I don’t know what you mean semantics is hard: Challenges in evaluation of semantic phenomena

Ellie Pavlick
Department of Computer Science
Brown University

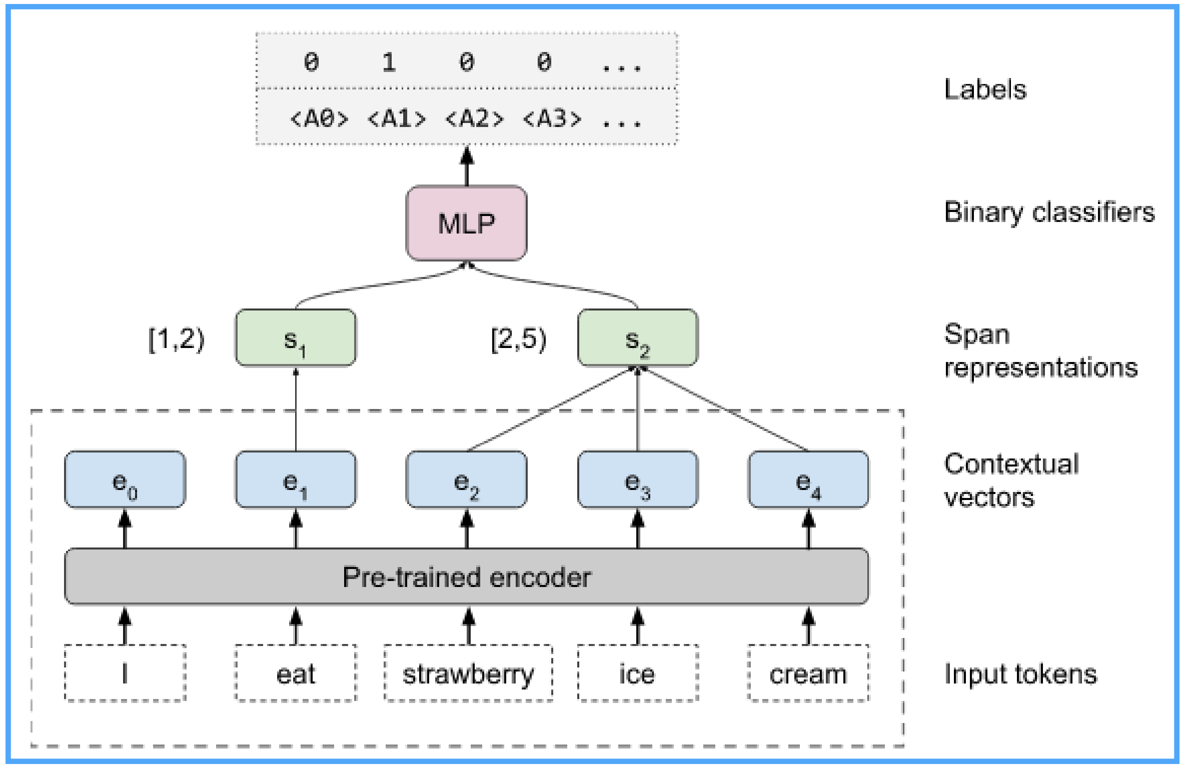
Past ~2 years:
What do deep LMs know about language?

Past ~2 years:
What do deep LMs know about language?

Probing Classifiers: What types of linguistic structures do representations encode?

Past ~2 years:
What do deep LMs know about language?

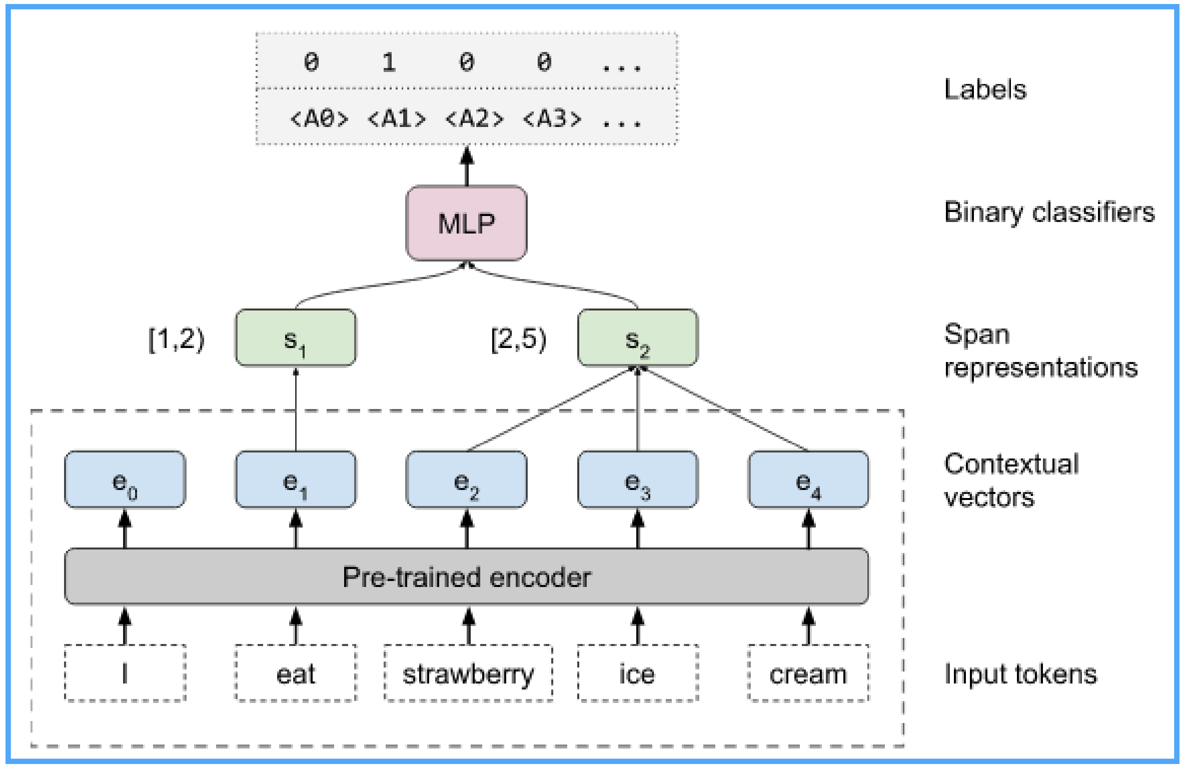
Probing Classifiers: What types of linguistic structures do representations encode?



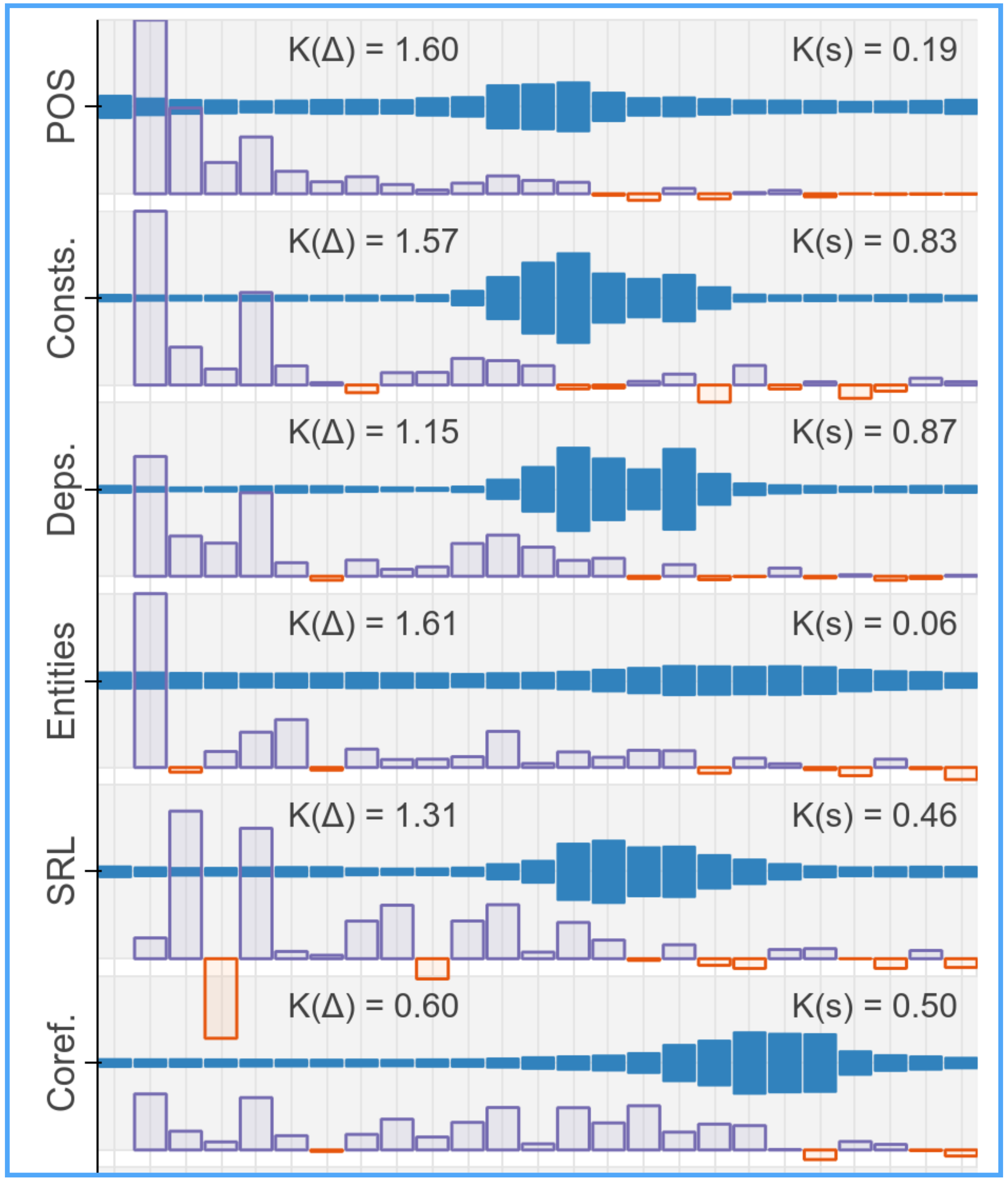
Tenney et al (ICLR 2018)

Past ~2 years:
What do deep LMs know about language?

Probing Classifiers: What types of linguistic structures do representations encode?



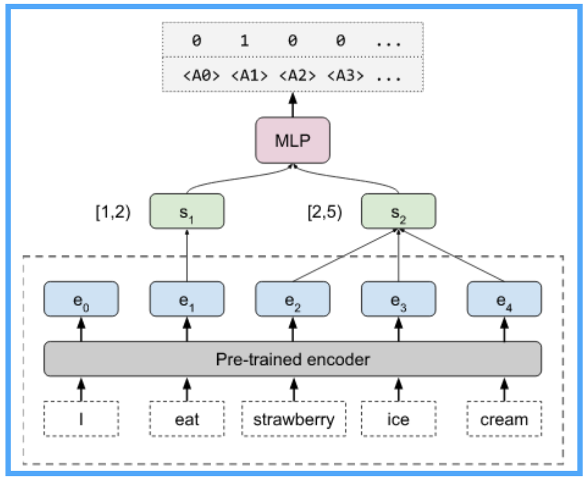
Tenney et al (ICLR 2018)



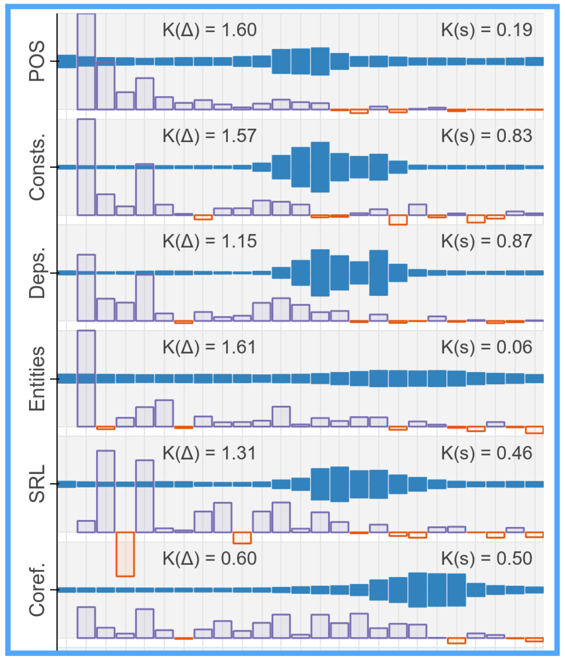
Tenney et al (ACL 2019)

Past ~2 years:
What do deep LMs know about language?

Probing Classifiers: What types of linguistic structures do representations encode?



Tenney et al (ICLR 2018)

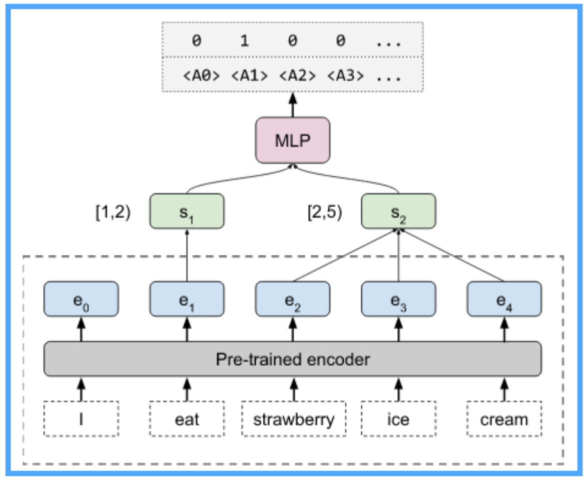


Tenney et al (ACL 2019)

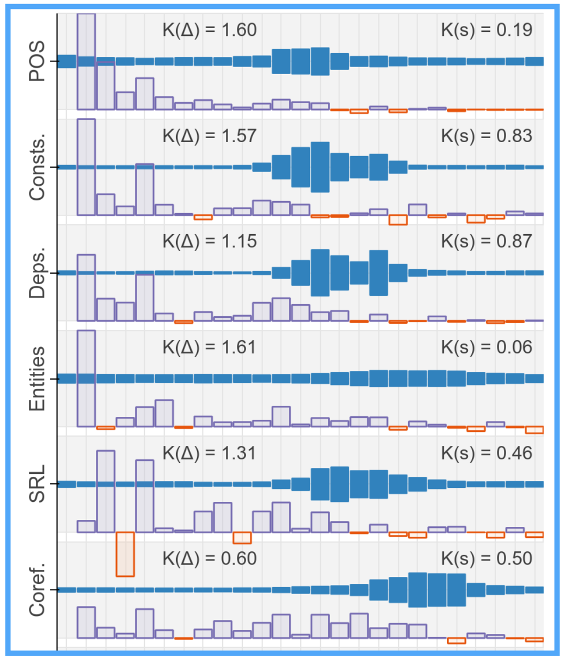
Past ~2 years:
What do deep LMs know about language?

Probing Classifiers: What types of linguistic structures do representations encode?

Challenge Tasks: How well do models perform on difficult “tail” events?

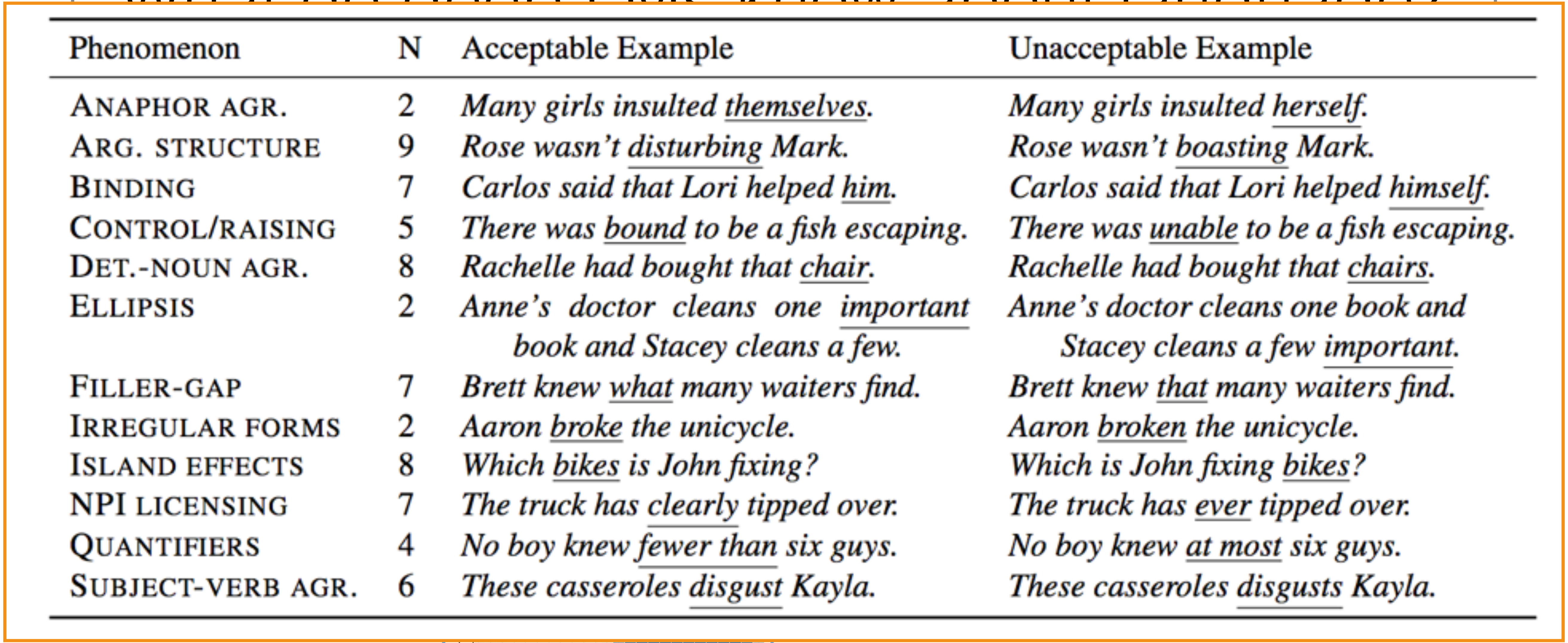


Tenney et al (ICLR 2018)



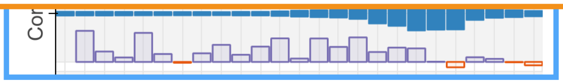
Tenney et al (ACL 2019)

Past ~2 years:
What do deep LMs know about language?



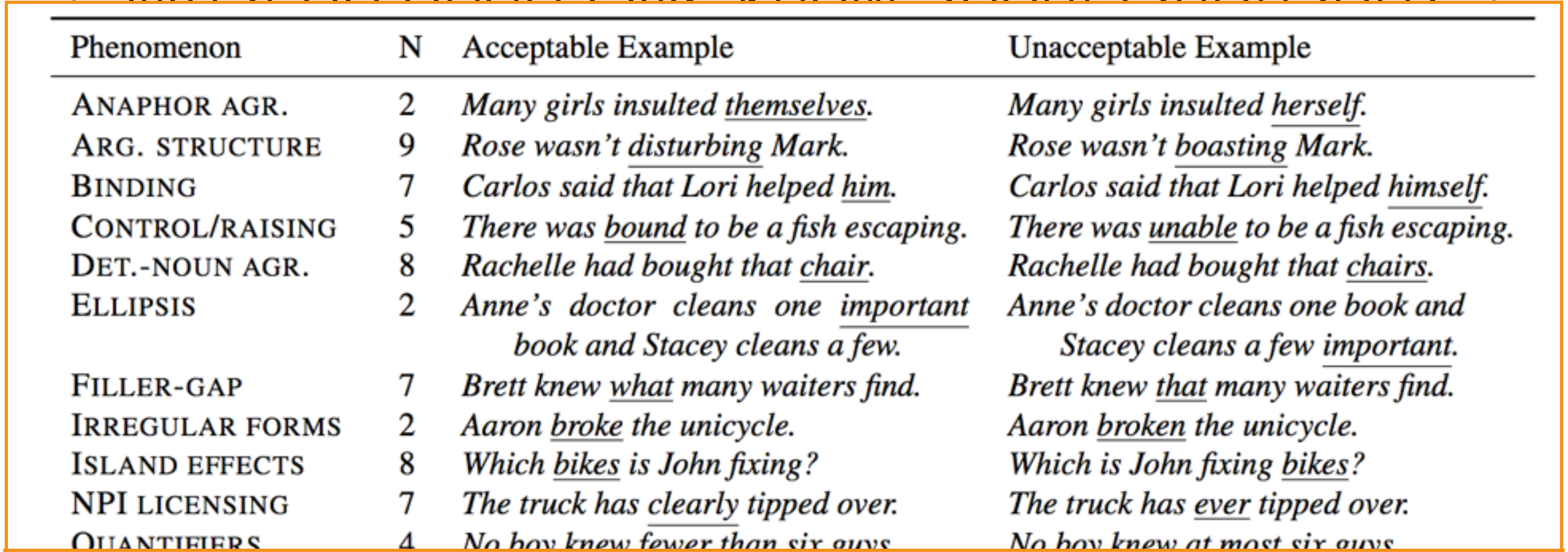
Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Warstadt et al (TACL 2020)

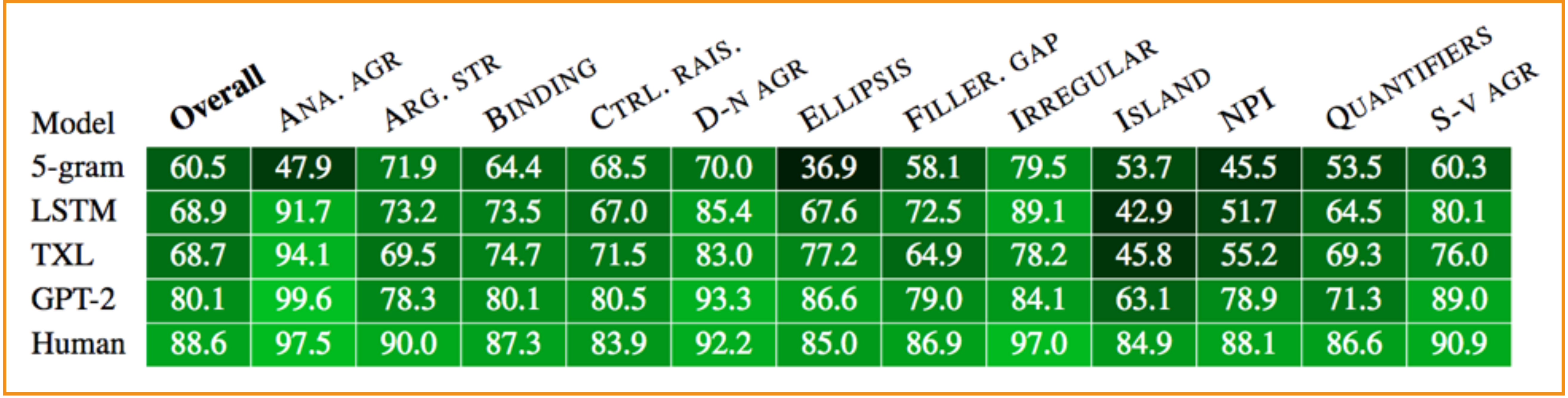


Tenney et al (ACL 2019)

Past ~2 years:
What do deep LMs know about language?



Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*

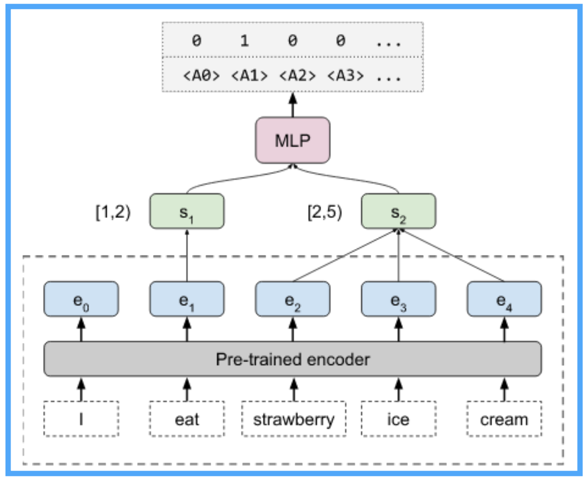


Model	Overall	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
5-gram	60.5	47.9	71.9	64.4	68.5	70.0	36.9	58.1	79.5	53.7	45.5	53.5	60.3
LSTM	68.9	91.7	73.2	73.5	67.0	85.4	67.6	72.5	89.1	42.9	51.7	64.5	80.1
TXL	68.7	94.1	69.5	74.7	71.5	83.0	77.2	64.9	78.2	45.8	55.2	69.3	76.0
GPT-2	80.1	99.6	78.3	80.1	80.5	93.3	86.6	79.0	84.1	63.1	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

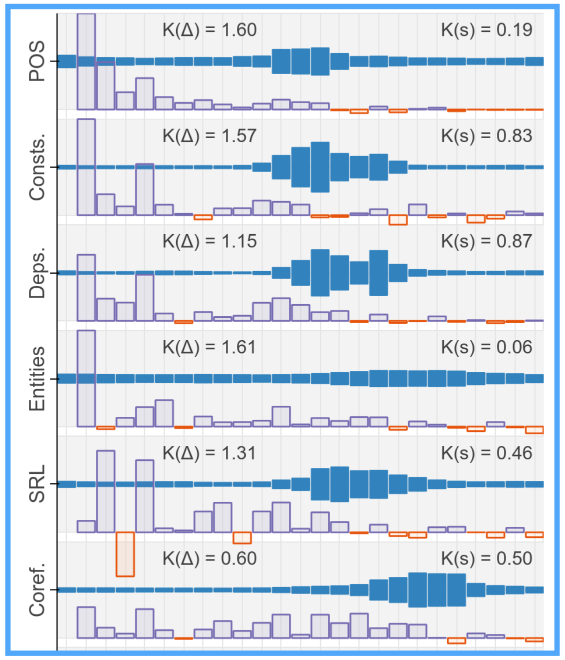
Past ~2 years:
What do deep LMs know about language?

Probing Classifiers: What types of linguistic structures do representations encode?

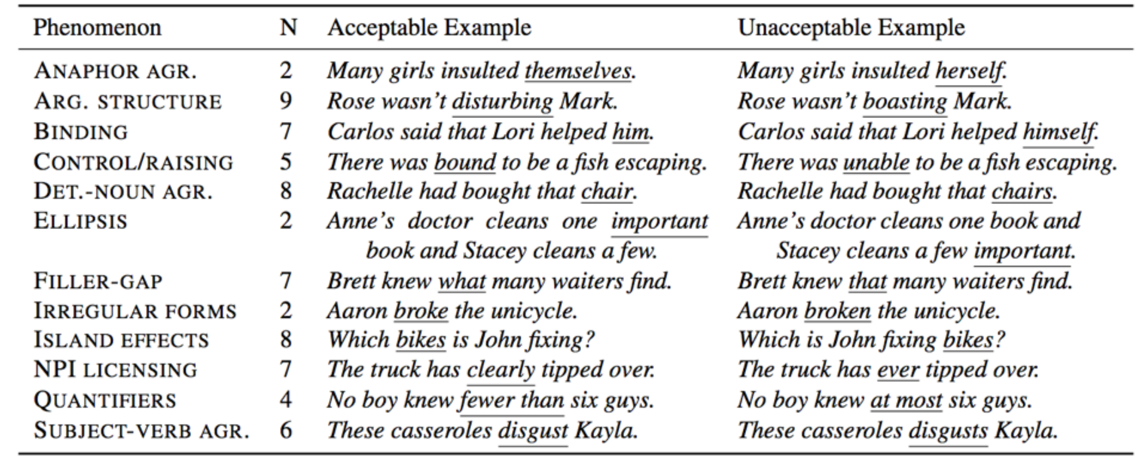
Challenge Tasks: How well do models perform on difficult “tail” events?



