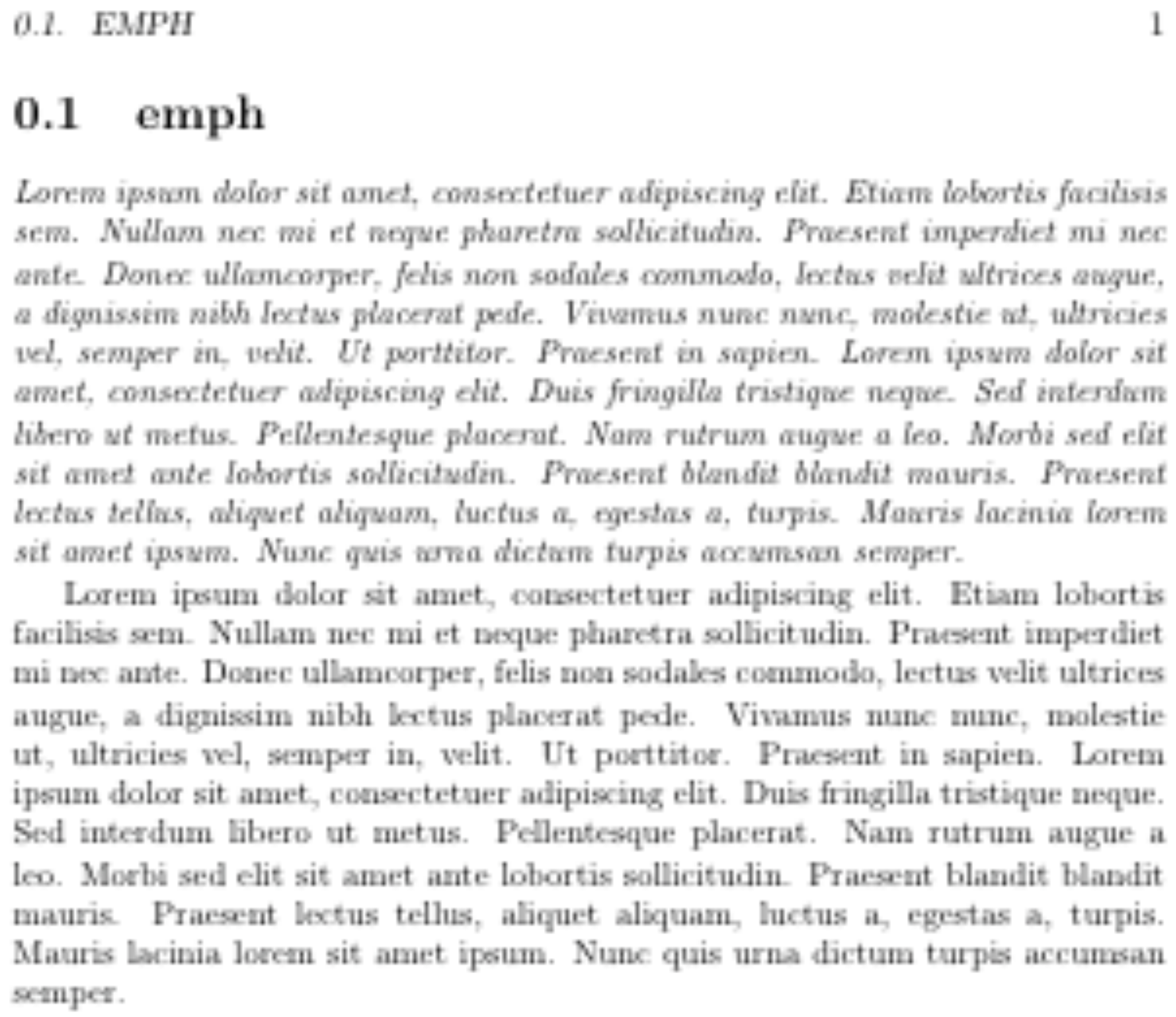
Can perception resolve syntactic/anaphoric/discursive ambiguity?   Yes.

Can tests in a simulator capture distinctions that MC struggles with?   Yes.

…and consider where signal for your task comes from!