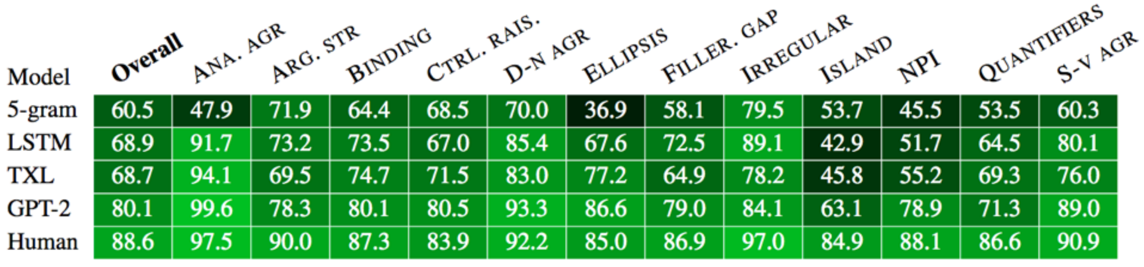
Tenney et al (ICLR 2018)



Tenney et al (ACL 2019)



Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn’t disturbing Mark.*	*Rose wasn’t boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne’s doctor cleans one important book and Stacey cleans a few.*	*Anne’s doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

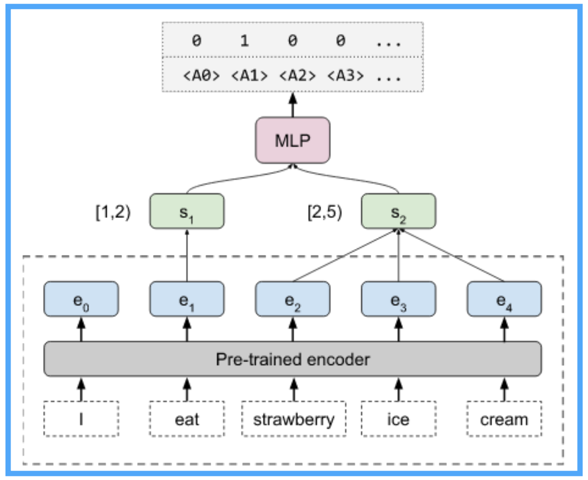


Model	**Overall**	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
5-gram	60.5	47.9	71.9	64.4	68.5	70.0	36.9	58.1	79.5	53.7	45.5	53.5	60.3
LSTM	68.9	91.7	73.2	73.5	67.0	85.4	67.6	72.5	89.1	42.9	51.7	64.5	80.1
TXL	68.7	94.1	69.5	74.7	71.5	83.0	77.2	64.9	78.2	45.8	55.2	69.3	76.0
GPT-2	80.1	99.6	78.3	80.1	80.5	93.3	86.6	79.0	84.1	63.1	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

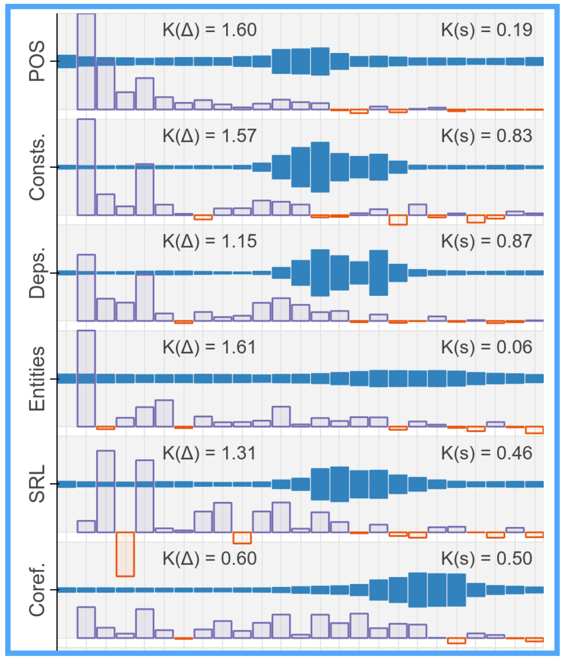
Warstadt et al (TACL 2020)

Past ~2 years:
What do deep LMs know about language?

Probing Classifiers: What types of linguistic structures do representations encode?



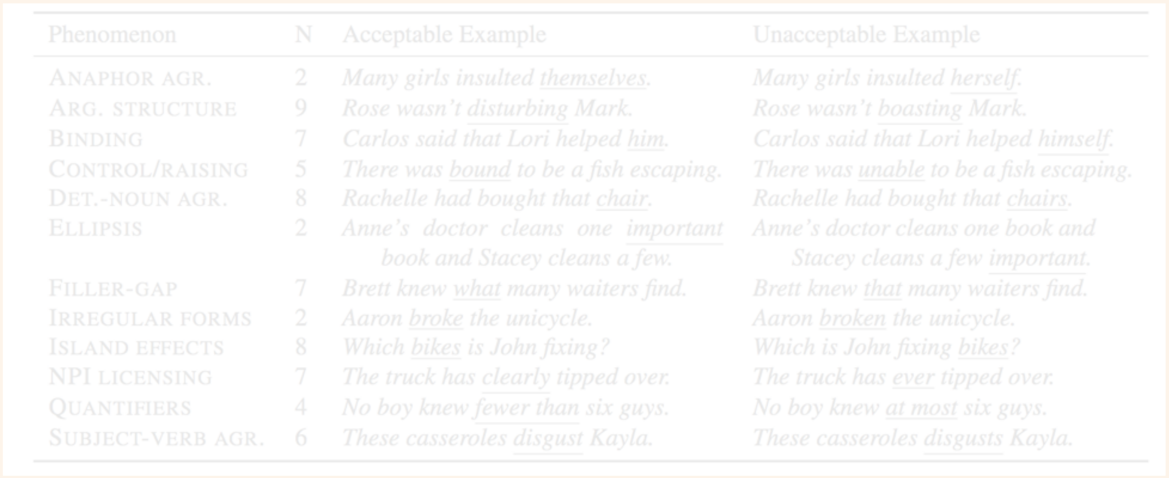
Tenney et al (ICLR 2018)



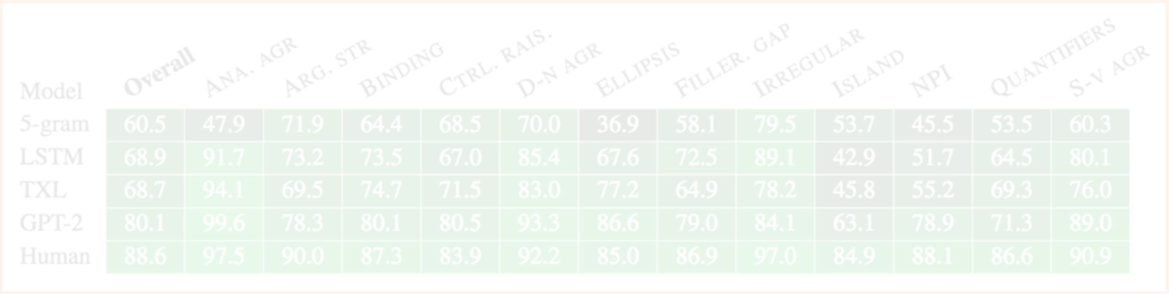
Tenney et al (ACL 2019)

Clear model of which structures should be represented.

Challenge Tasks: How well do models perform on difficult "tail" events?



Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

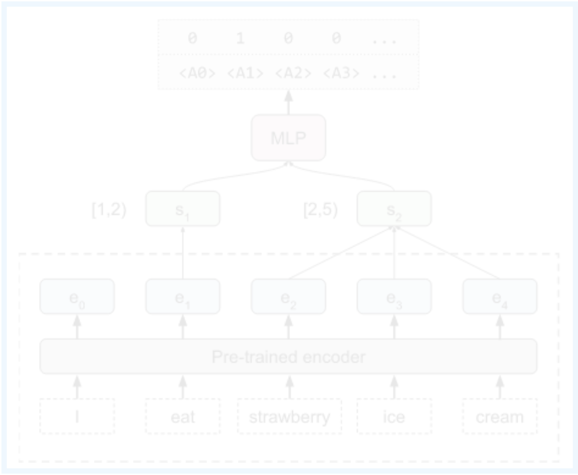


Model	Overall	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
5-gram	60.5	47.9	71.9	64.4	68.5	70.0	36.9	58.1	79.5	53.7	45.5	53.5	60.3
LSTM	68.9	91.7	73.2	73.5	67.0	85.4	67.6	72.5	89.1	42.9	51.7	64.5	80.1
TXL	68.7	94.1	69.5	74.7	71.5	83.0	77.2	64.9	78.2	45.8	55.2	69.3	76.0
GPT-2	80.1	99.6	78.3	80.1	80.5	93.3	86.6	79.0	84.1	63.1	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

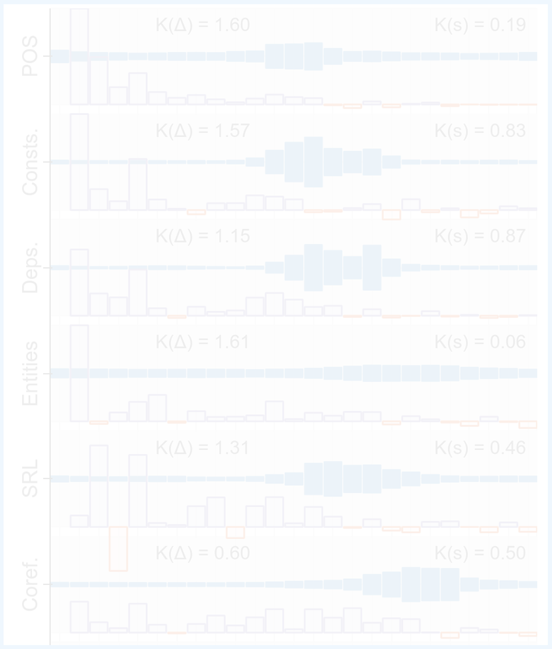
Warstadt et al (TACL 2020)

Past ~2 years:
What do deep LMs know about language?

Probing Classifiers: What types of linguistic structures do representations encode?

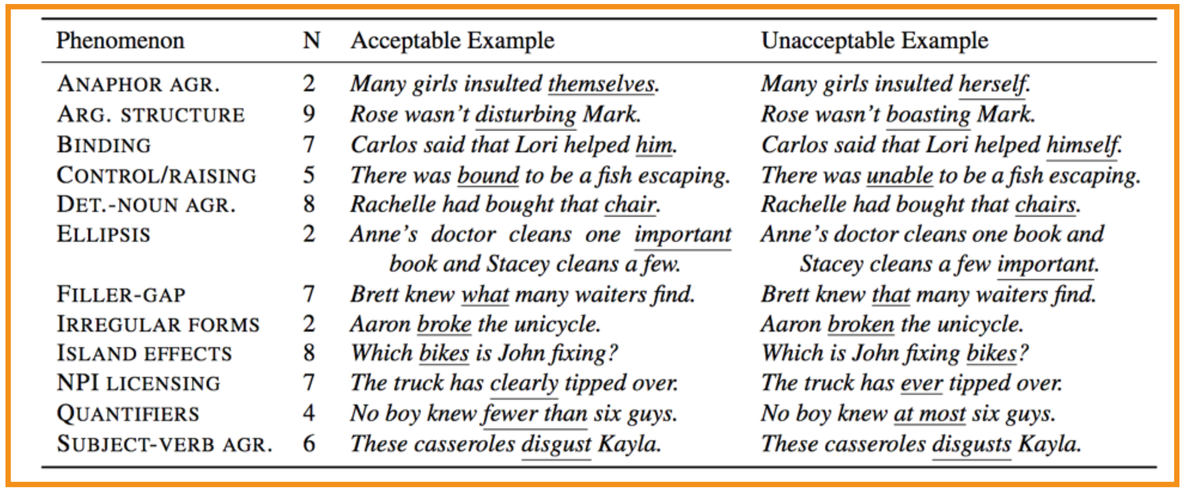


Tenney et al (ICLR 2018)

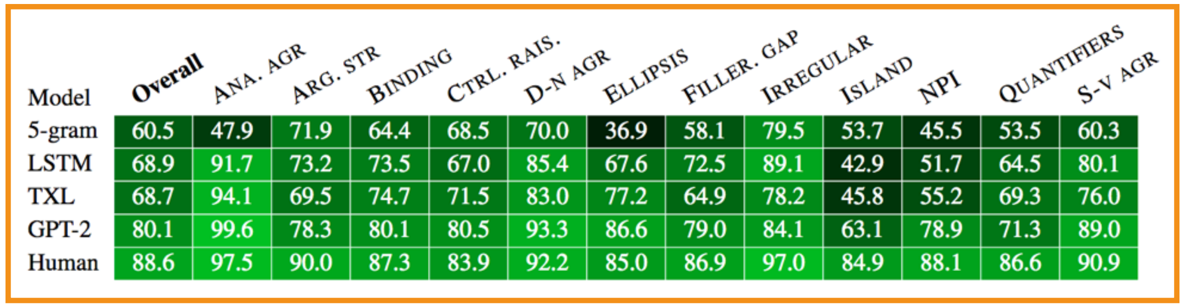


Tenney et al (ACL 2019)

Challenge Tasks: How well do models perform on difficult "tail" events?



Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*



Model	**Overall**	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
5-gram	60.5	47.9	71.9	64.4	68.5	70.0	36.9	58.1	79.5	53.7	45.5	53.5	60.3
LSTM	68.9	91.7	73.2	73.5	67.0	85.4	67.6	72.5	89.1	42.9	51.7	64.5	80.1
TXL	68.7	94.1	69.5	74.7	71.5	83.0	77.2	64.9	78.2	45.8	55.2	69.3	76.0
GPT-2	80.1	99.6	78.3	80.1	80.5	93.3	86.6	79.0	84.1	63.1	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Warstadt et al (TACL 2020)

Clear manifestation of phenomenon in the grammar.

Semantics, Pragmatics, “Common Sense”

Semantics, Pragmatics, “Common Sense”

- Do these models encode basic lexical concepts?
- Can these models compose those concepts?
- Do these model reason about context and “question under discussion”?

Semantics, Pragmatics, “Common Sense”

- Do these models encode basic lexical concepts?

The ball rolled down the hill

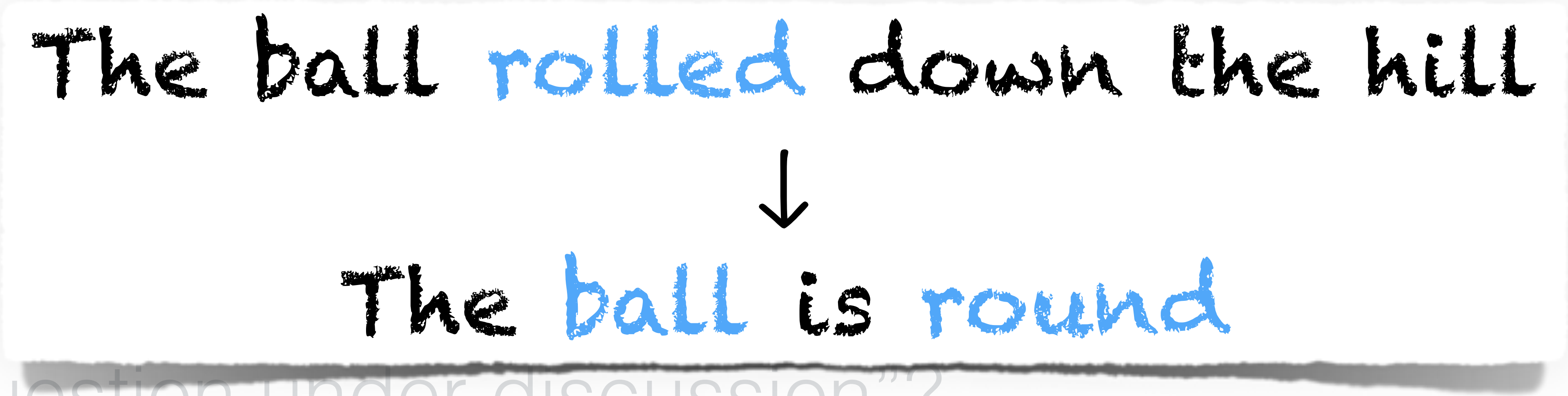
↓

The ball is round

- Ca
- Do
“question under discussion”?

Semantics, Pragmatics, “Common Sense”

- Do these models encode basic lexical concepts?



The ball rolled down the hill

↓

The ball is round

- Ca

- Do
  “question under discussion”?

Semantics, Pragmatics, “Common Sense”

- Do these models encode basic lexical concepts?

The dax rolled down the hill

↓

The dax is round

- Ca
- Do
  “question under discussion”?

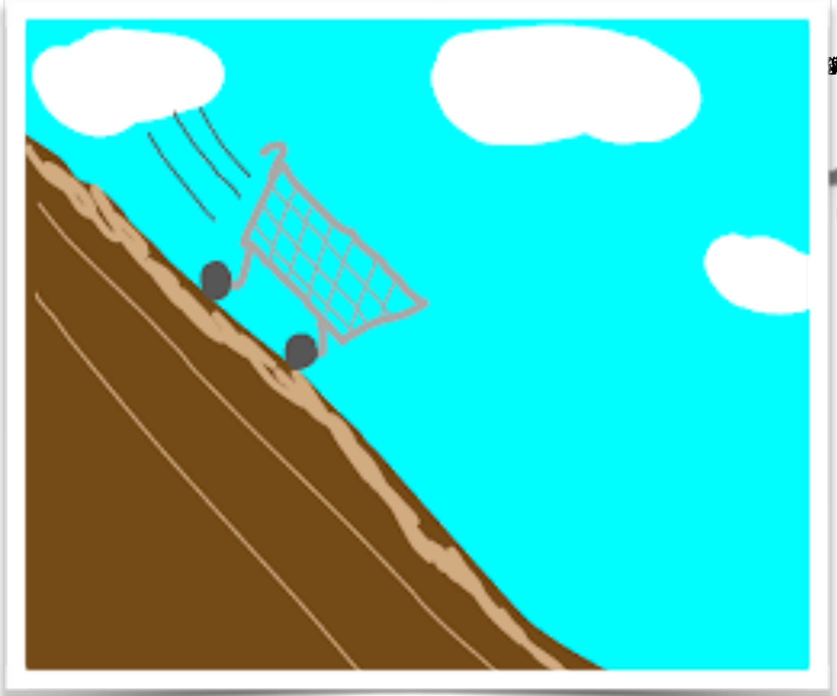
Semantics, Pragmatics, "Common Sense"

- Do these models encode basic lexical concepts?

The dax rolled down the hill

↓

The dax is round



- Ca

der discussion"?

Semantics, Pragmatics, “Common Sense”

- Do these models encode basic lexical concepts?
- Can these models compose those concepts?
- Do these model reason about context and “question under discussion”?

Semantics, Pragmatics, “Common Sense”

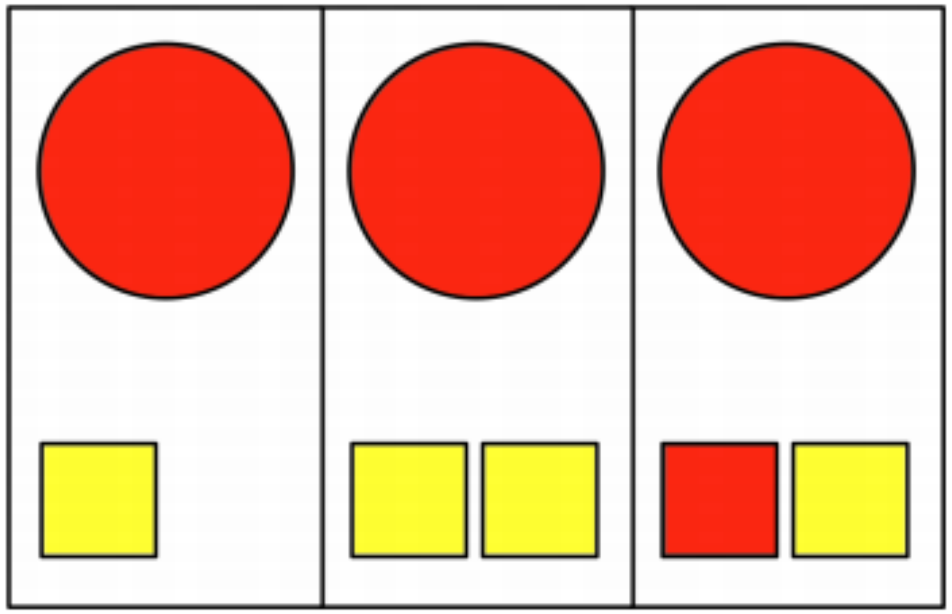
- Do these models encode basic lexical concepts?
- Can these models compose those concepts?
- Do
  “q

The ball that rolled is not red

↓

??

None of these three circles have the same color as both of the squares in their own cell



On the semantics of phi features on pronouns. Sudo (2012).

- Do … pts?

- Can these models compose those concepts?

- D… “q…

The ball that rolled is not red

↓

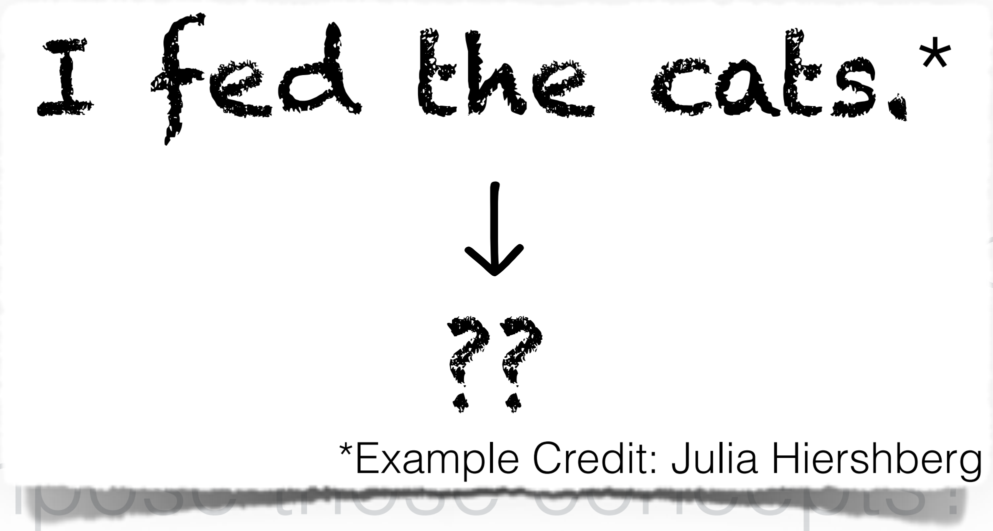
??

Semantics, Pragmatics, “Common Sense”

- Do these models encode basic lexical concepts?
- Can these models compose those concepts?
- Do these model reason about context and “question under discussion”?

Semantics, Pragmatics, “Common Sense”

- Do these models enco pts?
- Can these models compose these concepts?



I fed the cats.*

↓

??

*Example Credit: Julia Hiershberg

- Do these model reason about context and “question under discussion”?

Semantics, Pragmatics, “Common Sense”

Did you feed the animals?

↓

I fed the cats…*

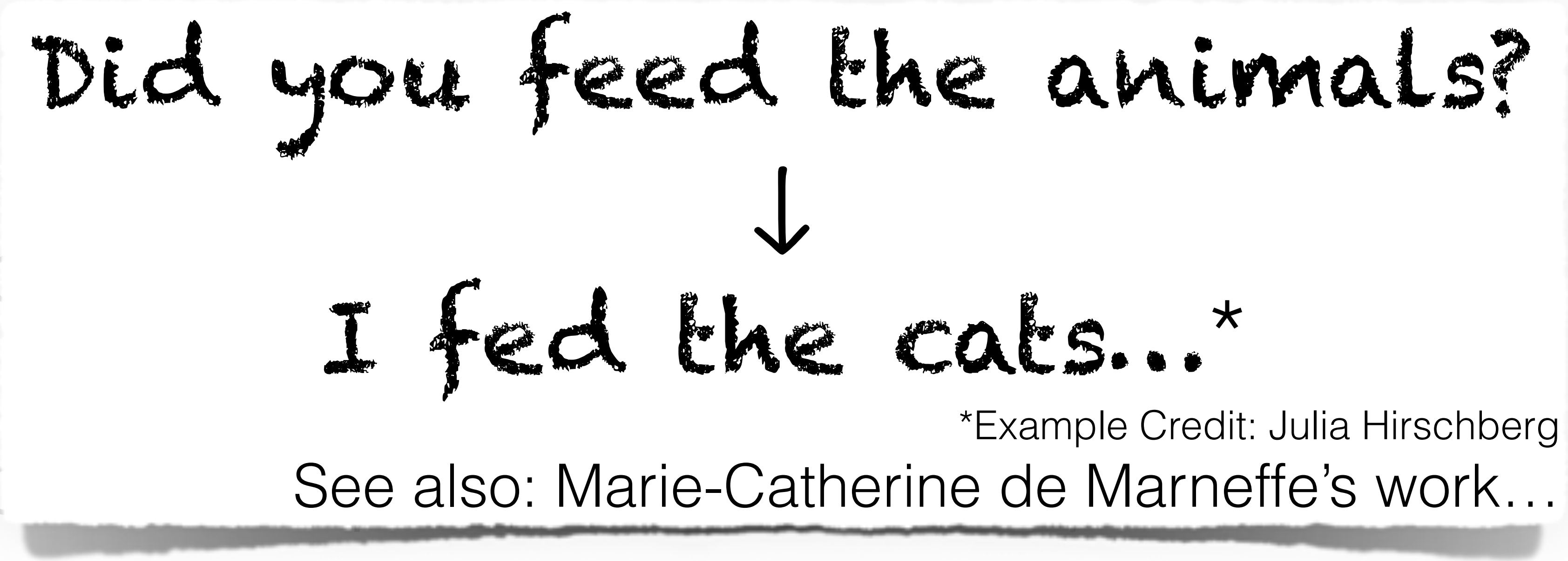
*Example Credit: Julia Hirschberg

- Do these mode
- Can these models compose these concepts?
- Do these model reason about context and “question under discussion”?

Semantics, Pragmatics, “Common Sense”

- Do these mode

- Can these mod



Did you feed the animals?

↓

I fed the cats...*

*Example Credit: Julia Hirschberg

See also: Marie-Catherine de Marneffe’s work…

- Do these model reason about context and “question under discussion”?

Semantics, Pragmatics, “Common Sense”

Is the King of France bald?

↓

There is no King of France!

al concepts?

- Can these models compose those concepts?

- Do these model reason about context and “question under discussion”?

Major Challenges

Major Challenges

- Living area of research—we can’t ask linguistics to just lend us some ready-to-go evaluations

Major Challenges

- Living area of research—we can't ask linguistics to just lend us some ready-to-go evaluations
- Good "probing tasks" require situation and grounding—to vision, dialog, etc—which makes error attribution very difficult

Major Challenges

- Living area of research—we can't ask linguistics to just lend us some ready-to-go evaluations

- Good "probing tasks" require situation and grounding—to vision, dialog, etc—which makes error attribution very difficult

- Human baselines are hard pin down. Variation is high and agreement often low. Experimental designs are usually carefully and highly contrived.

Three Case Studies

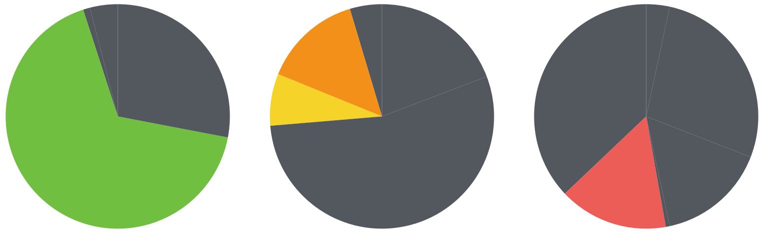
Three Case Studies

Modifier-Noun Composition

fake gun

↓

gun



Most babies are little and most problems are huge: Compositional Entailment in Adjective-Nouns. Pavlick and Callison-Burch (2016)

So-Called Nonsubsective Adjectives. Pavlick and Callison-Burch (2016)

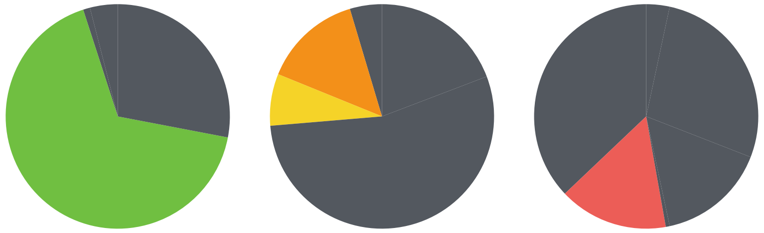
Three Case Studies

Modifier-Noun Composition

fake gun

↓

gun



Most babies are little and most problems are huge: Compositional Entailment in Adjective-Nouns. Pavlick and Callison-Burch (2016)

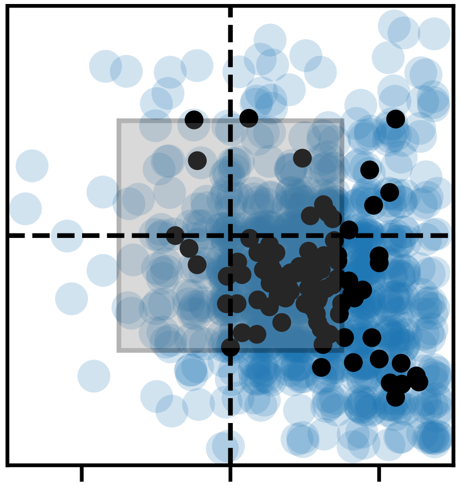
So-Called Nonsubsective Adjectives. Pavlick and Callison-Burch (2016)

Verb-Complement Composition

attempt to sing

↓

sing



Do NLI models capture verb veridicality? Ross and Pavlick (2019)

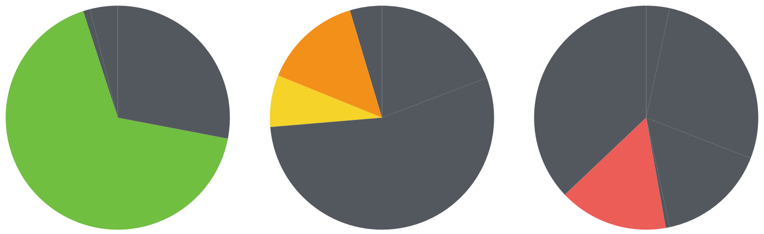
Three Case Studies

Modifier-Noun Composition

fake gun

↓

gun



Most babies are little and most problems are huge: Compositional Entailment in Adjective-Nouns. Pavlick and Callison-Burch (2016)

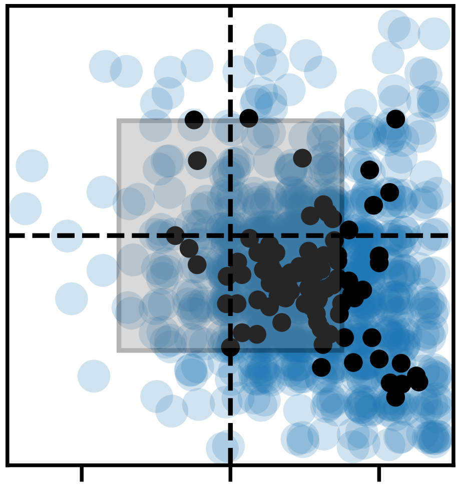
So-Called Nonsubsective Adjectives. Pavlick and Callison-Burch (2016)

Verb-Complement Composition

attempt to sing

↓

sing



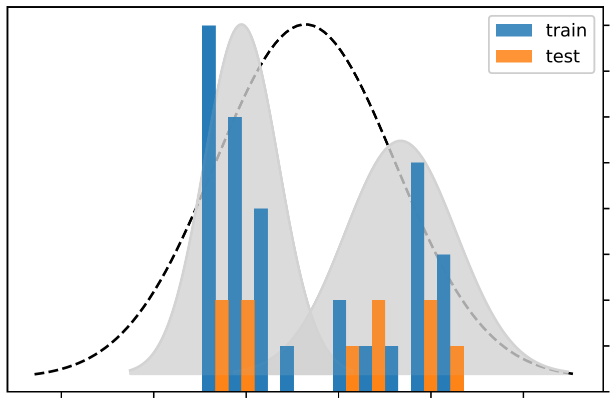
Do NLI models capture verb veridicality? Ross and Pavlick (2019)

Sentence-Level Inference

A man is standing under a tree

↓

A person is outside.



Inherent Disagreements in Human Textual Inferences. Pavlick and Kwiatkowski (2020)

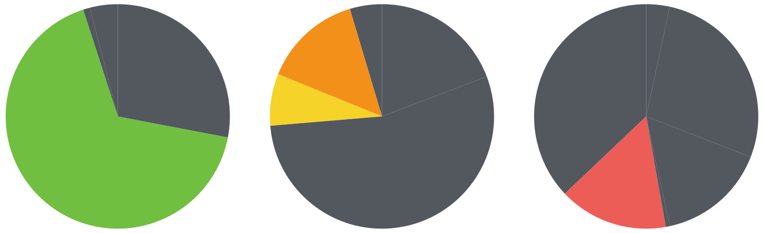
Three Case Studies

Modifier-Noun Composition

fake gun

↓

gun



Most babies are little and most problems are huge: Compositional Entailment in Adjective-Nouns. Pavlick and Callison-Burch (2016)

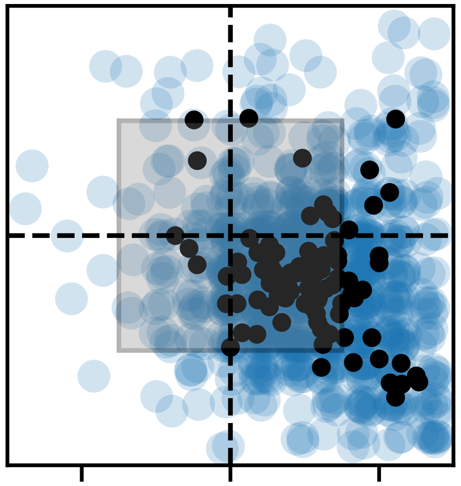
So-Called Nonsubsective Adjectives. Pavlick and Callison-Burch (2016)

Verb-Complement Composition

attempt to sing

↓

sing



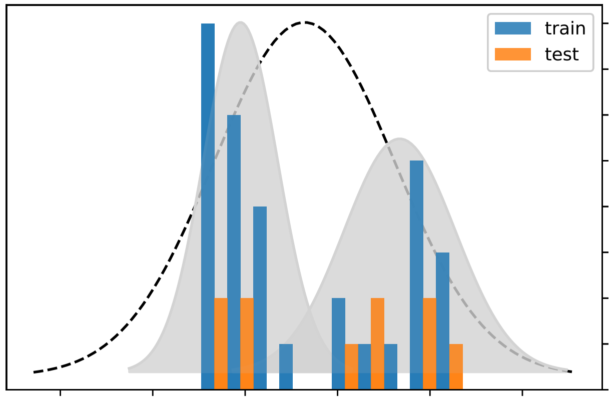
Do NLI models capture verb veridicality? Ross and Pavlick (2019)

Sentence-Level Inference

A man is standing under a tree

↓

A person is outside.

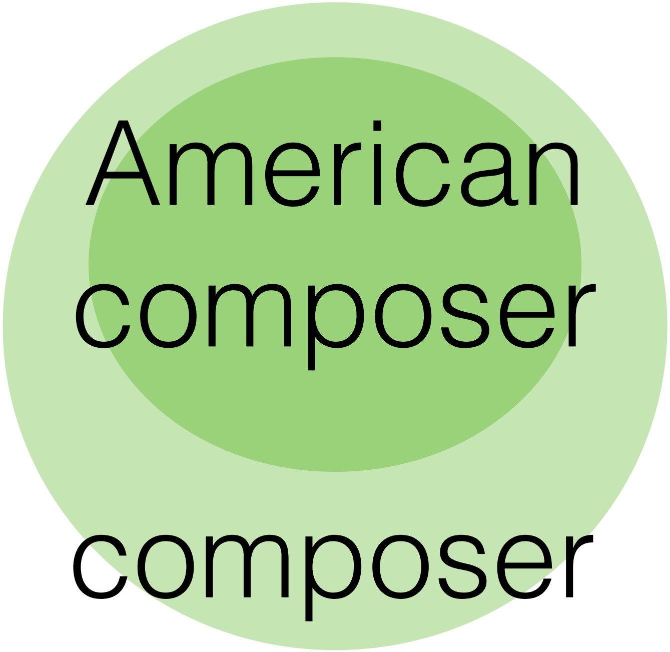


Inherent Disagreements in Human Textual Inferences. Pavlick and Kwiatkowski (2020)

Classes of Modifiers

Classes of Modifiers

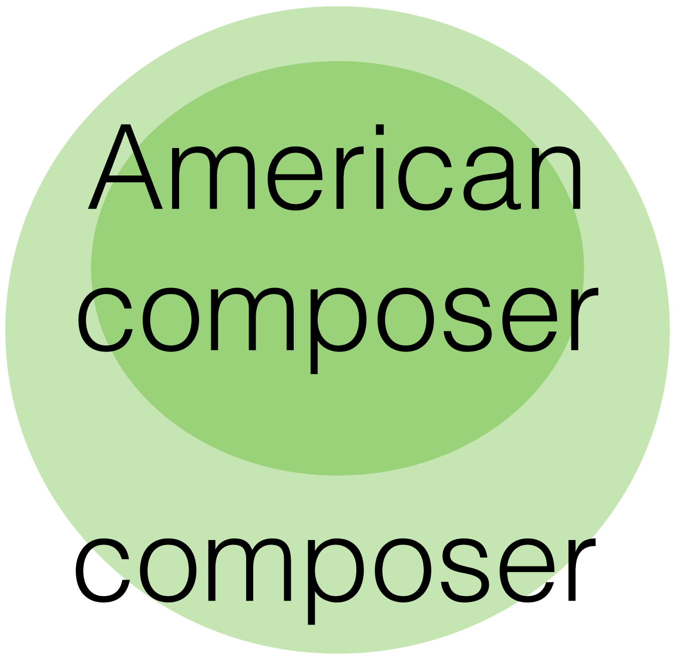
MH ⇒ H



Subsective

Classes of Modifiers

$MH \Rightarrow H$

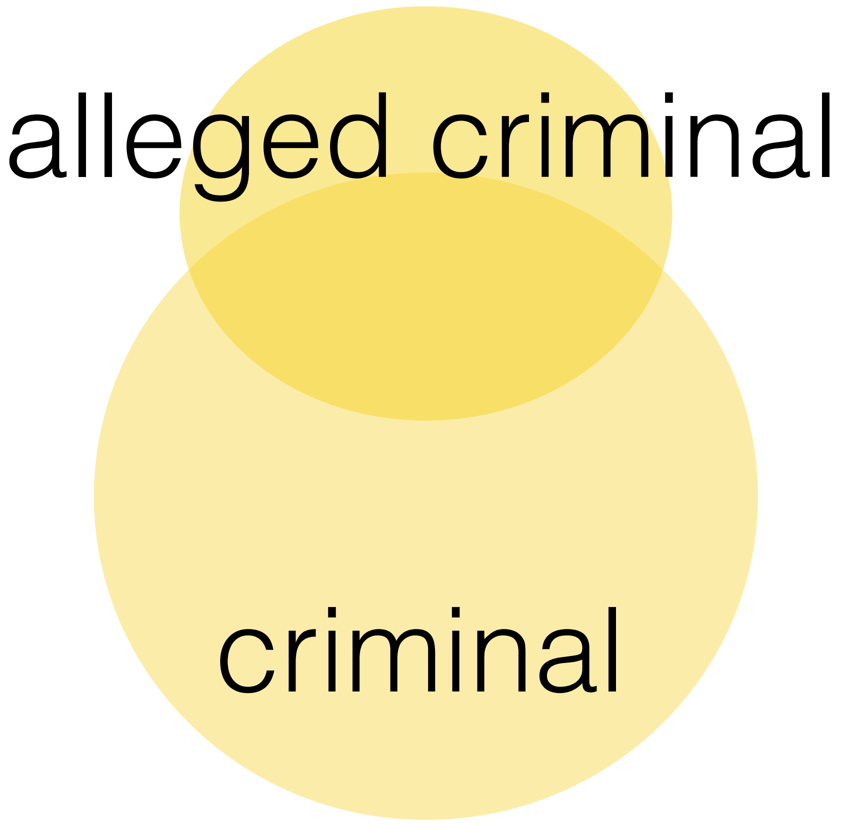


American composer

composer

Subsective

$MH \nRightarrow H$



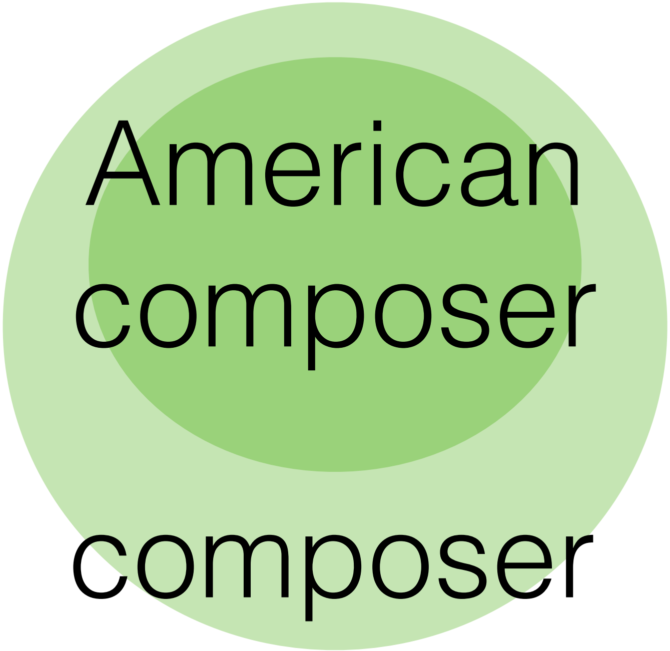
alleged criminal

criminal

Plain Non-Subsective

Classes of Modifiers

$MH \Rightarrow H$

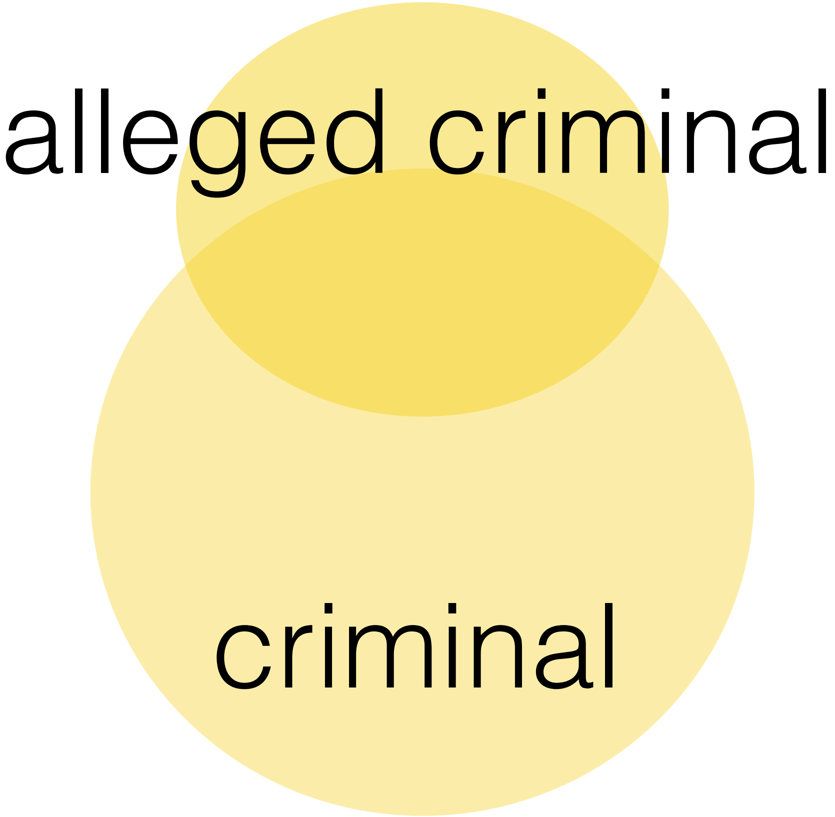


American composer

composer

Subsective

$MH \nRightarrow H$

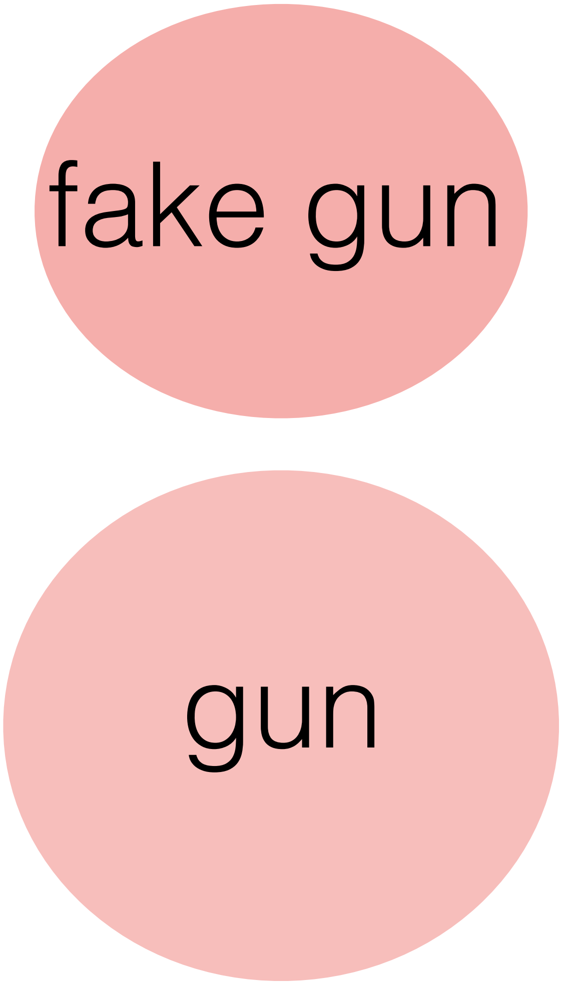


alleged criminal

criminal

Plain Non-Subsective

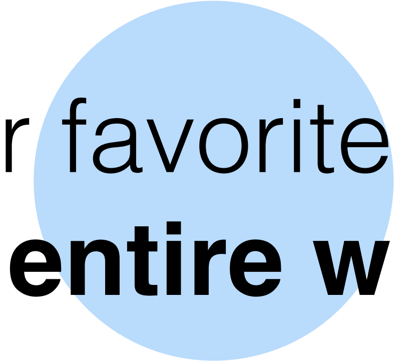
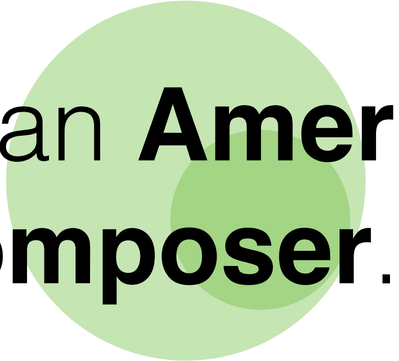
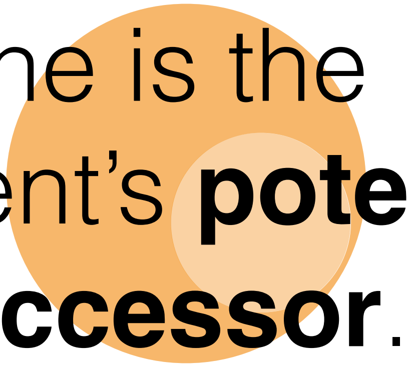
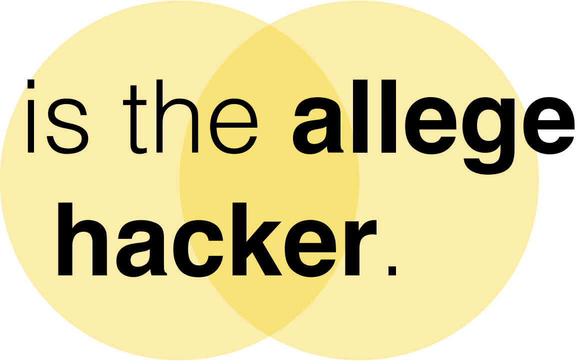
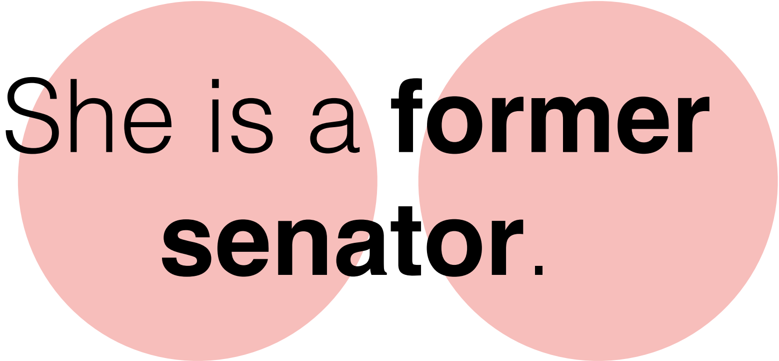
$MH \Rightarrow \neg H$



fake gun

gun

Privative

Equivalence	$MH \Longleftrightarrow H$	 It is her favorite book in the **entire world**.
Reverse Entailment	$MH \Rightarrow H \land$ $H \nRightarrow MH$	 She is an **American composer**.
Forward Entailment	$MH \nRightarrow H \land$ $H \Rightarrow MH$	 She is the president's **potential successor**.
Independence	$MH \nRightarrow H \land$ $H \nRightarrow MH$	 She is the **alleged hacker**.
Exclusion	$MH \Rightarrow \neg H \land$ $H \Rightarrow \neg MH$	 She is a **former senator**.

Experimental Design

Experimental Design

$H \Rightarrow MH?$

Eddy is a **cat**.

Eddy is a **domestic cat**.

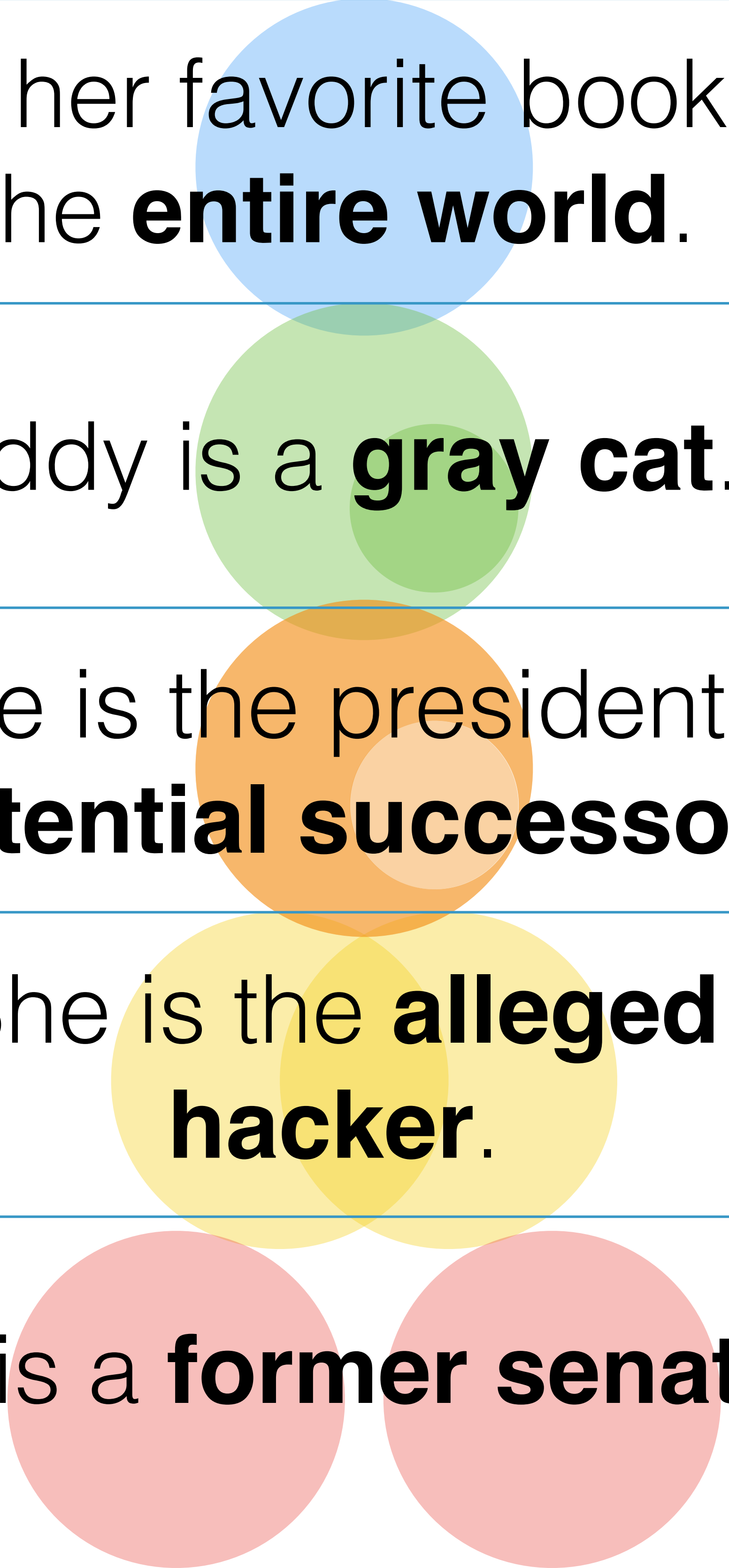
Experimental Design

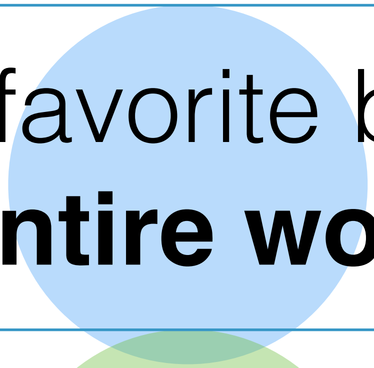
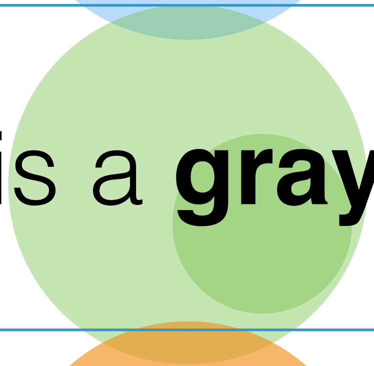
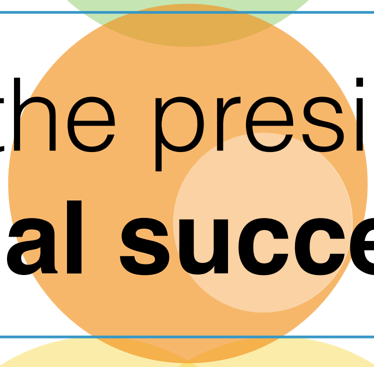
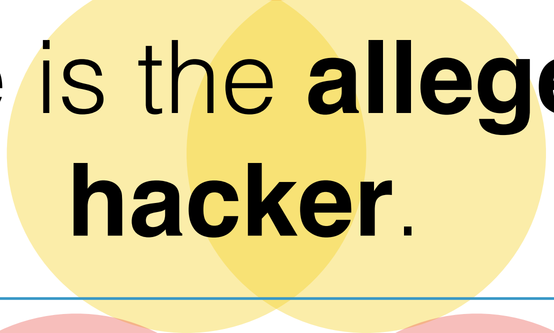
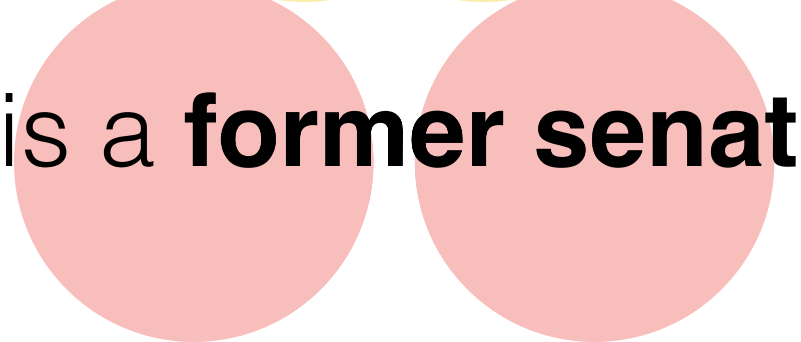
$$MH \Rightarrow H?$$

Eddy is a **domestic cat**.