Is knowledge about task X most richly encoded in:

Is knowledge about task X most richly encoded in:



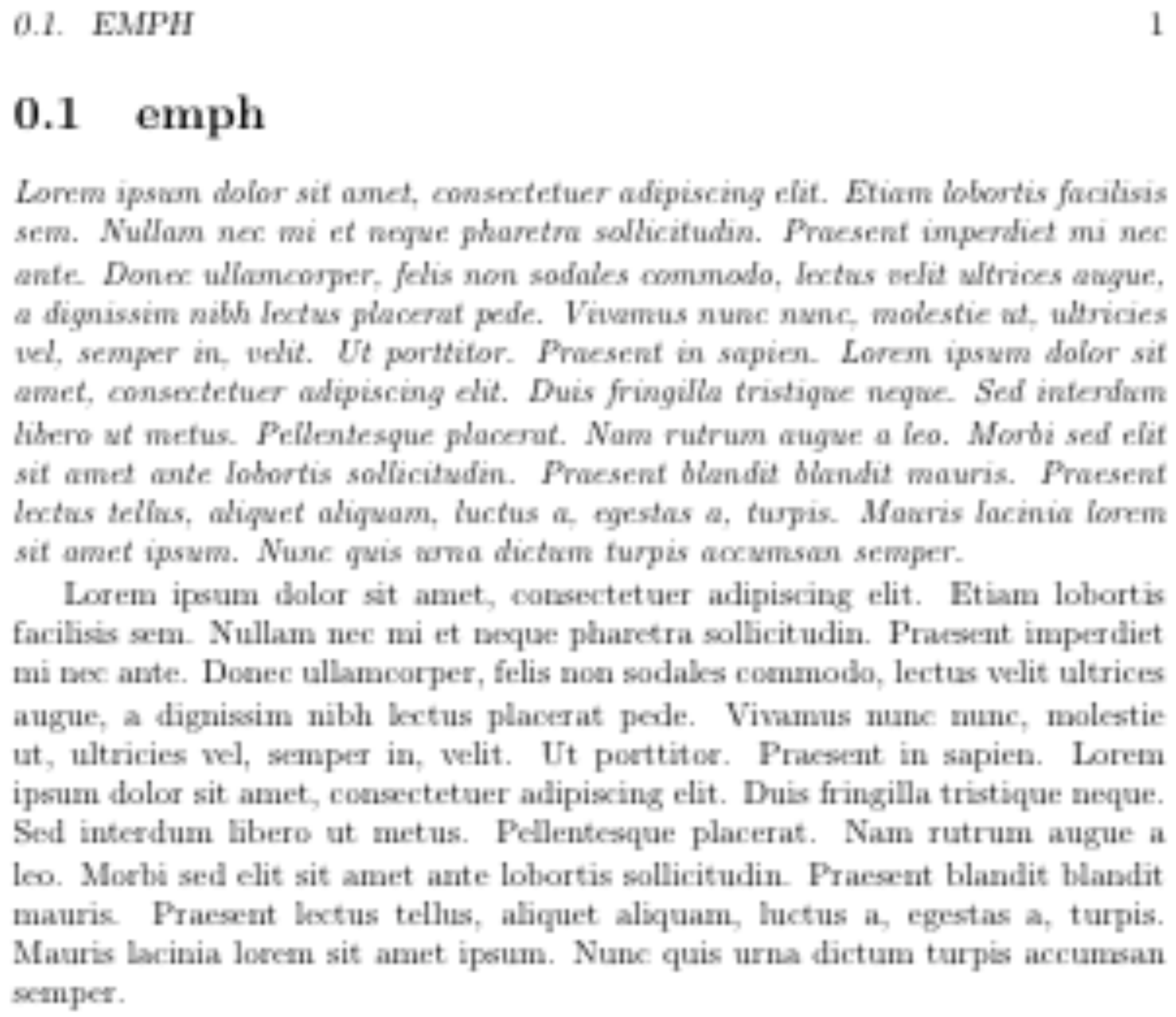
0.1. EMPH 1

0.1 emph

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Is knowledge about task X most richly encoded in:



0.1. EMPH 1

0.1 emph

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.