Eddy is a **cat**.

	MH ⇒ H	H ⇒ MH	
Equiv.	Yes	Yes	It is her favorite book in the **entire world**.
Rev. Ent.	Yes	Unk	Eddy is a **gray cat**.
For. Ent.	Unk	Yes	She is the president's **potential successor**.
Indep.	Unk	Unk	She is the **alleged hacker**.
Excl.	No	No	She is a **former senator**.



	MH ⇒ H	H ⇒ MH	
Equiv.	Yes	Yes	 It is her favorite book in the **entire world**.
Rev. Ent.	Yes	Unk	 Eddy is a **gray cat**.
For. Ent.	Unk	Yes	 She is the president's **potential successor**.
Indep.	Unk	Unk	 She is the **alleged hacker**.
Excl.	No	No	 She is a **former senator**.

~200 human annotators
~5,000 sentences

	MH ⇒ H	H ⇒ MH	
Equiv.	Yes	Yes	It is her favorite book in the **entire world**.
Rev. Ent.	Yes	Unk	Eddy is a **gray cat**.
For. Ent.	Unk	Yes	She is the president's **potential successor**.
Indep.	Unk	Unk	She is the **alleged** [illegible].
Excl.	No	No	She is a **former senator**.

~200 human annotators
~5,000 sentences

4 Genres

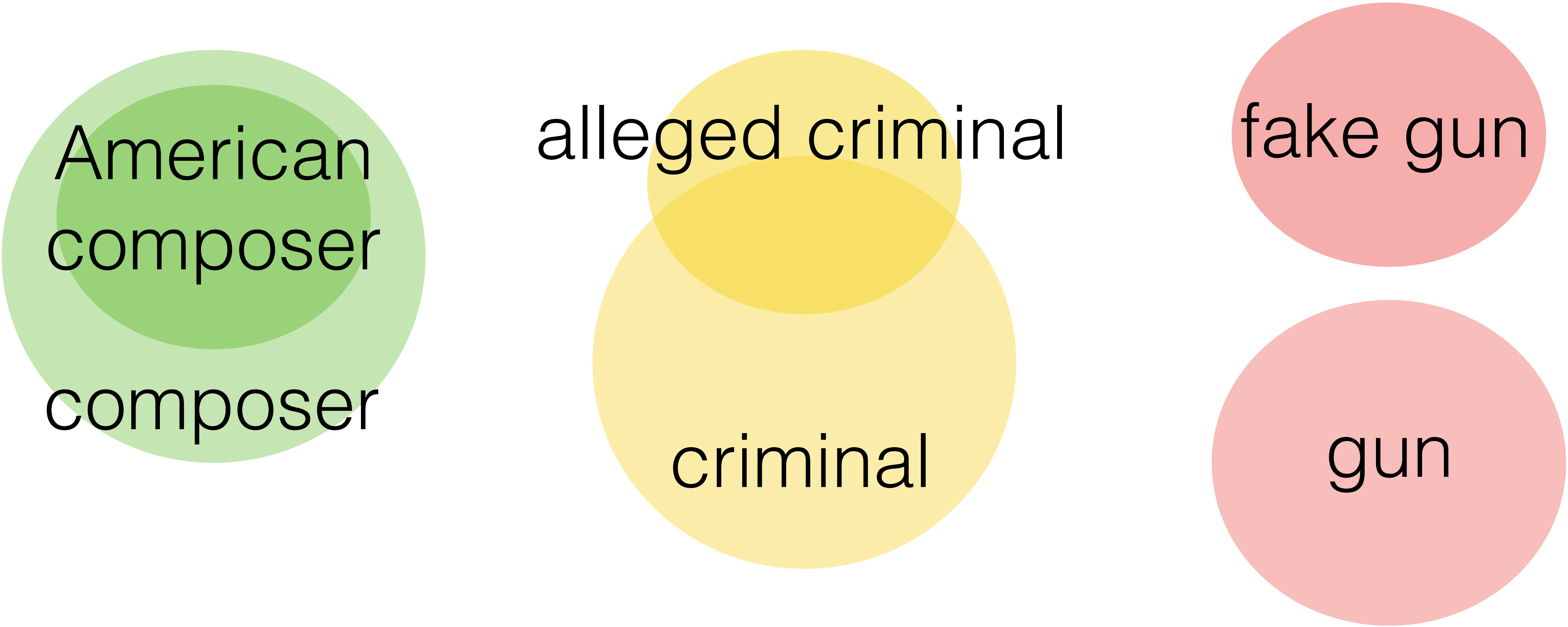
- News (Gigaword)
- Forums (Internet Argumentation Corpus)
- Image Captions (Denotation Graph)
- Literature (Gutenberg Prose Fiction)

Classes of Modifiers

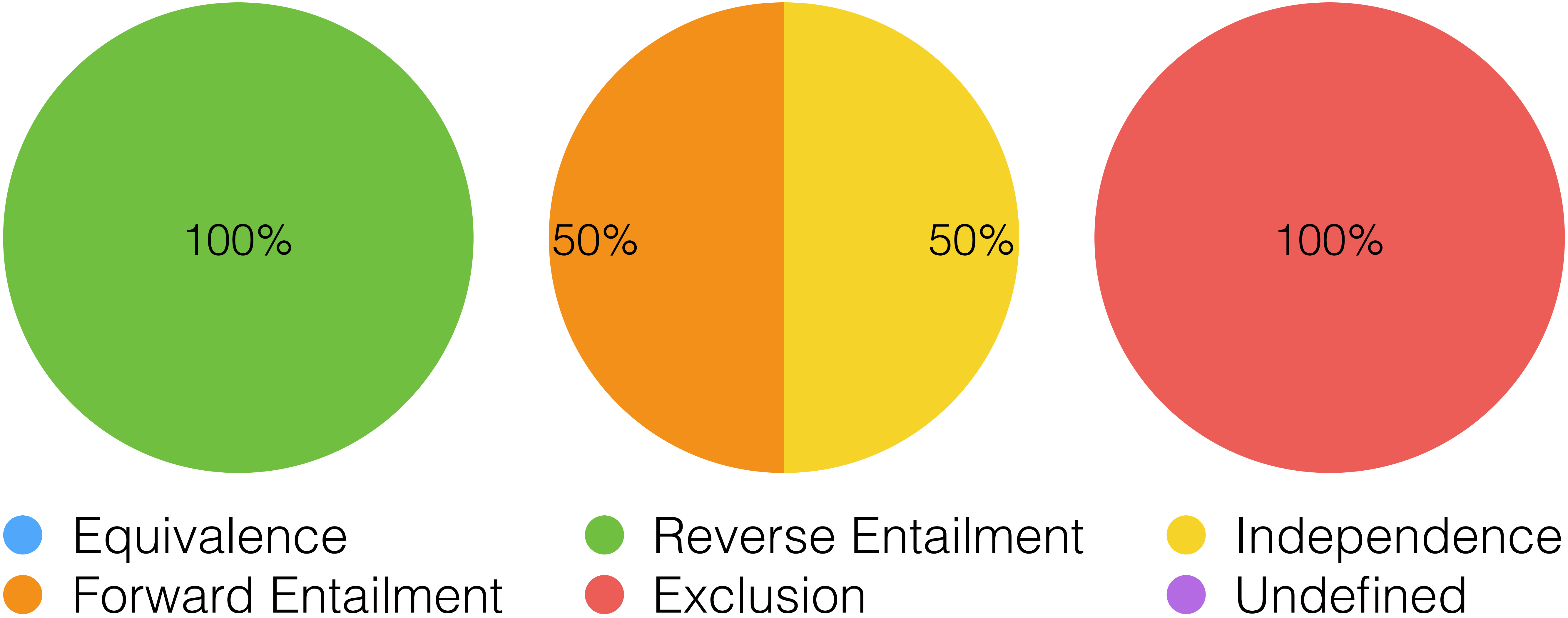
Subsective
$MH \Rightarrow H$

Plain Non-Subsective
$MH \nRightarrow H$

Privative
$MH \Rightarrow \neg H$



Classes of Modifiers

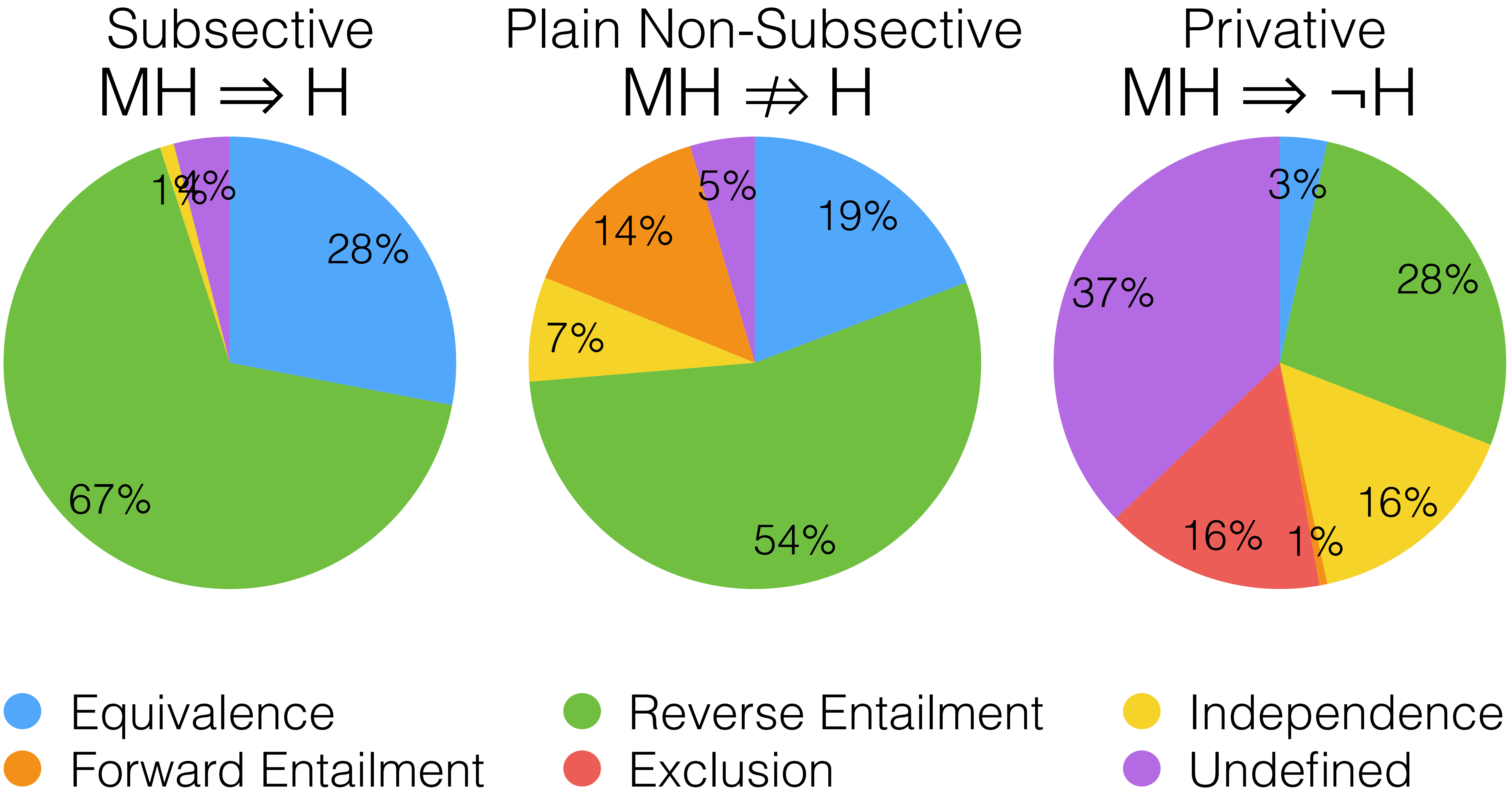


Subsective $MH \Rightarrow H$

Plain Non-Subsective $MH \not\Rightarrow H$

Privative $MH \Rightarrow \neg H$

Classes of Modifiers



Subsective

$MH \Rightarrow H$

Plain Non-Subsective

$MH \nRightarrow H$

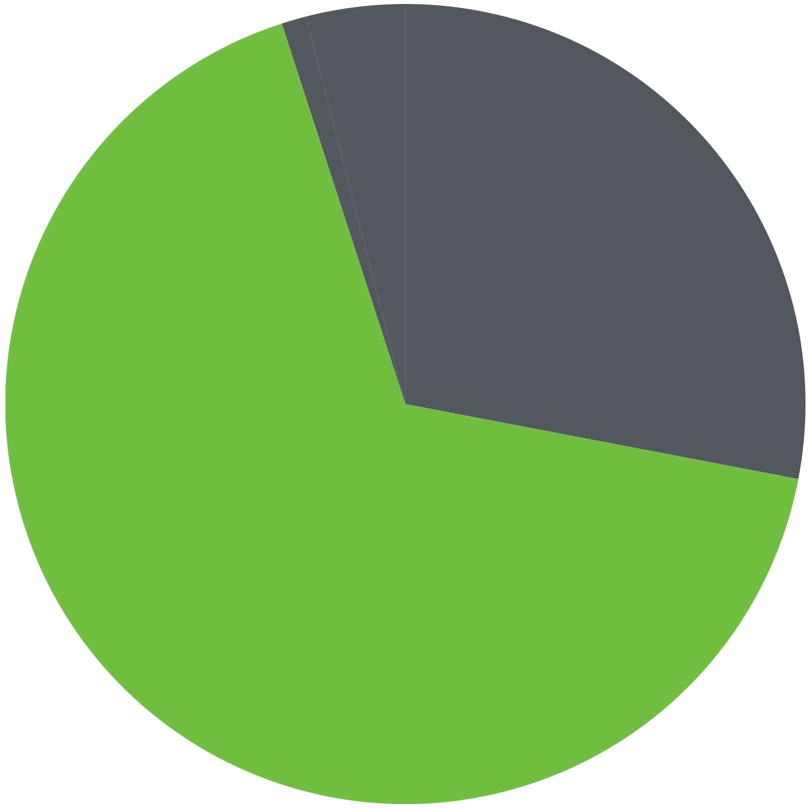
Privative

$MH \Rightarrow \neg H$

Classes of Modifiers

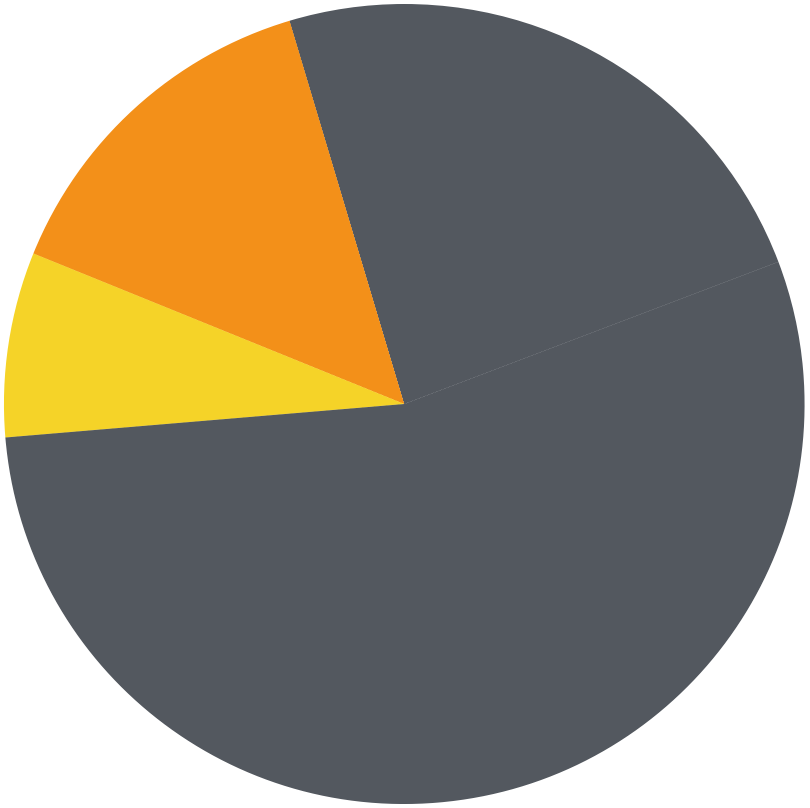
Subsective

$MH \Rightarrow H$



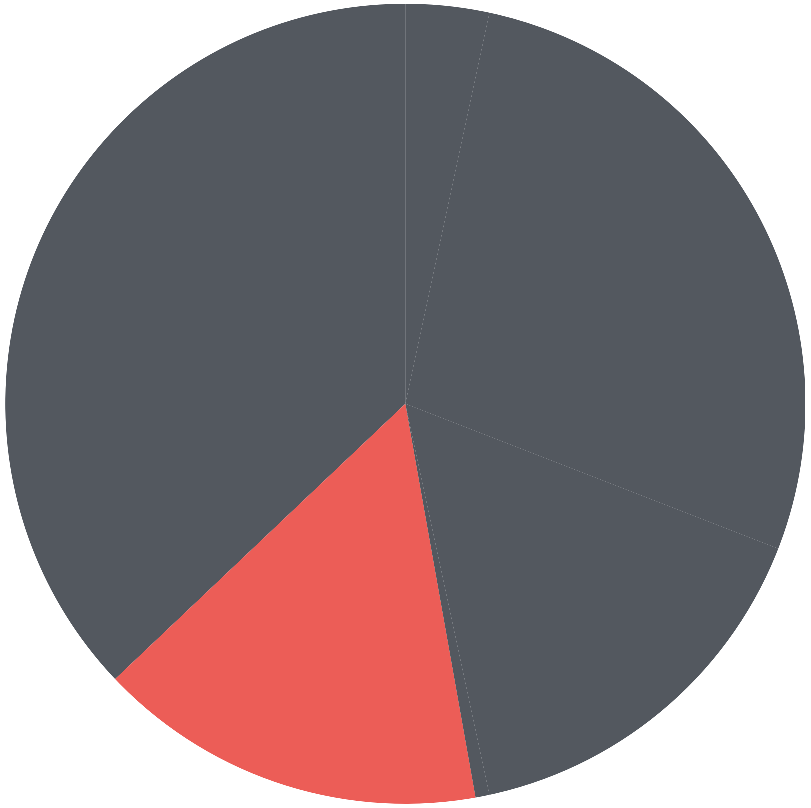
Plain Non-Subsective

$MH \nRightarrow H$



Privative

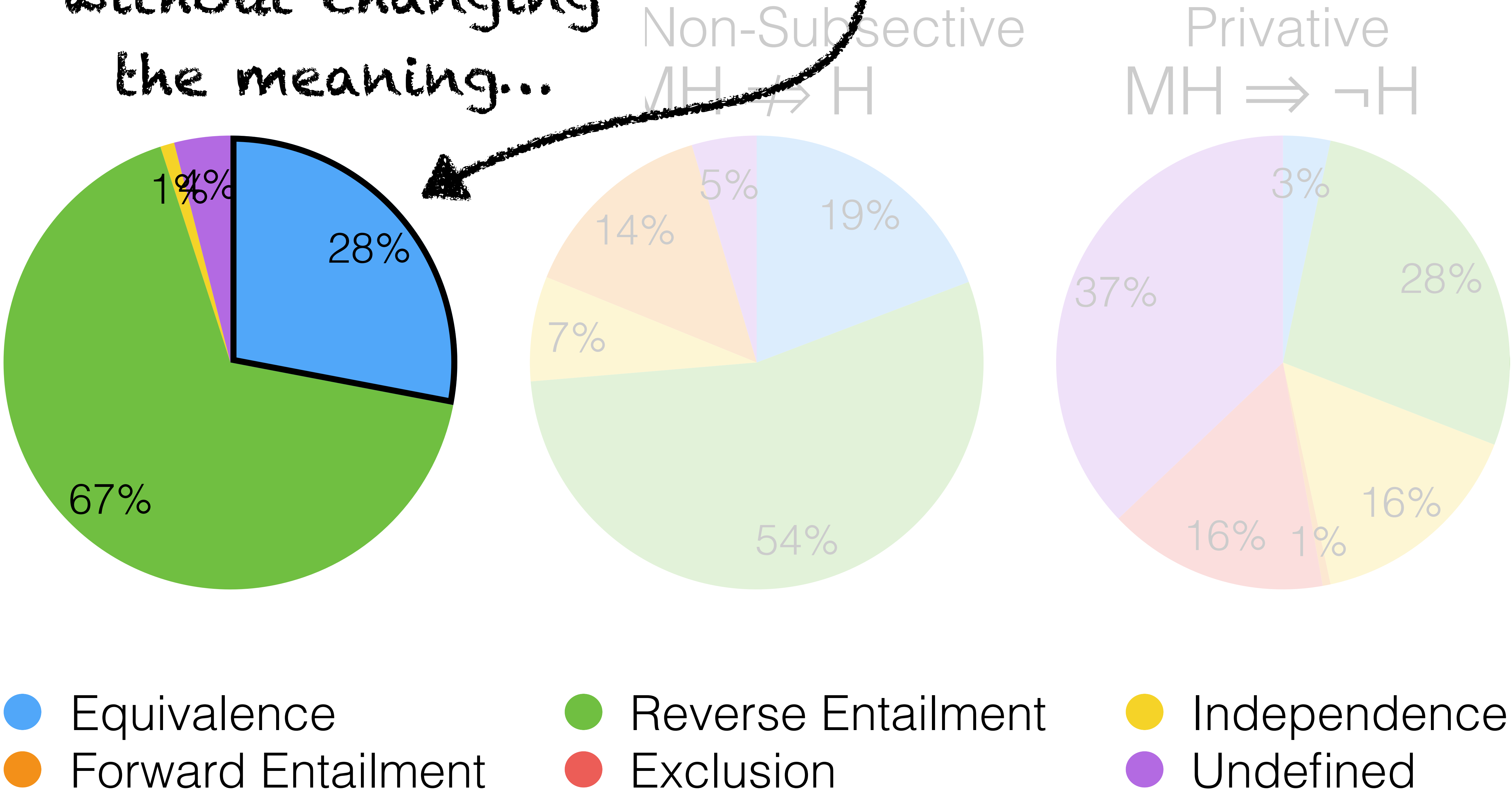
$MH \Rightarrow \neg H$



Generalizations based on the class of the modifier lead to incorrect predictions more often than not.

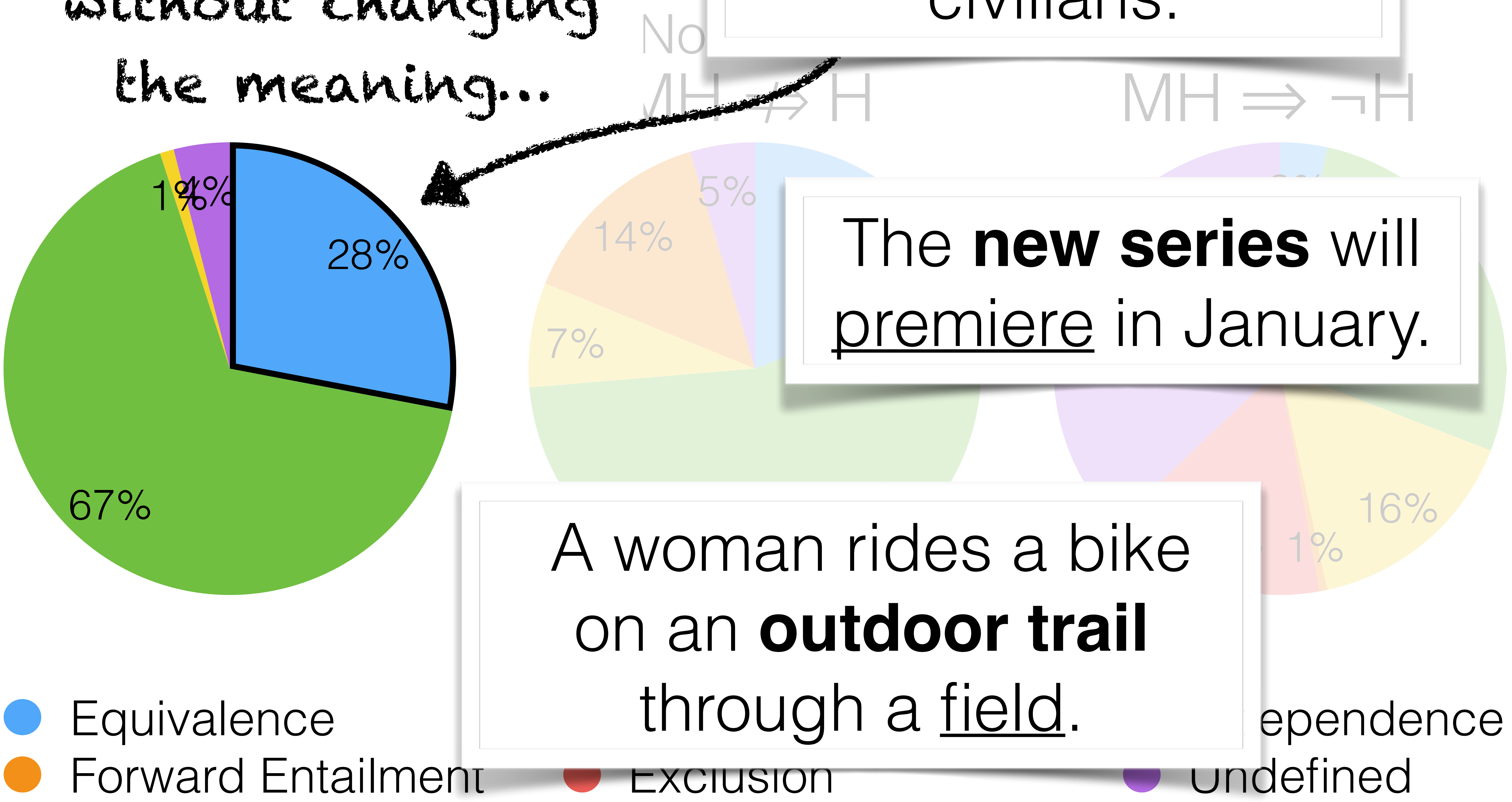
# of Modifiers

sometimes we can insert an adjective without changing the meaning...



sometimes we can insert an adjective without changing the meaning...

The **deadly attack** killed at least 12 civilians.



The **new series** will premiere in January.

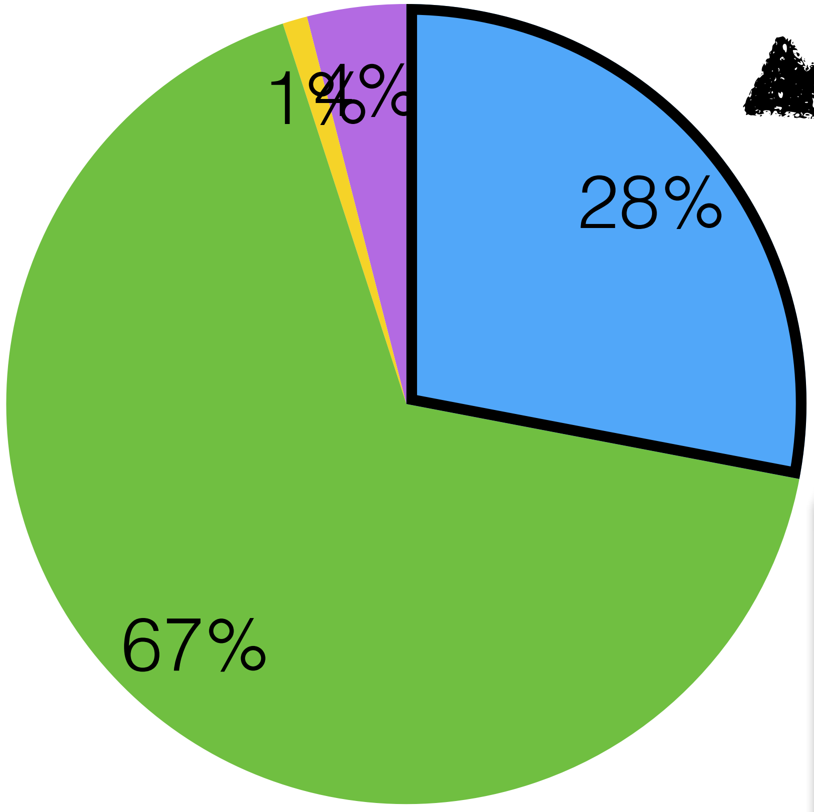
A woman rides a bike on an **outdoor trail** through a field.

sometimes we can insert an adjective without changing the meaning...

The **entire bill** is now subject to approval by the parliament.

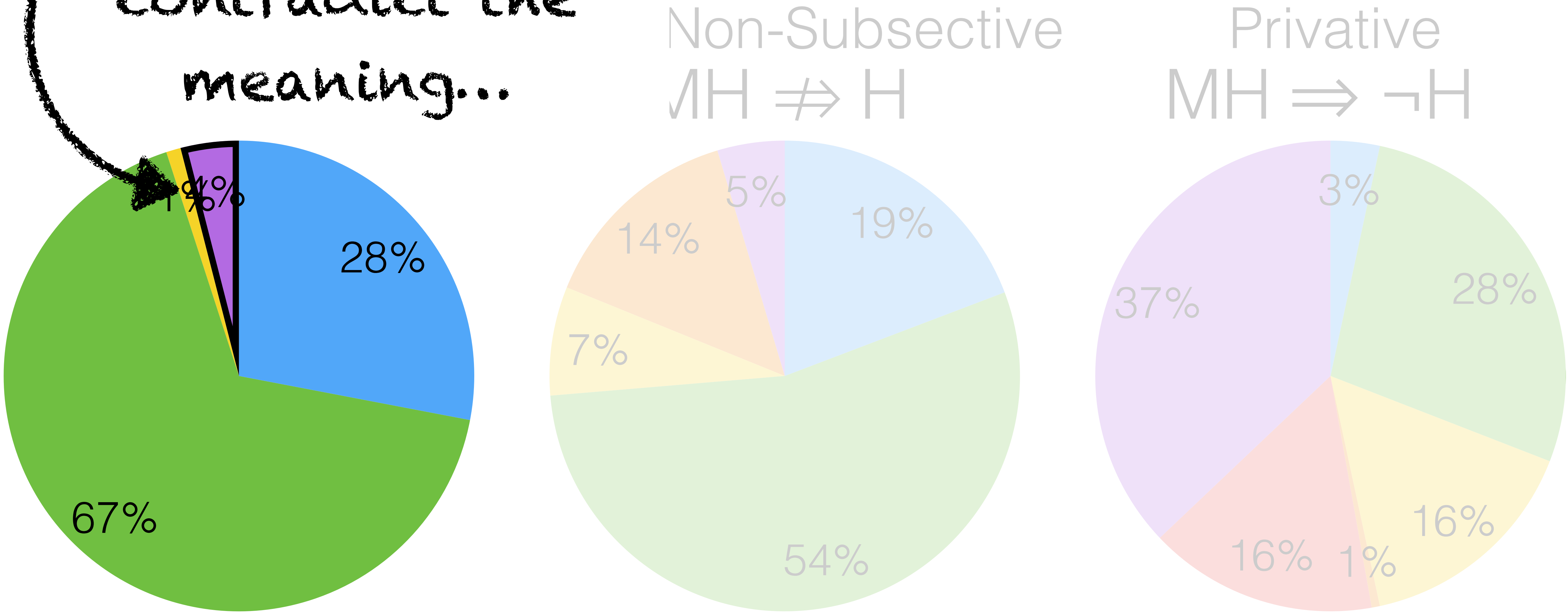
Greenberg also was put under investigation for his **crucial role** at the company.

I simply love the **actual experience** of being one with the ocean and the life in it.



# of Modifiers

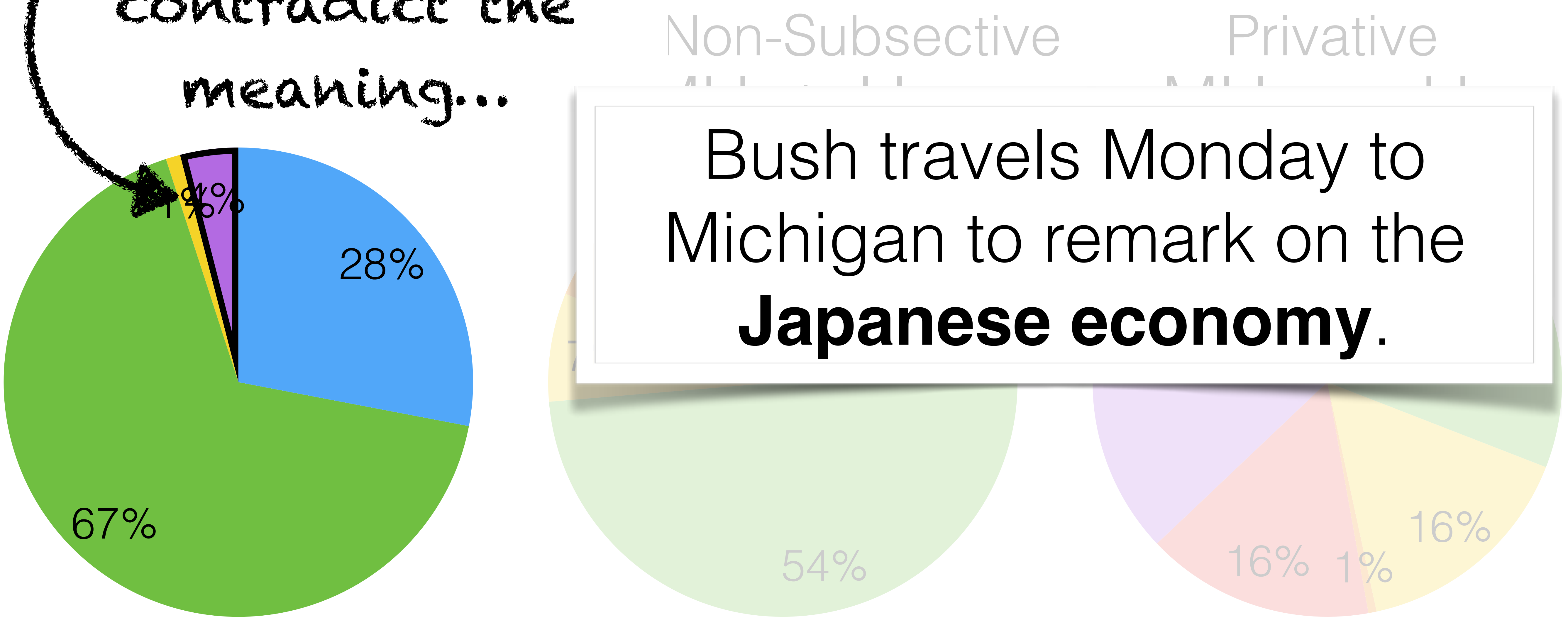
sometimes if we insert an adjective, we appear to contradict the meaning...



● Equivalence
● Forward Entailment
● Reverse Entailment
● Exclusion
● Independence
● Undefined

sometimes if we insert an adjective, we appear to contradict the meaning...

# of Modifiers



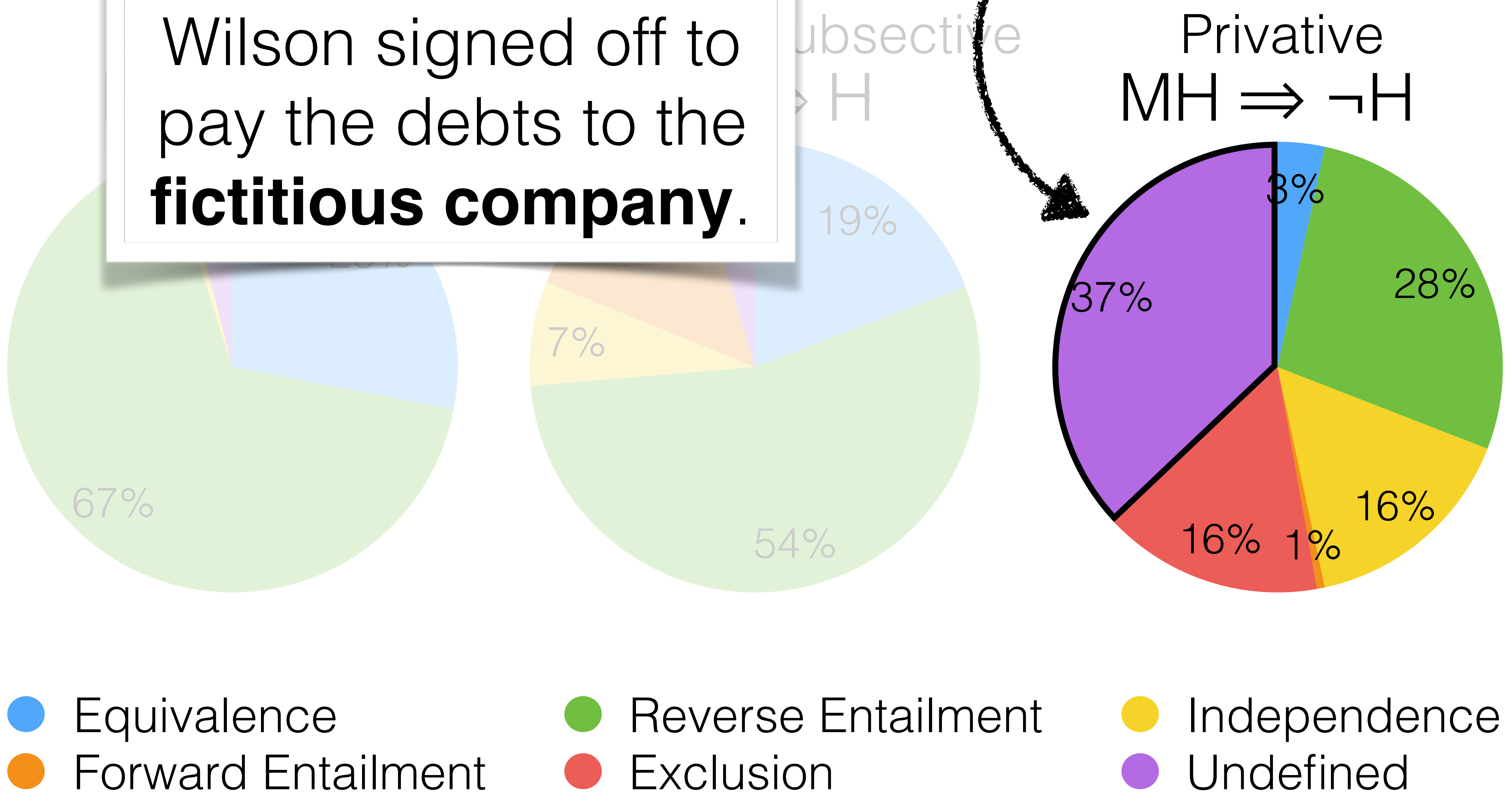
Bush travels Monday to Michigan to remark on the **Japanese economy**.

● Equivalence
● Forward Entailment
● Reverse Entailment
● Exclusion
● Independence
● Undefined

Classes of M

in fact this is how most privitives appear to behave...

Wilson signed off to pay the debts to the **fictitious company**.



ubsective

⇒ H

Privative

$MH \Rightarrow \neg H$

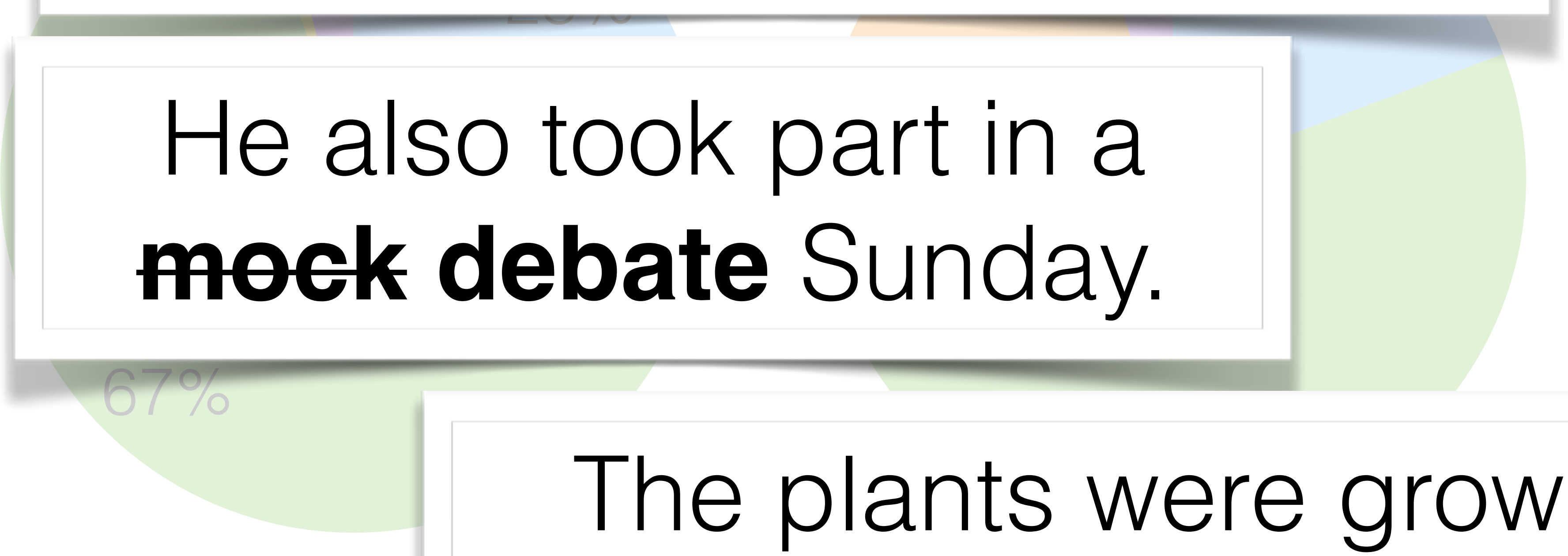
Classes

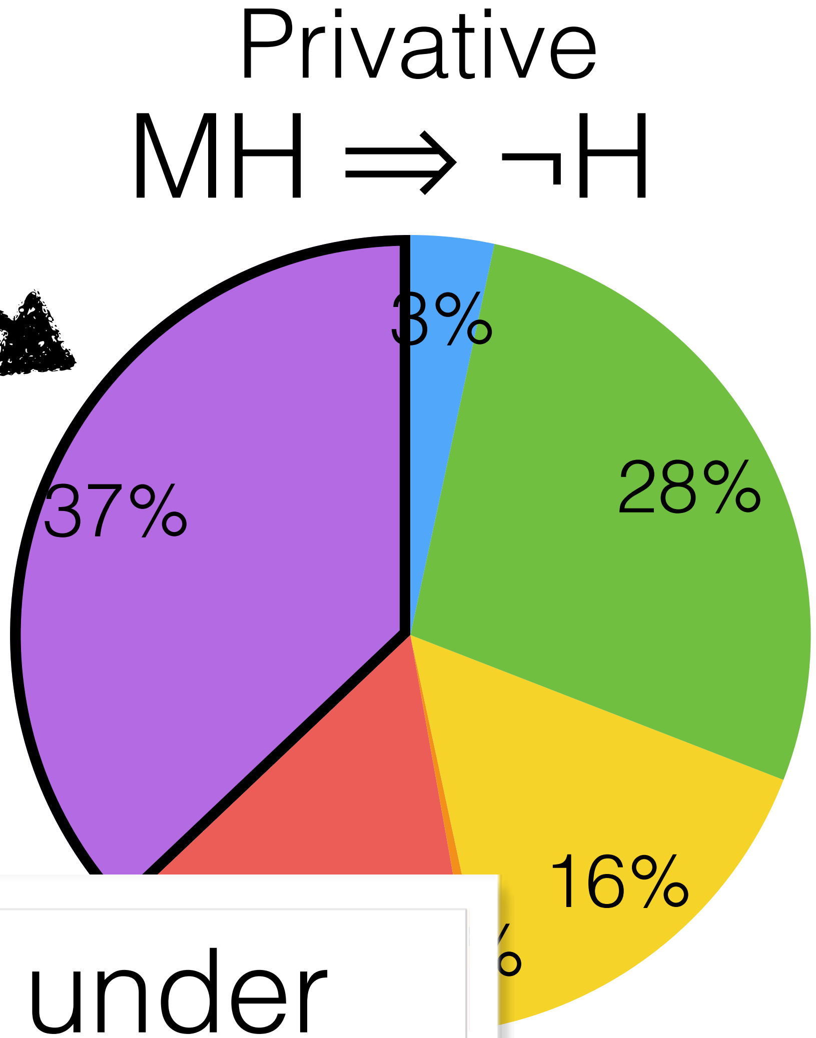
but in most cases,
deleting the adjective was
rated as okay/entailed

Flawed **~~counterfeit~~ software** can corrupt the information entrusted to it.

He also took part in a **~~mock~~ debate** Sunday.

The plants were grown under **~~artificial~~ light** and the whole operation was computerised.





● Equivalen

● Forward E

endence

ned

Takeaways

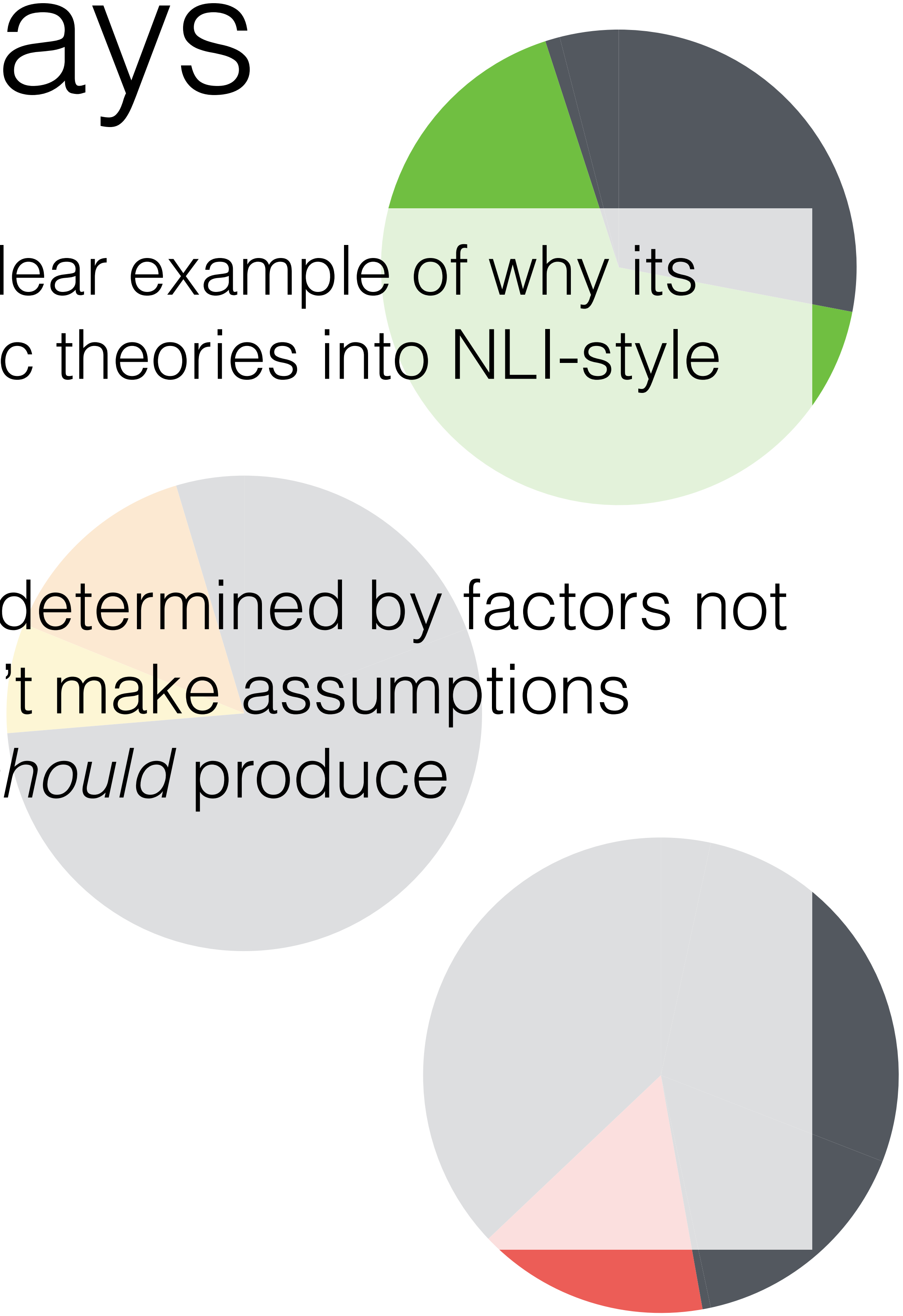
Takeaways

- Classes of modifiers provide a clear example of why its hard to naively translate semantic theories into NLI-style tasks

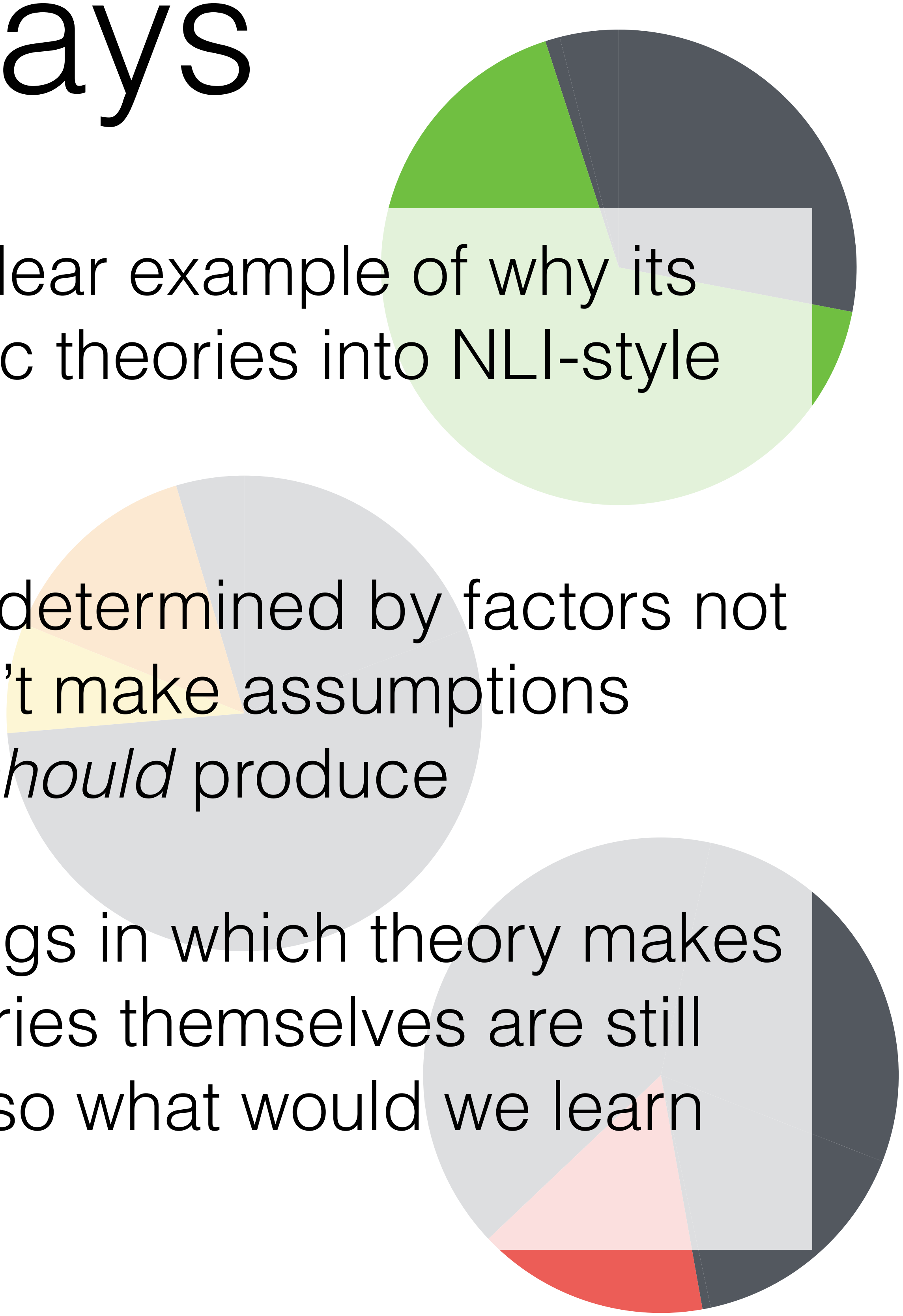
Takeaways

- Classes of modifiers provide a clear example of why its hard to naively translate semantic theories into NLI-style tasks

- Inferences “in practice” may be determined by factors not covered in the theory, so we can’t make assumptions about which labels our models *should* produce



Takeaways



- Classes of modifiers provide a clear example of why its hard to naively translate semantic theories into NLI-style tasks

- Inferences “in practice” may be determined by factors not covered in the theory, so we can’t make assumptions about which labels our models *should* produce

- We could constrain eval to settings in which theory makes correct predictions, but the theories themselves are still under study and under debate, so what would we learn from these evaluations?

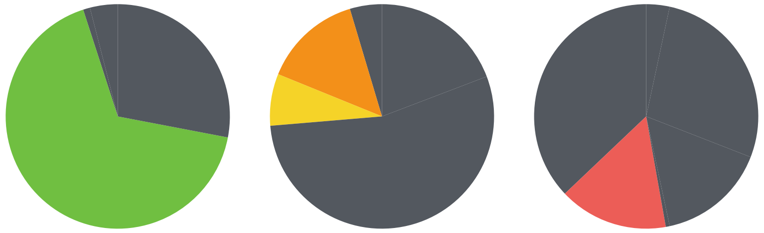
Three Case Studies

Modifier-Noun Composition

fake gun

↓

gun



Most babies are little and most problems are huge: Compositional Entailment in Adjective-Nouns. Pavlick and Callison-Burch (2016)

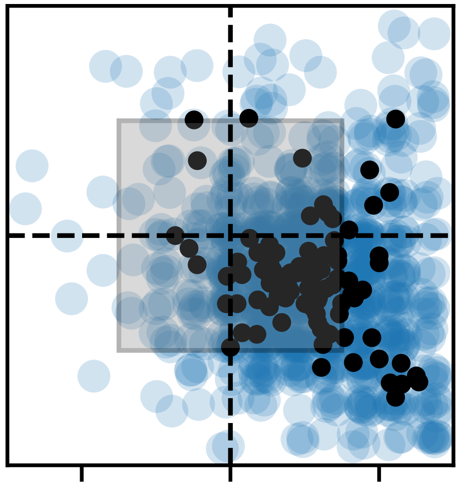
So-Called Nonsubsective Adjectives. Pavlick and Callison-Burch (2016)

Verb-Complement Composition

attempt to sing

↓

sing



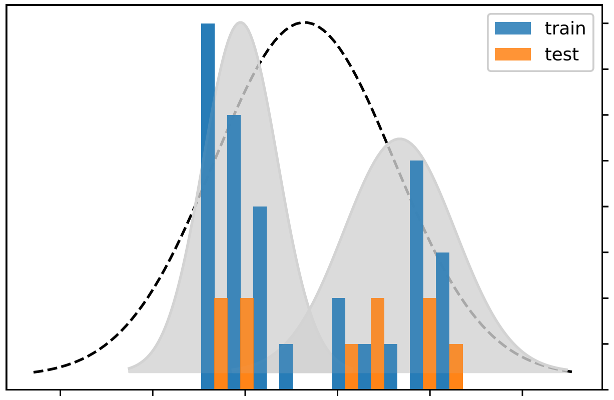
Do NLI models capture verb veridicality? Ross and Pavlick (2019)

Sentence-Level Inference

A man is standing under a tree

↓

A person is outside.



Inherent Disagreements in Human Textual Inferences. Pavlick and Kwiatkowski (2020)

Classes of Verbs

Classes of Verbs

They **know that** the answer is 5.

↓

The answer is 5.

Classes of Verbs

They **know that** the answer is 5.

+

↓

The answer is 5.

Classes of Verbs

They **know that** the answer is 5.

+ ↓

The answer is 5.

They **do not know that** the answer is 5.

+ ↓

The answer is 5.

Classes of Verbs

Positive Context	Negative Context	Example
+	+	They **know that** the answer is 5.

Classes of Verbs

They **managed to** get it right.
↓
They got it right.

+

They **did not manage to** get it right.
↓
They got it right.

−

Classes of Verbs

Positive Context	Negative Context	Example
+	+	They **know that** the answer is 5.
+	-	They **managed to** get it right.

Classes of Verbs

They **think that** the answer is 5.

o ↓

The answer is 5.

They **do not think that** the answer is 5.

o ↓

The answer is 5.

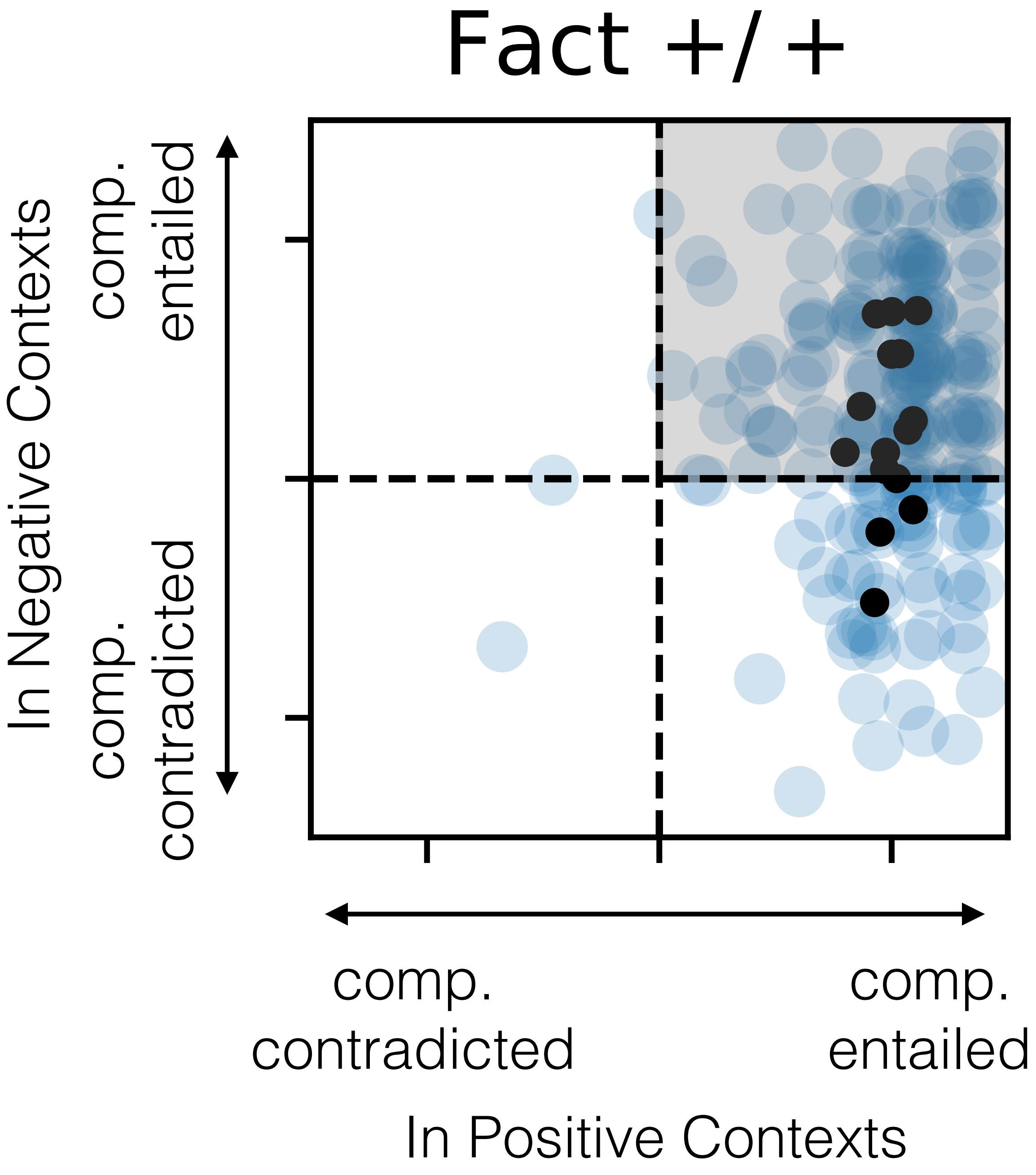
Classes of Verbs

Positive Context	Negative Context	Example
+	+	They **know that** the answer is 5.
+	-	They **managed to** get it right.
o	o	They **think that** the answer is 5.

Positive Context	Negative Context	Example
+	+	They **know that** the answer is 5.
+	-	They **managed to** get it right.
-	+	They **failed to** get it right.
o	+	They **suspect that** the answer is 5.
o	-	They **attempted to** get it right.
-	o	They **refused to** answer.
+	o	They **confirmed that** the answer is 5.
o	o	They **think that** the answer is 5.

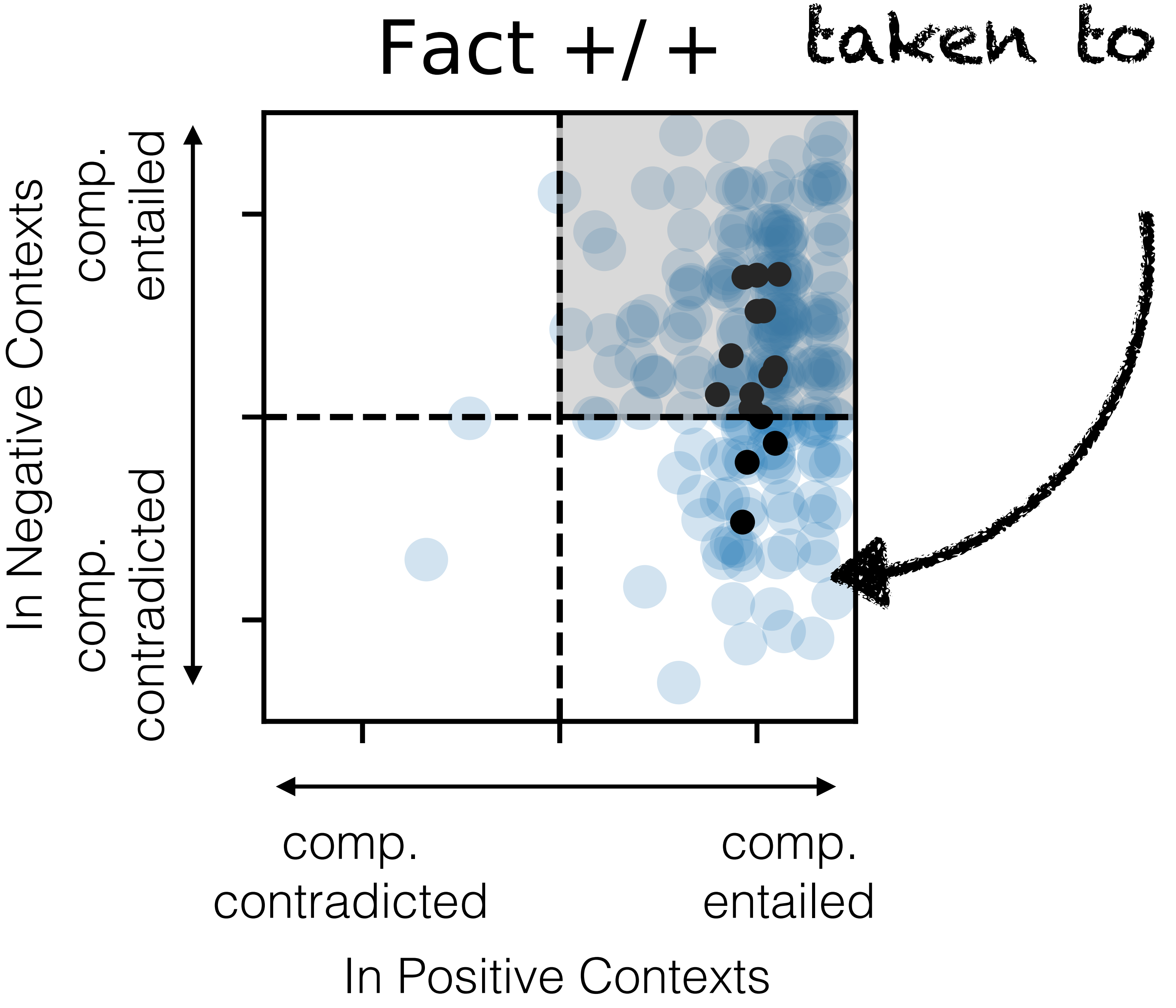
see: Karttunen (2012),
http://web.stanford.edu/group/csli_ lnr/Lexical_Resources/

Human Inferences



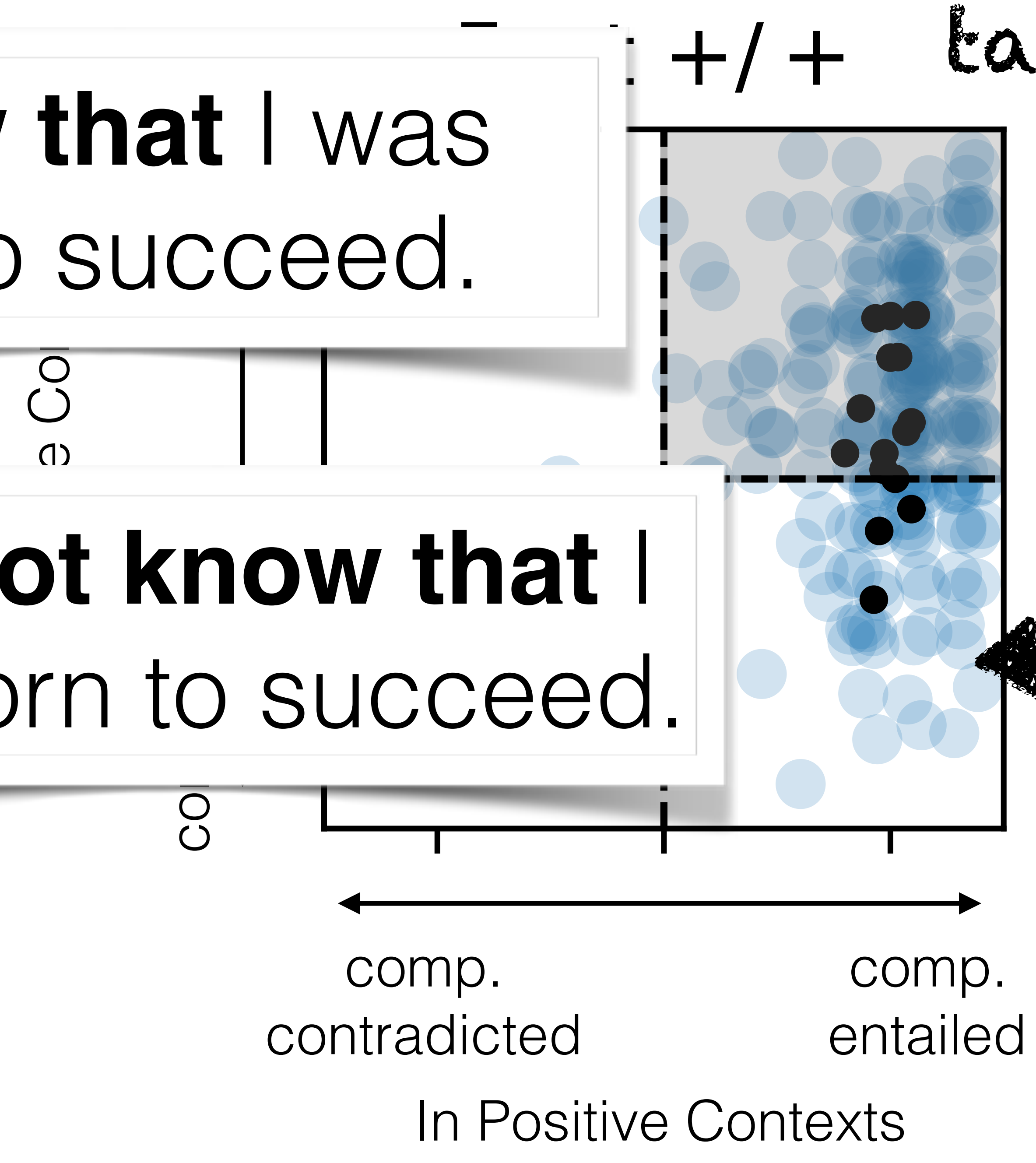
Human Infe

in many negative contexts, compliment is not taken to be true



Human Infe

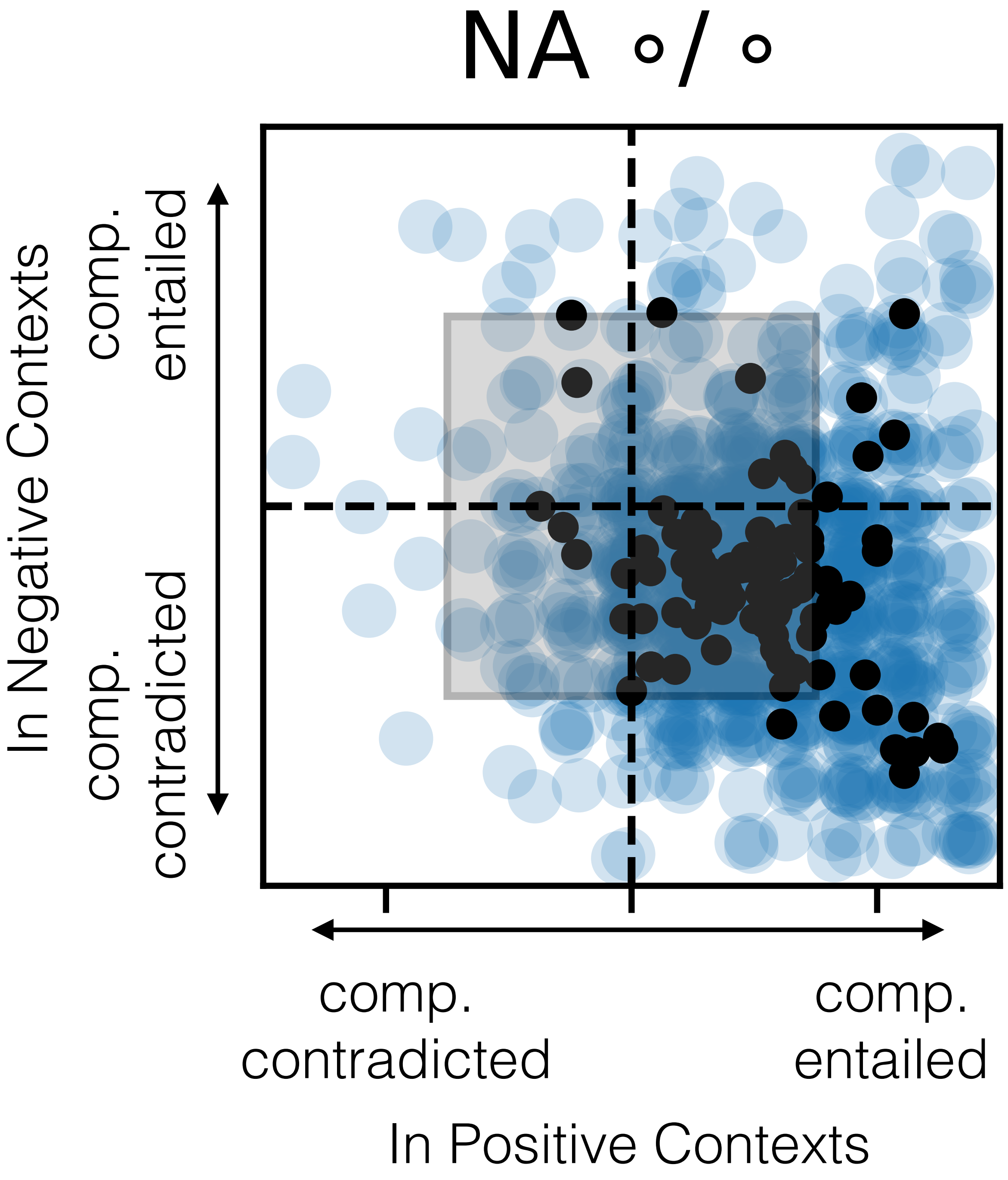
in many negative contexts, compliment is not taken to be true



+ I **know that** I was born to succeed.

o I **do not know that** I was born to succeed.

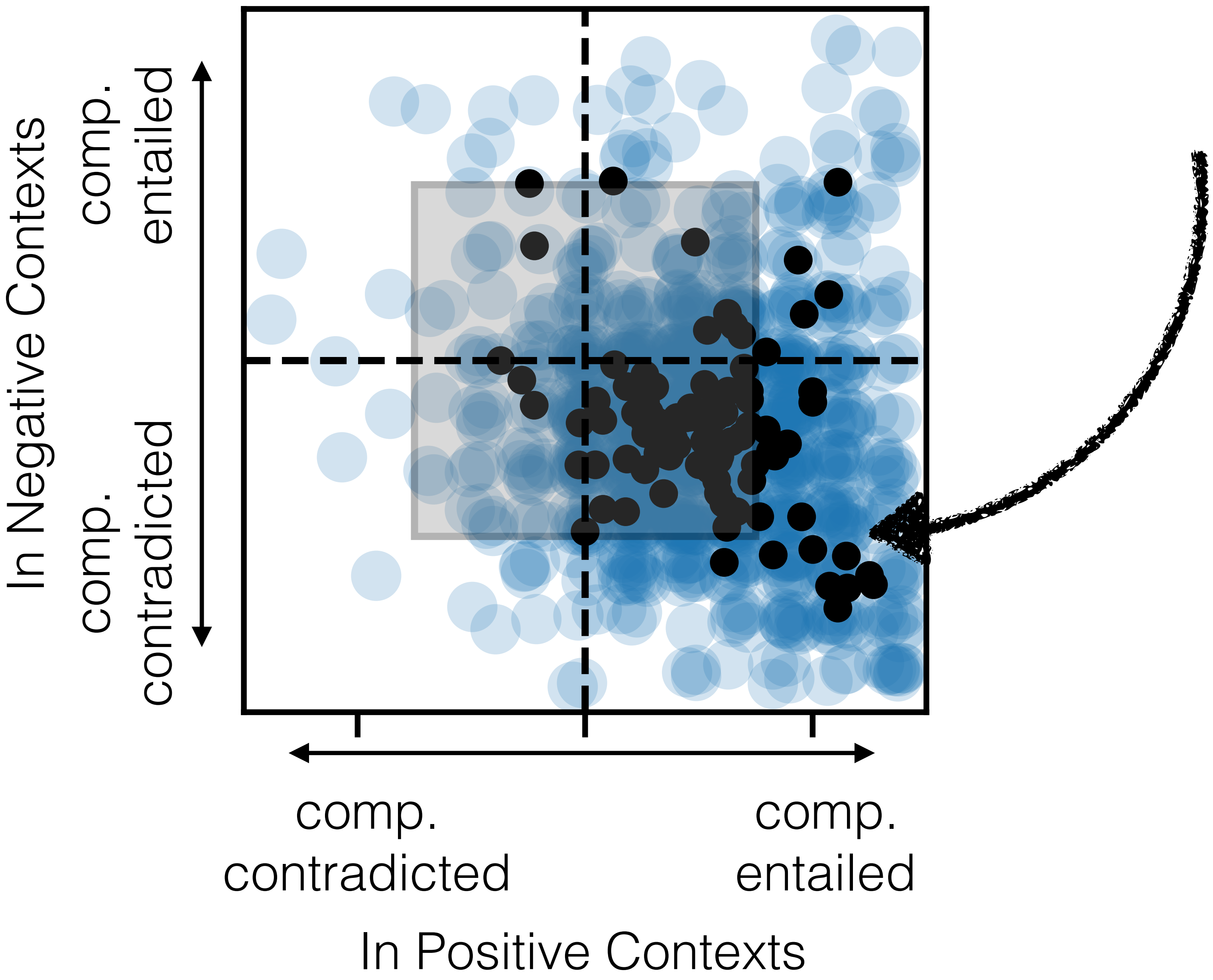
Human Inferences



Human Infe

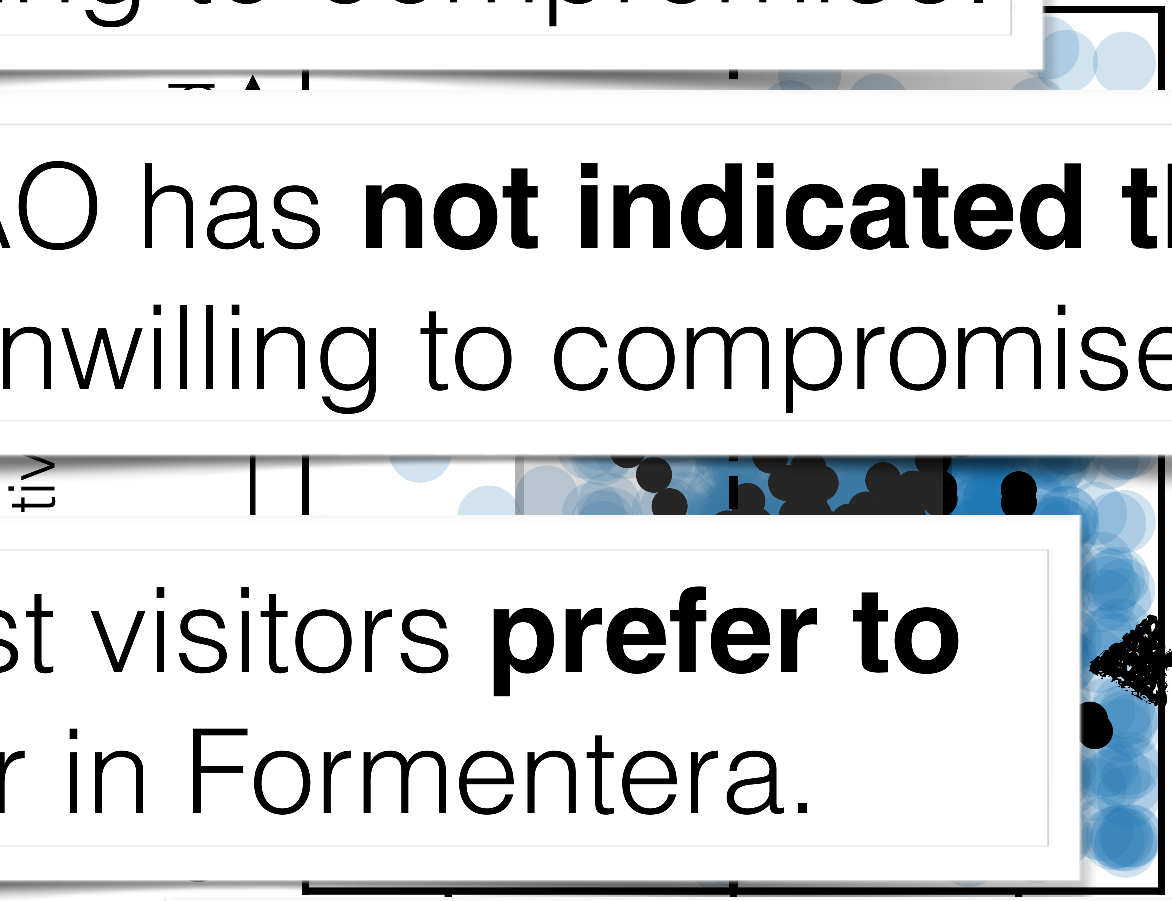
verbs often permit inferences, even when they arent "supposed to"

NA ∘/ ∘



Human Infe

verbs often permit inferences, even when they arent "supposed to"



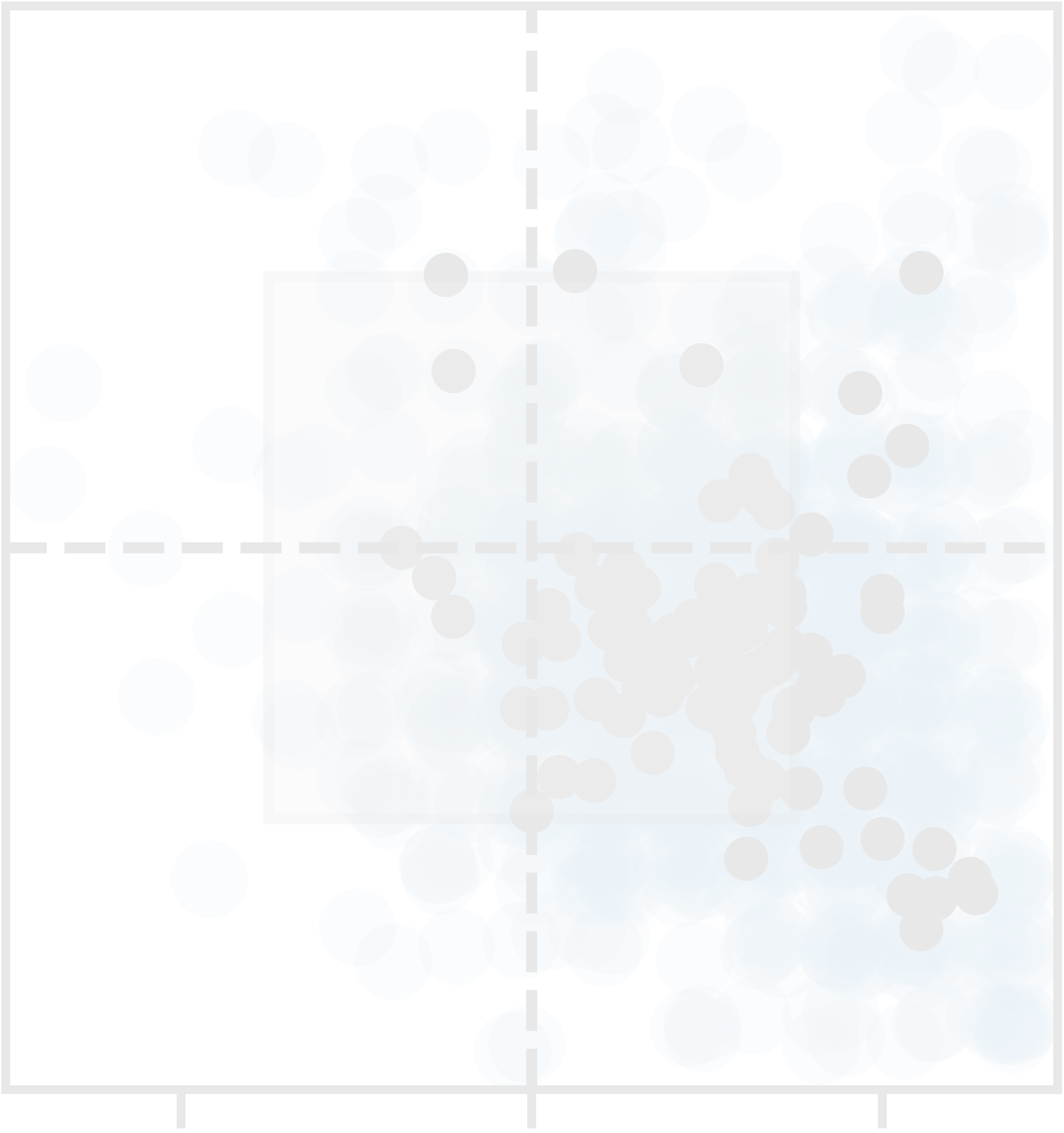
+ The GAO has **indicated that** it is unwilling to compromise.

− The GAO has **not indicated that** it is unwilling to compromise.

+ But most visitors **prefer to** linger in Formentera.

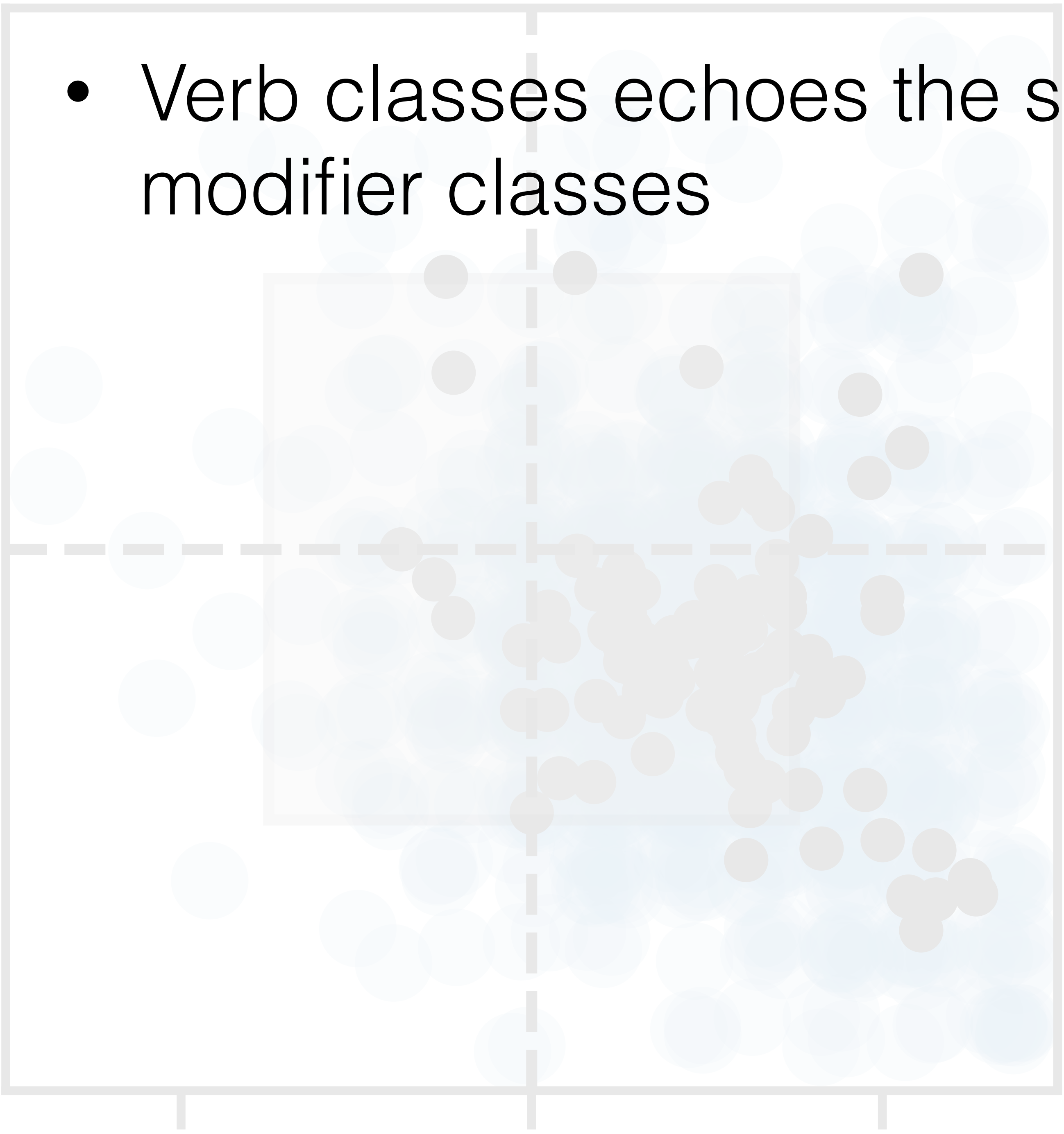
− But most visitors **do not prefer to** linger in Formentera.

Takeaways

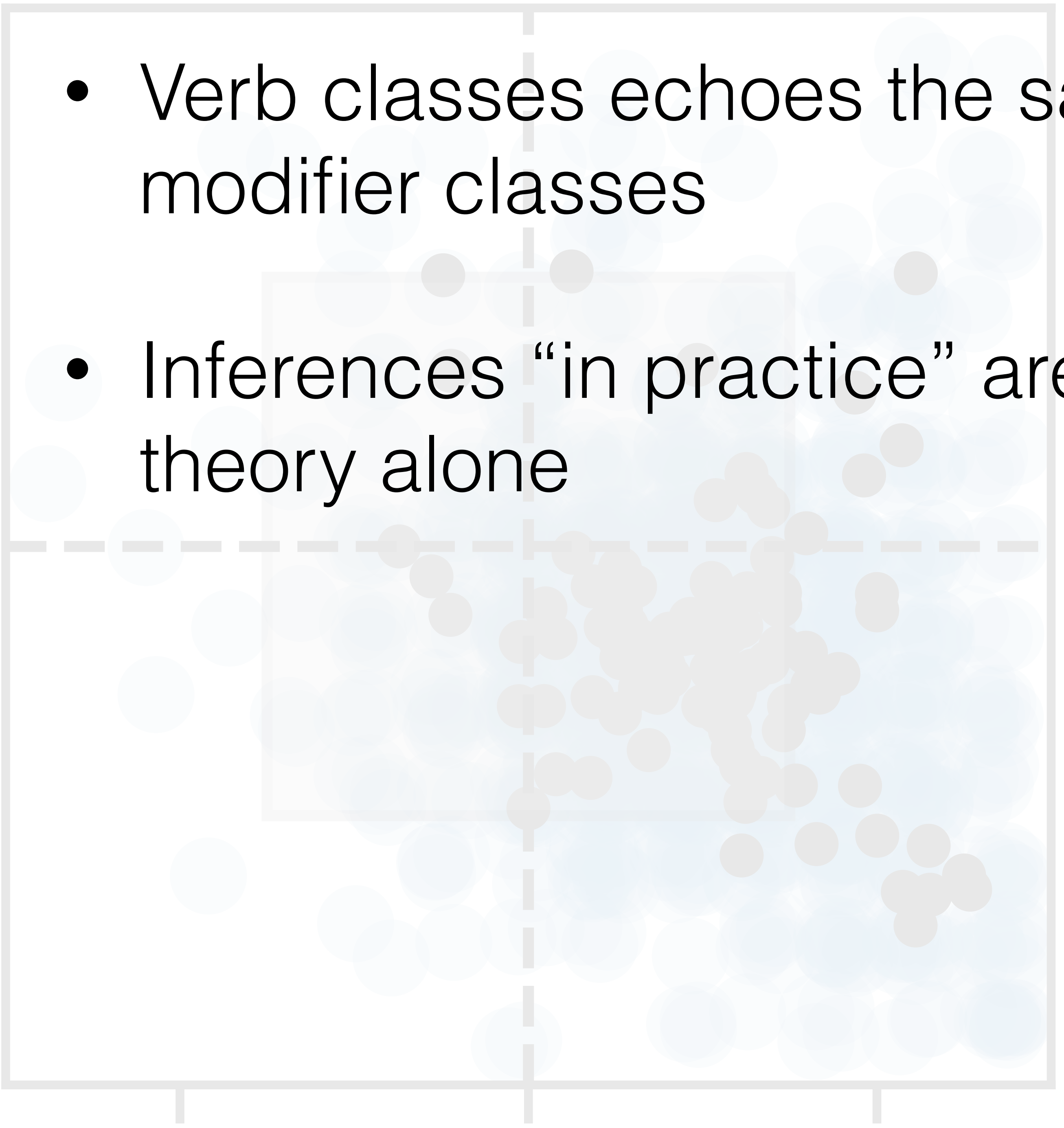


Takeaways

- Verb classes echoes the same themes seen with modifier classes

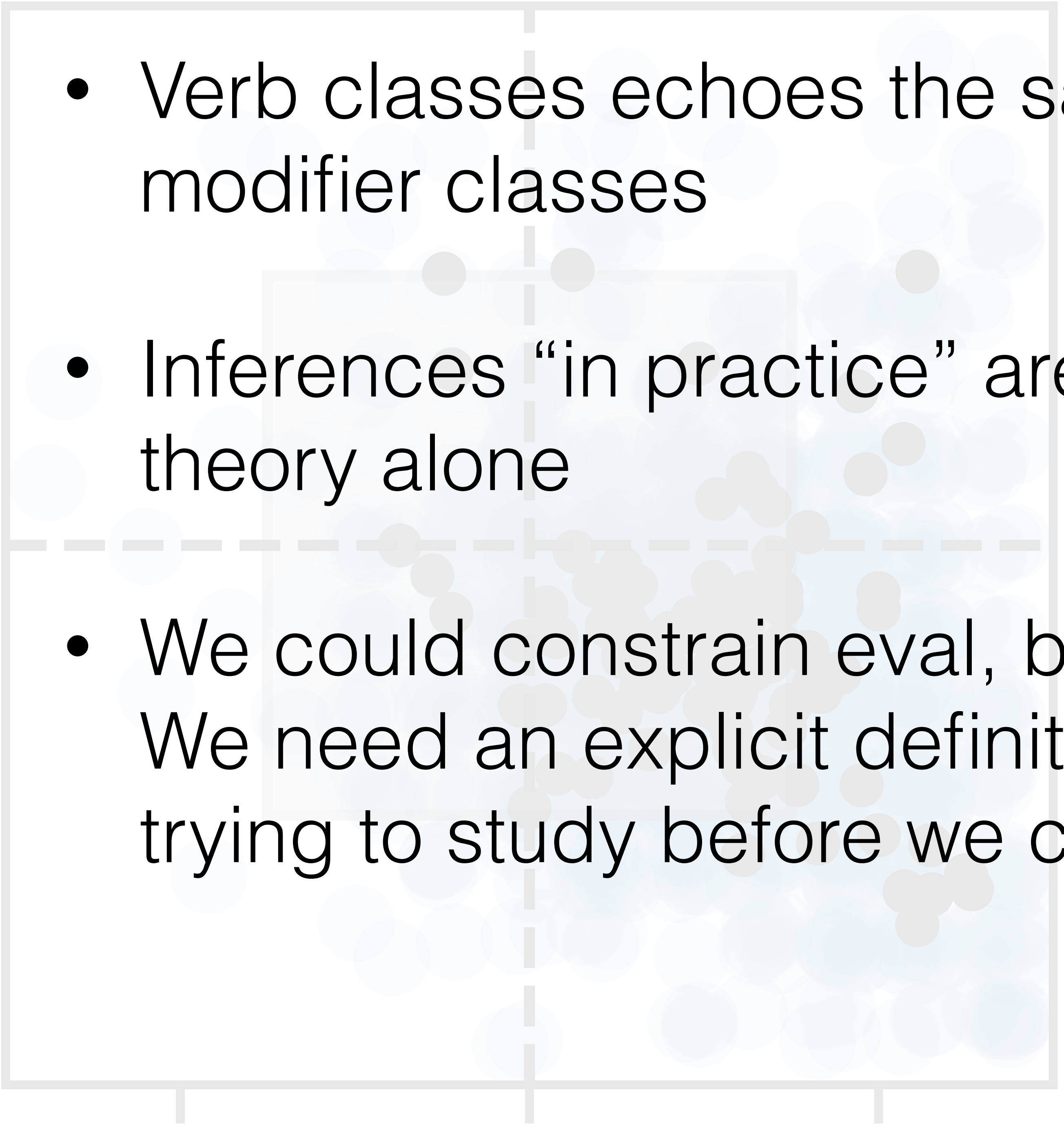


Takeaways



- Verb classes echoes the same themes seen with modifier classes
- Inferences “in practice” are not governed by the theory alone

Takeaways



- Verb classes echoes the same themes seen with modifier classes

- Inferences “in practice” are not governed by the theory alone

- We could constrain eval, but is this what we want? We need an explicit definition of what it is we are trying to study before we can define these tasks.

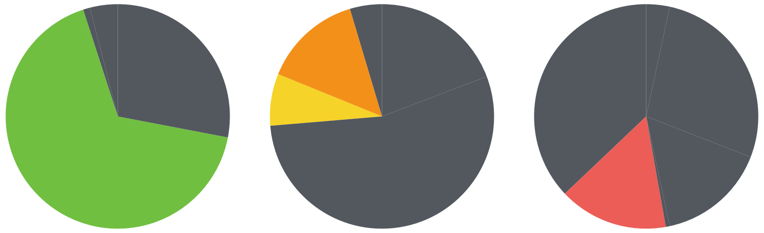
Three Case Studies

Modifier-Noun Composition

fake gun

↓

gun



Most babies are little and most problems are huge: Compositional Entailment in Adjective-Nouns. Pavlick and Callison-Burch (2016)

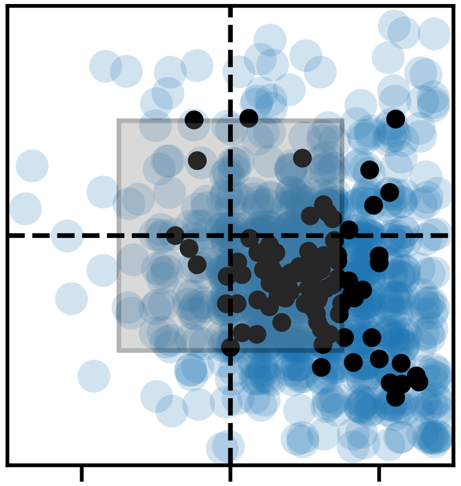
So-Called Nonsubsective Adjectives. Pavlick and Callison-Burch (2016)

Verb-Complement Composition

attempt to sing

↓

sing



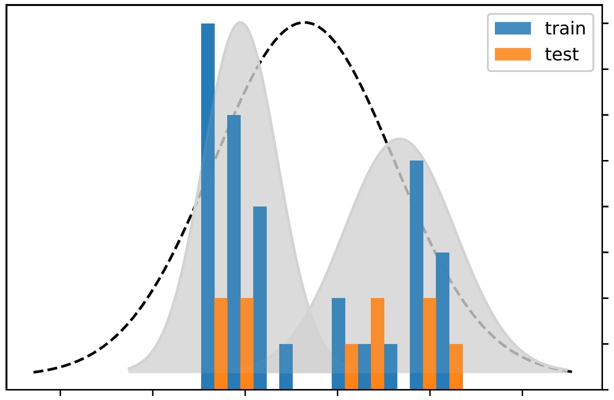
Do NLI models capture verb veridicality? Ross and Pavlick (2019)

Sentence-Level Inference

A man is standing under a tree

↓

A person is outside.



Inherent Disagreements in Human Textual Inferences. Pavlick and Kwiatkowski (2020)

Annotating “Ground Truth”

Annotating “Ground Truth”

A guy in a yellow shirt performs a balancing act on a taught chain near a canal.

⇓

A boy is doing a trick by water.

A young woman stands by a barbecue.

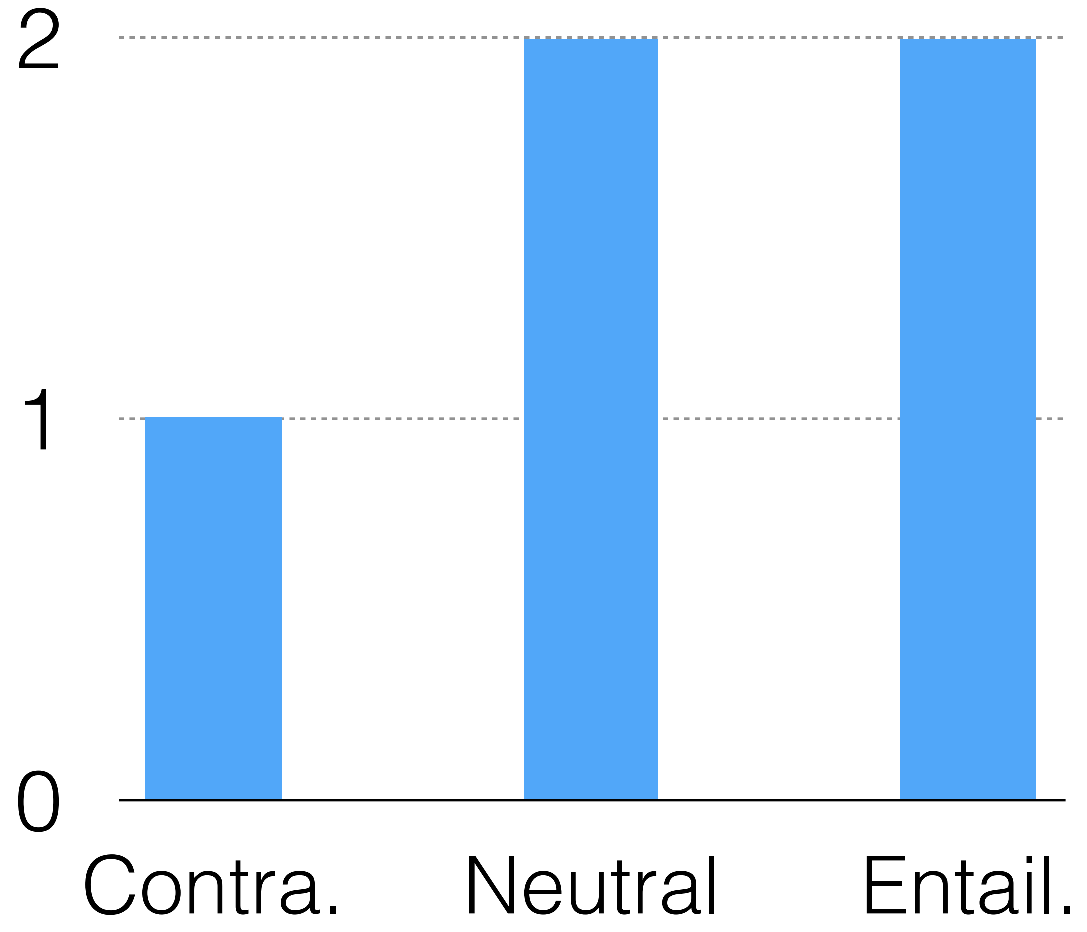
⇓

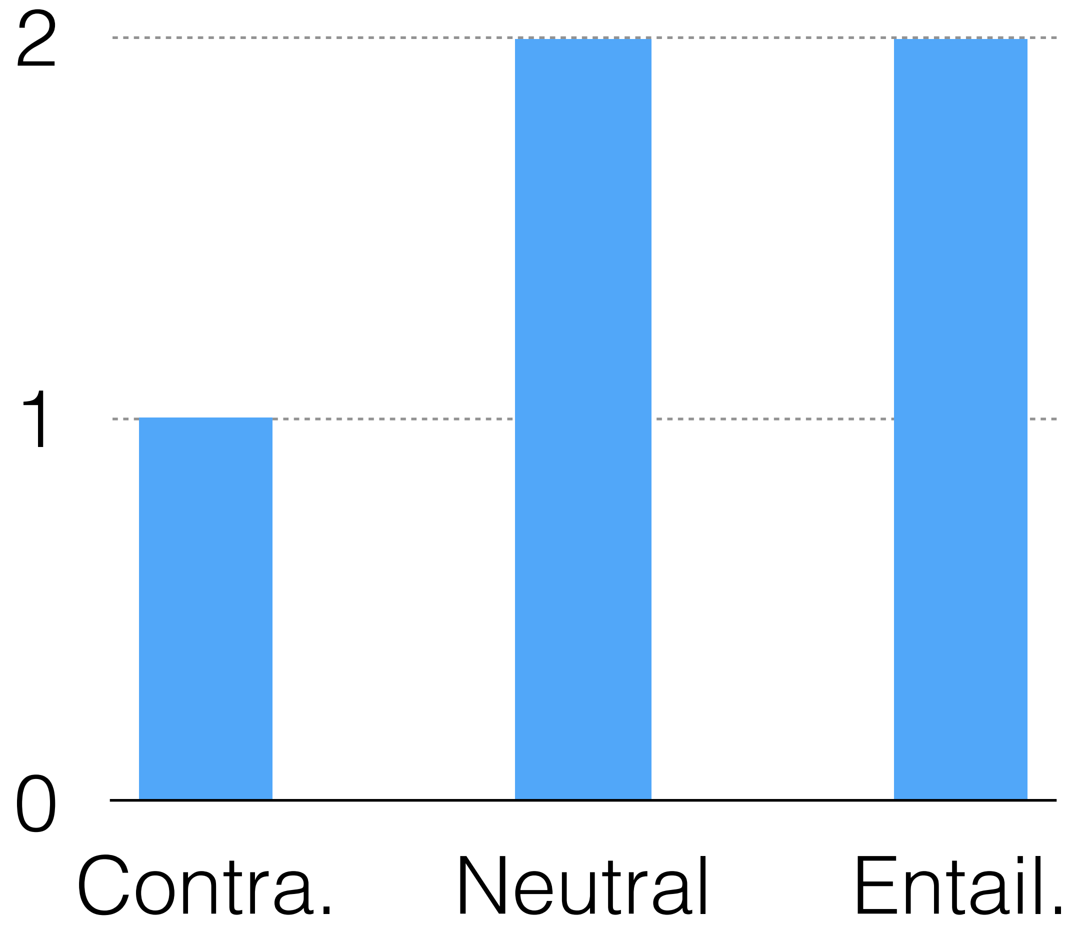
The young female is near a machine.

Annotating “Ground Truth”

A guy in a yellow shirt performs a balancing act on a taught chain near a canal.
⇓
A boy is doing a trick by water.

A young woman stands by a barbecue.
⇓
The young female is near a machine.

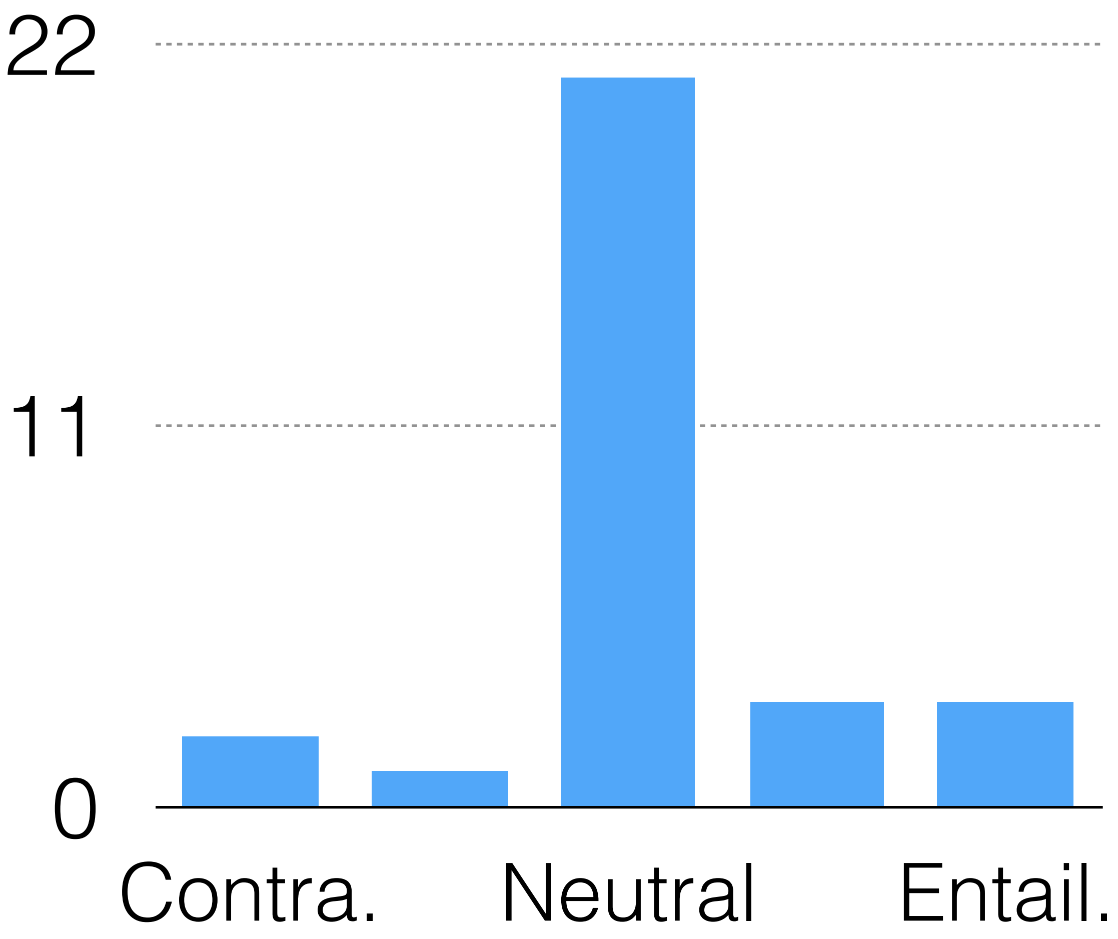




Annotating "Ground Truth"

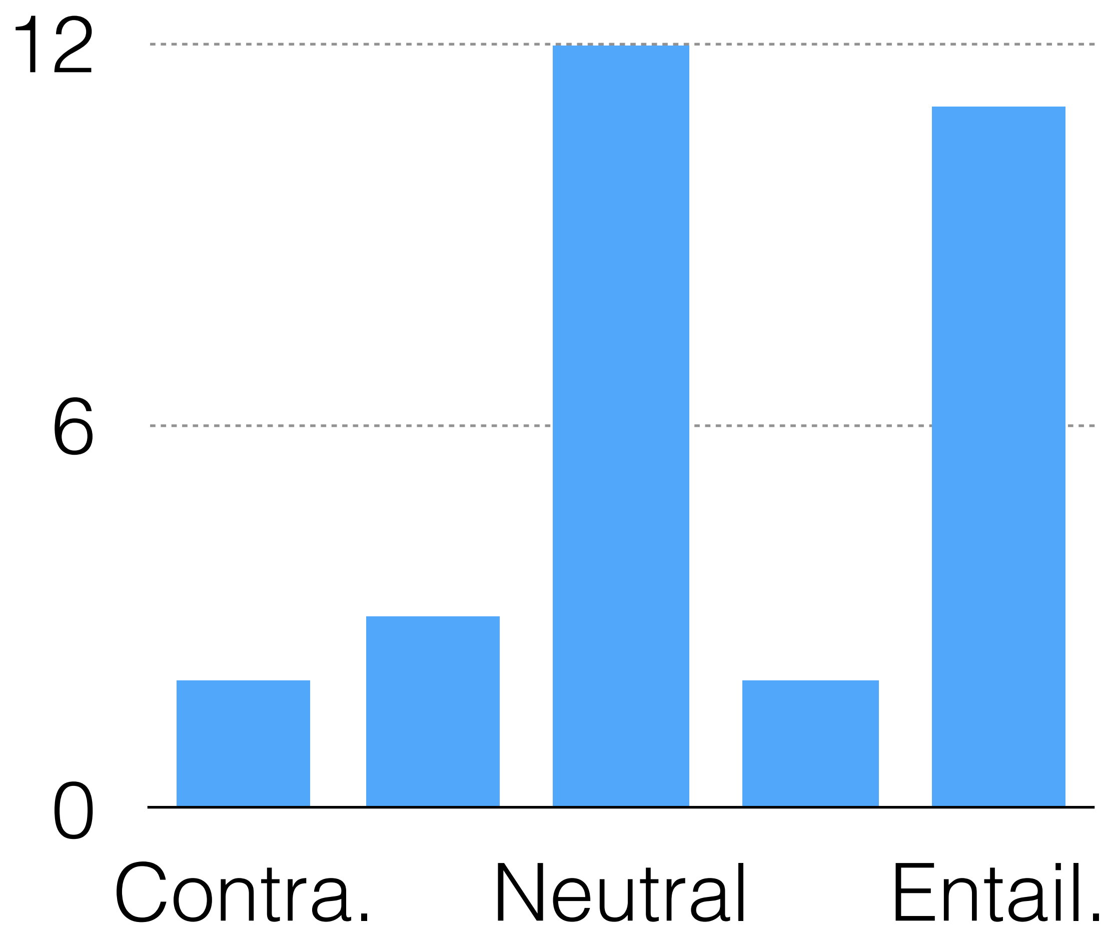
A guy in a yellow shirt performs a balancing act on a taught chain near a canal.

A boy is doing a trick by water.



A young woman stands by a barbecue.

The young female is near a machine.



Entailment Datasets

Stanford Natural Language Inference Dataset (SNLI)

+ Three dogs on a sidewalk → There are more than one dog here.

- A red rally car taking a slippery turn in a race → The car is stopped at a traffic light.

Multigenre Natural Language Inference Dataset (MNLI)

+ Historical heritage is very much the theme in Ichidani → Ichidani's historical heritage is important.

- okay i uh i have five children altogether → I do not have any children.

Recognizing Textual Entailment II (RTE2)

+ Self-sufficiency has been turned into a formal public awareness campaign in San Francisco, by Mayor Gavin Newsom. → Gavin Newsom is a politician of San Fransisco.

- The unconfirmed case concerns a rabies-like virus known only in bats → A case of rabies was confirmed.

Johns Hopkins Ordinal Common Sense Inference (JOCI)

+ It was Charlie's first day of work at the new firm. → The firm is a business.

- A young girl is holding her teddy bear while riding a pony. → The bear attacks.

Diverse Natural Language Inference Corpus (DNC)

+ Tony bent the rod. → Tony caused the bending.

- When asked about the restaurant, Jonah said “sauce was tasteless”. → Jonah liked the restaurant.

Entailment Datasets

Stanford Natural Language Inference Dataset (SNLI)

+ Three dogs on a sidewalk → There are more than one dog here.

- A red rally car taking a slippery turn in a race → The car is stopped at a traffic light.

Multigenre Natural Language Inference Dataset (MNLI)

+ Historical heritage is very much the theme in Ichidani → Ichidani’s historical heritage is important.

- 50 ratings each
- Continuous scale (-50 to 50)
- z-normalized by annotator (min 20 ratings each)

was confirmed.

Johns Hopkins Ordinal Common Sense Inference (JOCI)

+ It was Charlie’s first day of work at the new firm. → The firm is a business.

- A young girl is holding her teddy bear while riding a pony. → The bear attacks.

Diverse Natural Language Inference Corpus (DNC)

+ Tony bent the rod. → Tony caused the bending.

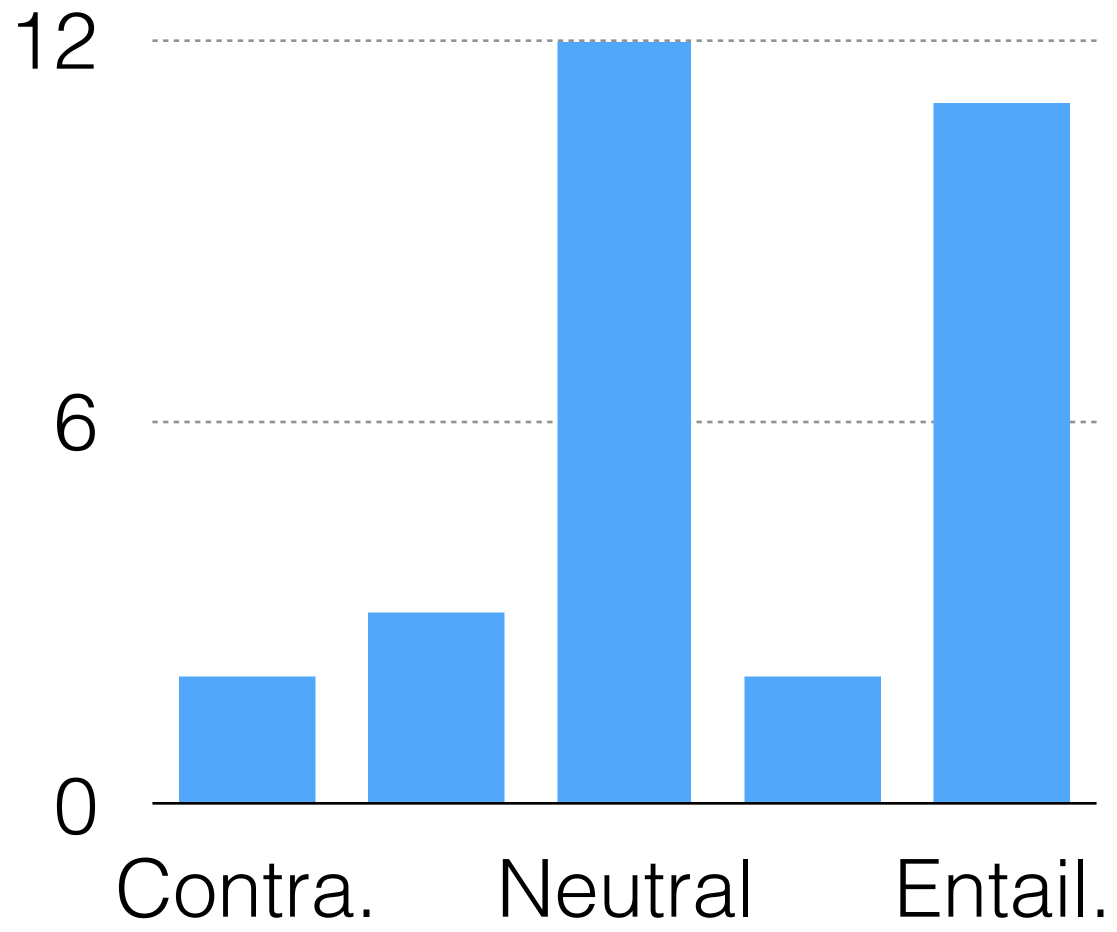
- When asked about the restaurant, Jonah said “sauce was tasteless”. → Jonah liked the restaurant.

Simple Gaussian Mixture Models

Simple Gaussian Mixture Models

A young woman stands by a barbecue.

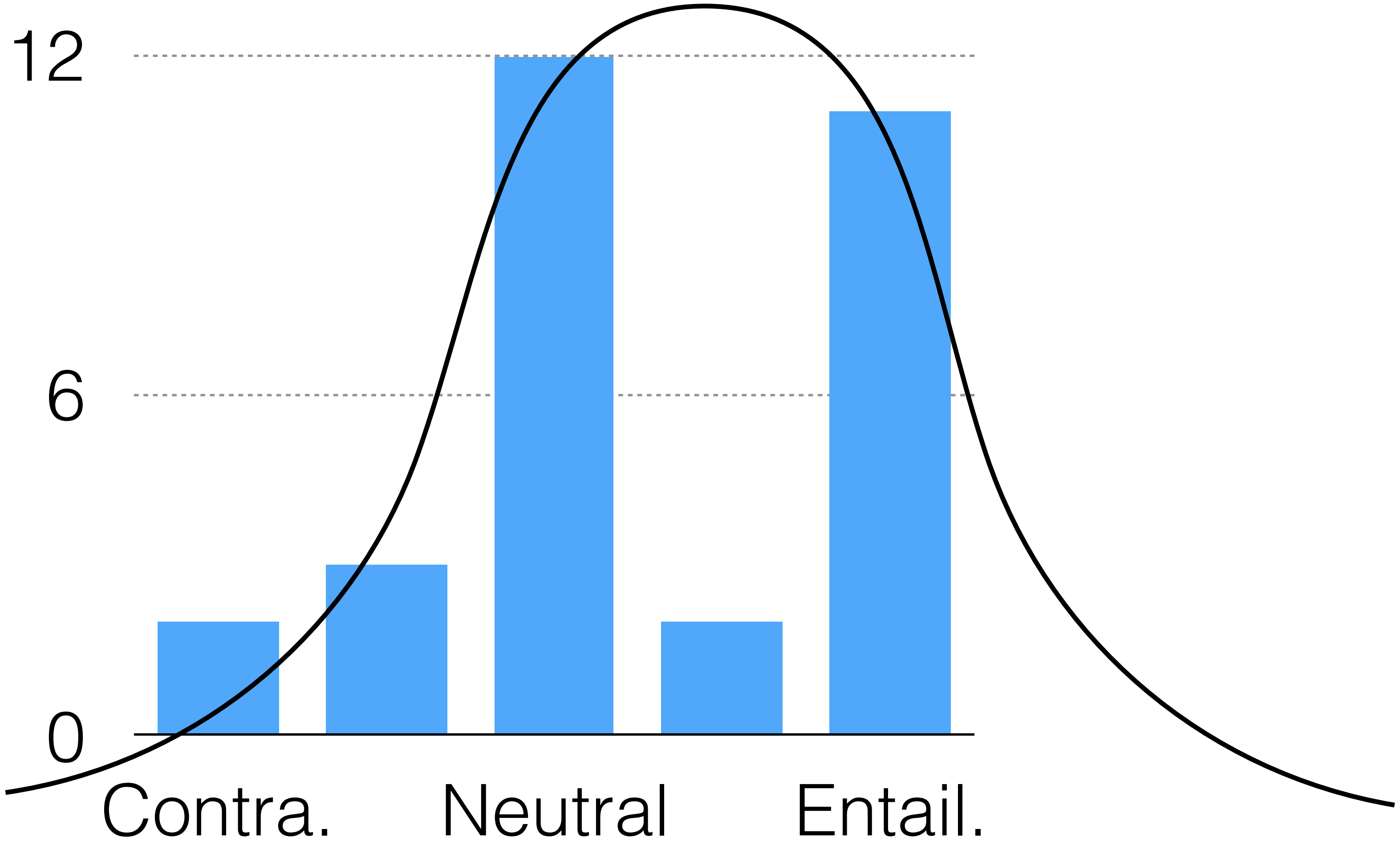
The young female is near a machine.



Simple Gaussian Mixture Models

A young woman stands by a barbecue.

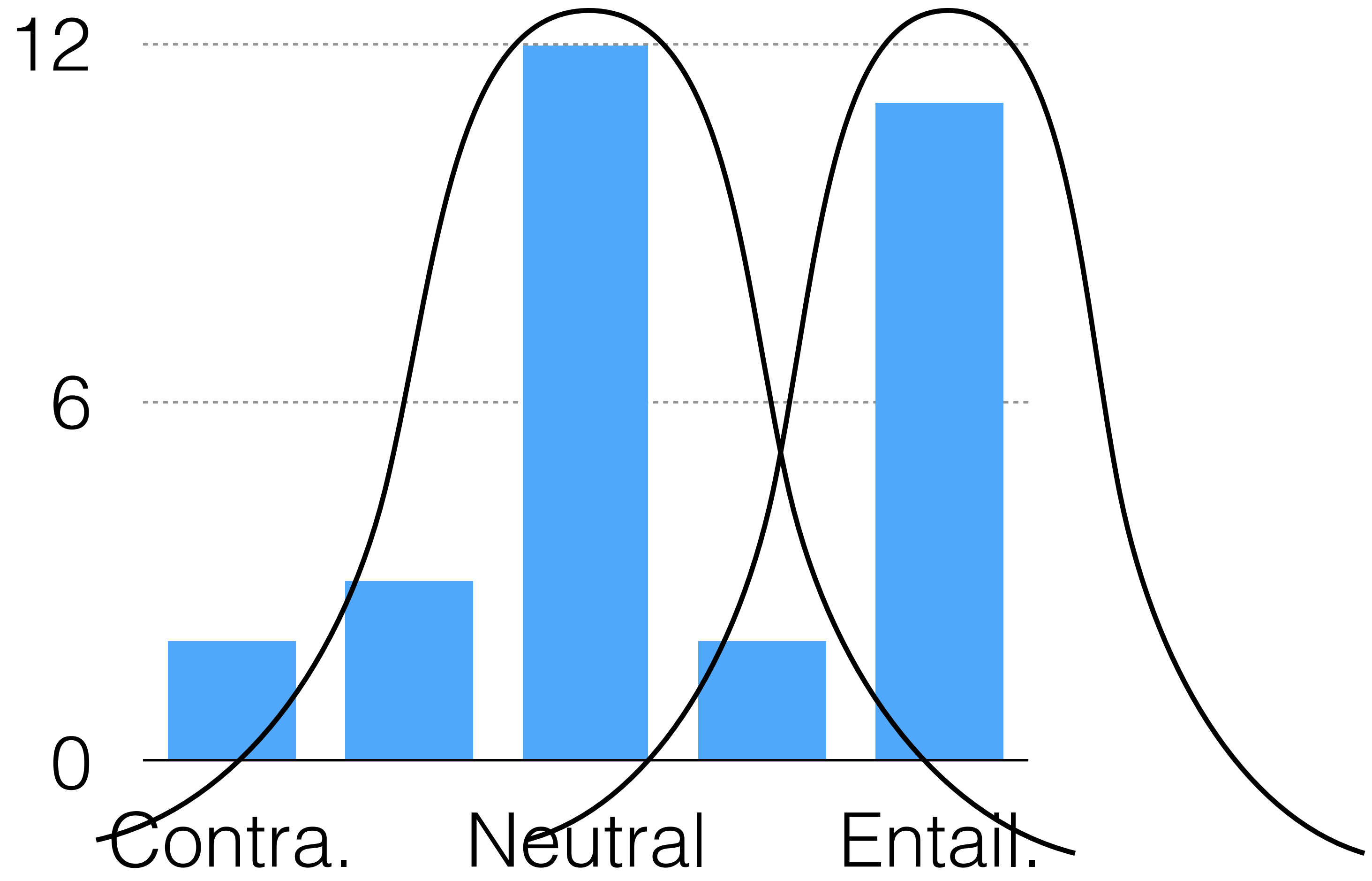
The young female is near a machine.



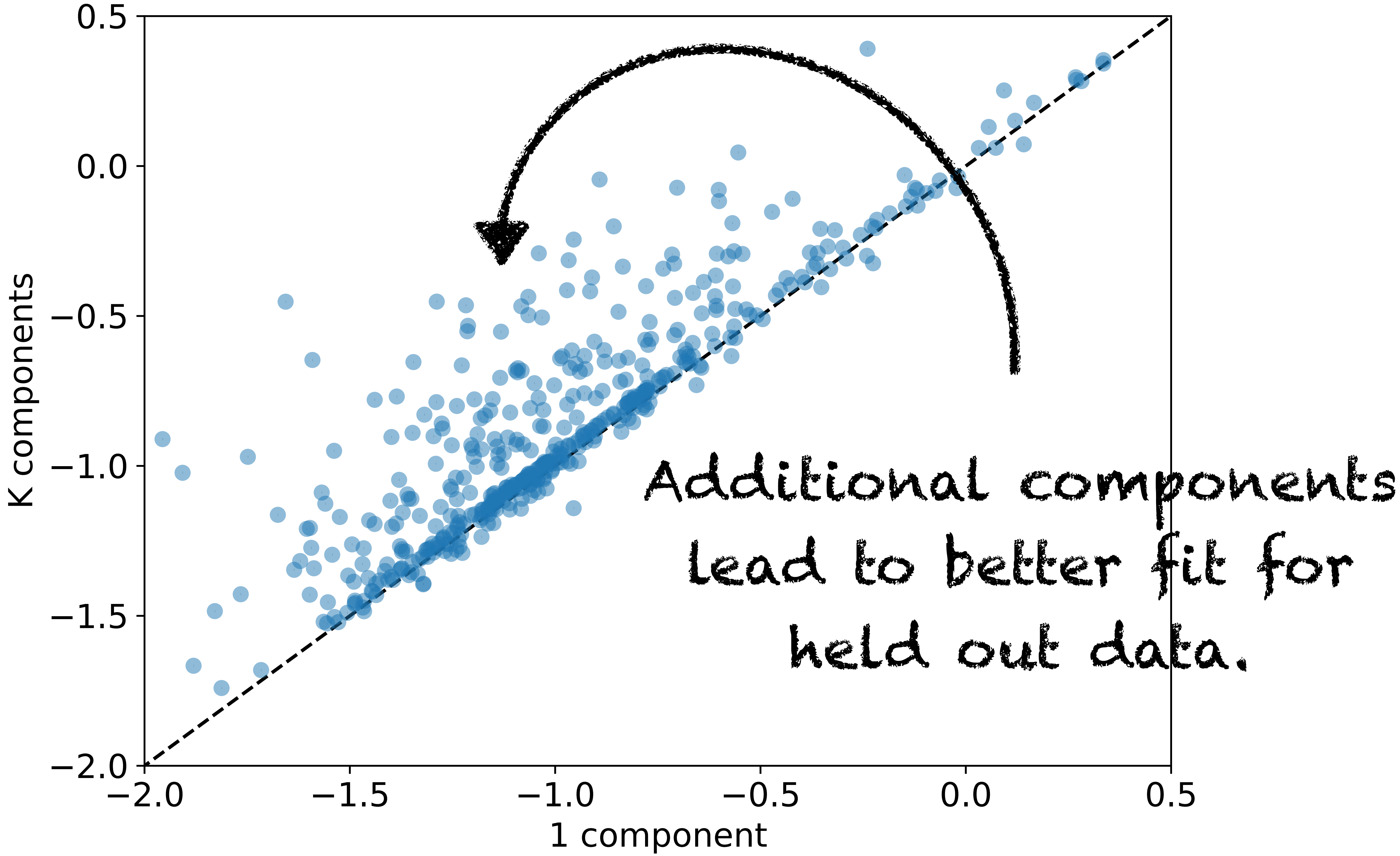
Simple Gaussian Mixture Models

A young woman stands by a barbecue.

The young female is near a machine.

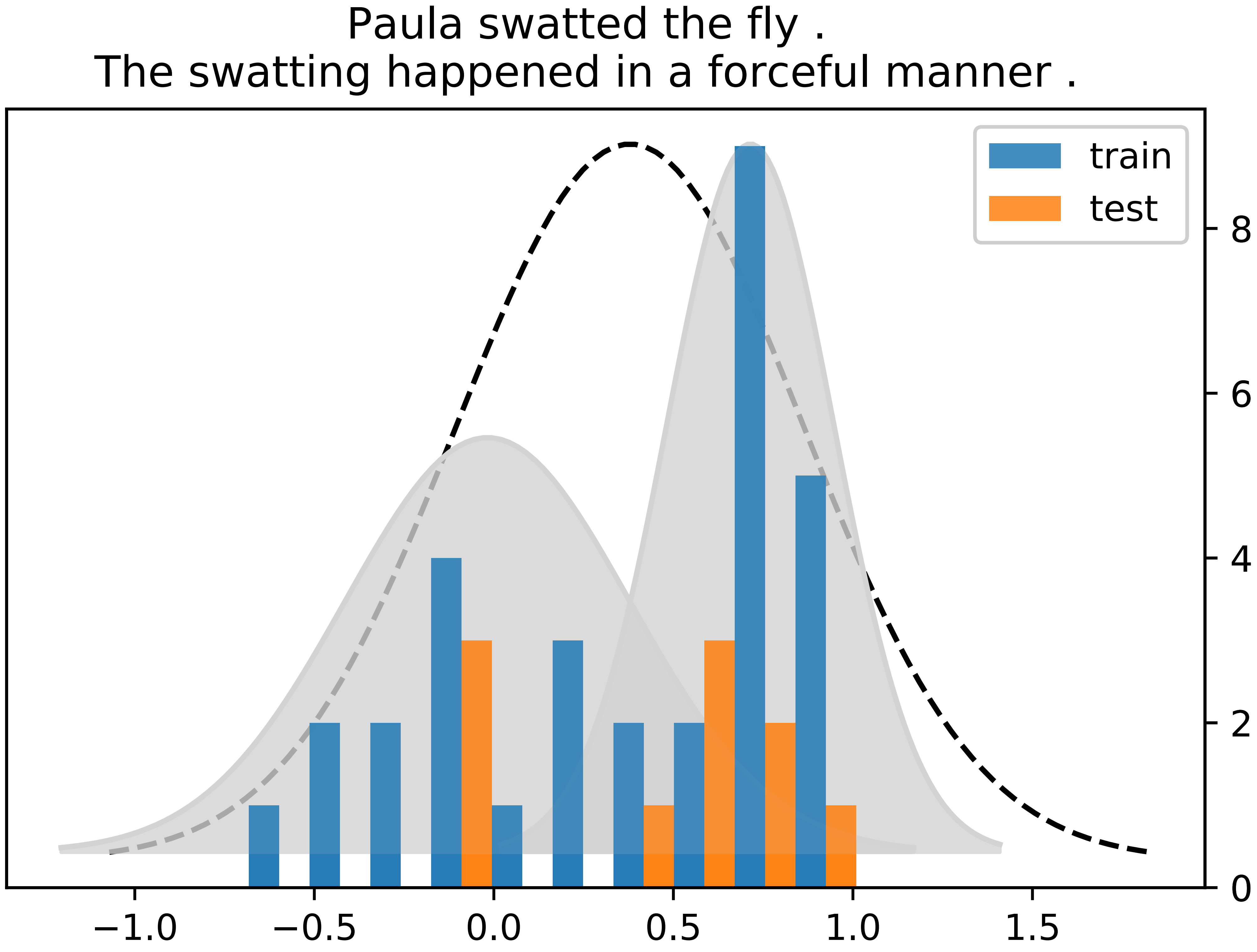


Simple Gaussian Mixture Models



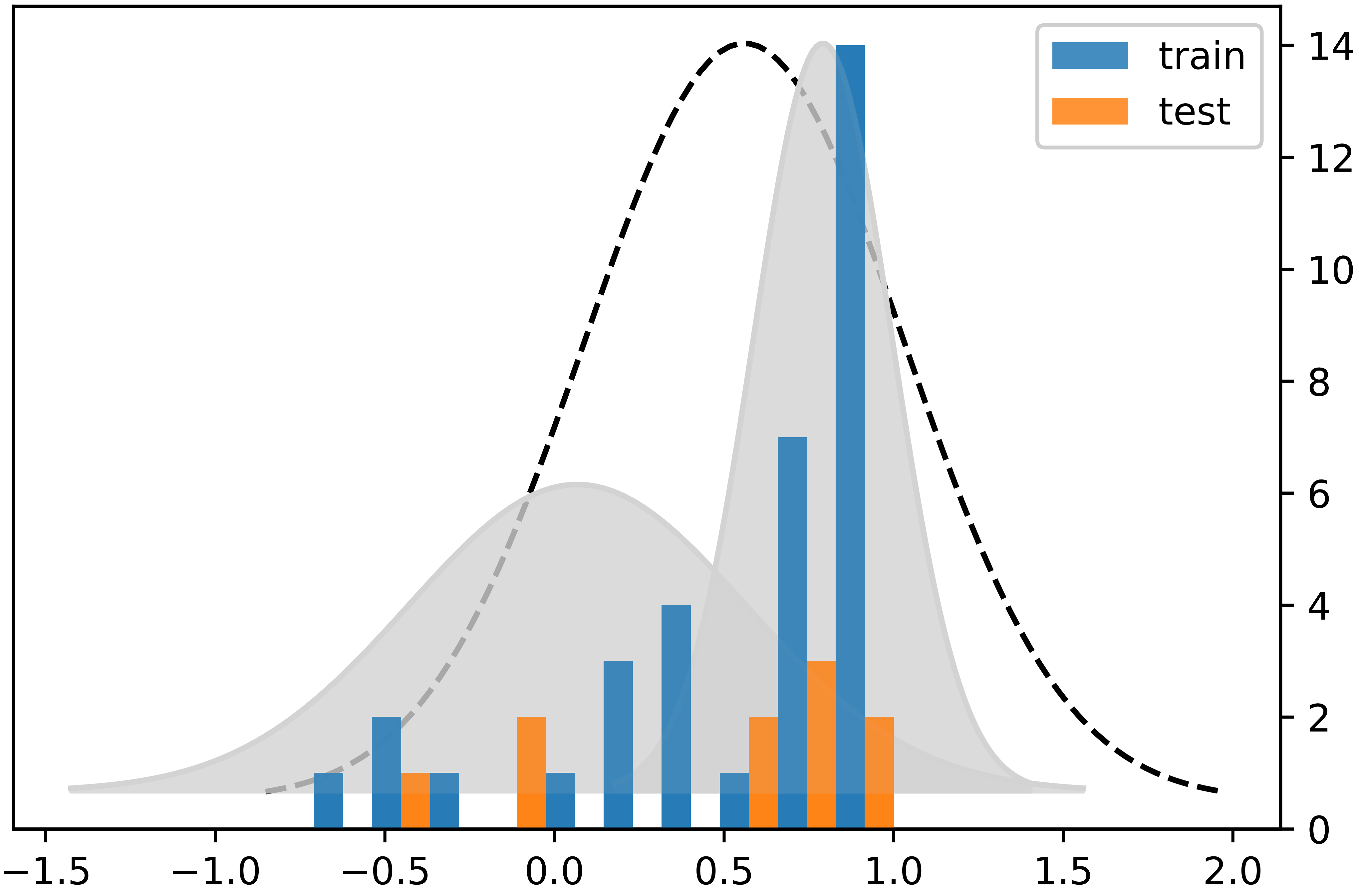
Additional components lead to better fit for held out data.

Simple Gaussian Mixture Models



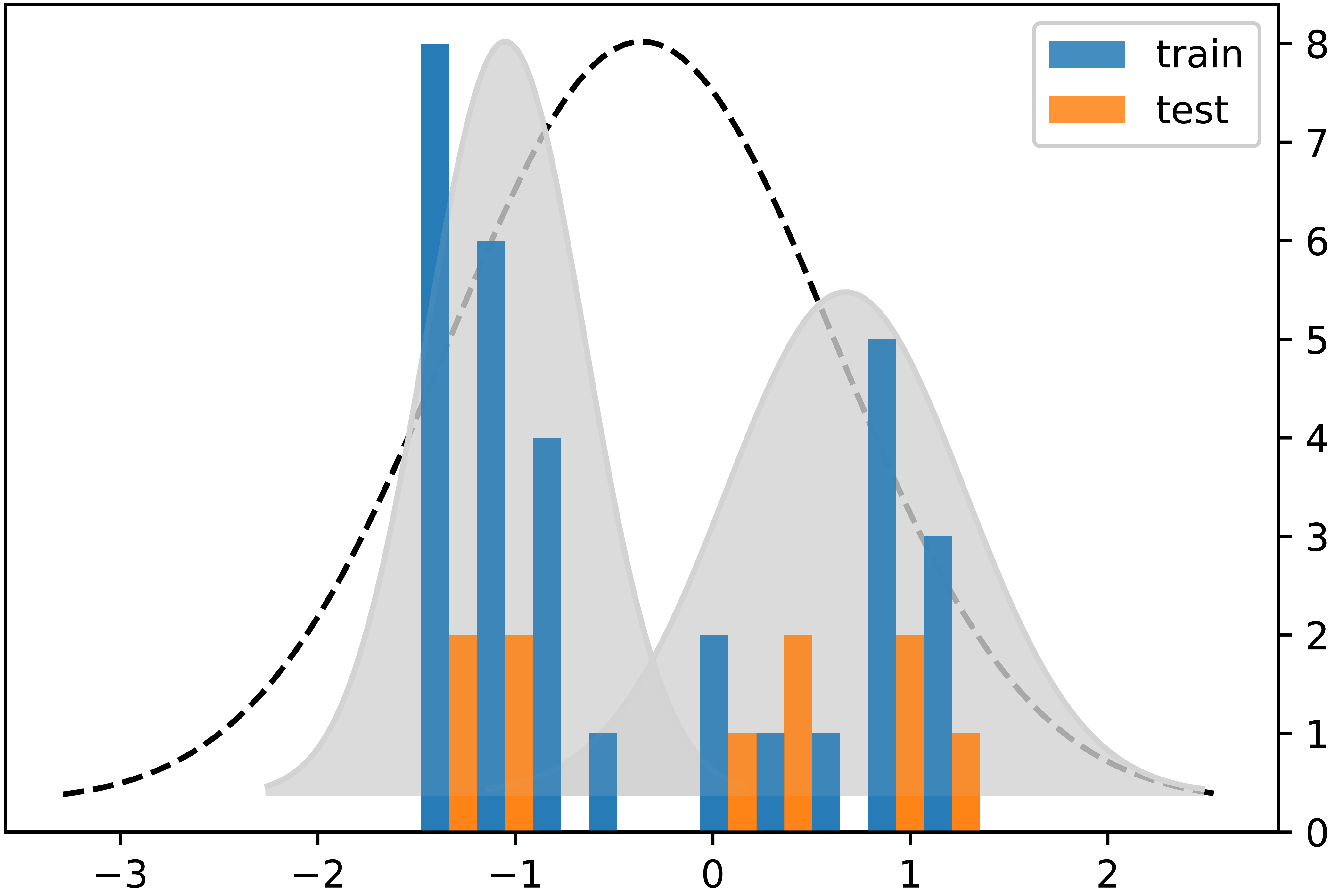
Simple Gaussian Mixture Models

someone confessed that a particular thing happened .
that thing happened .

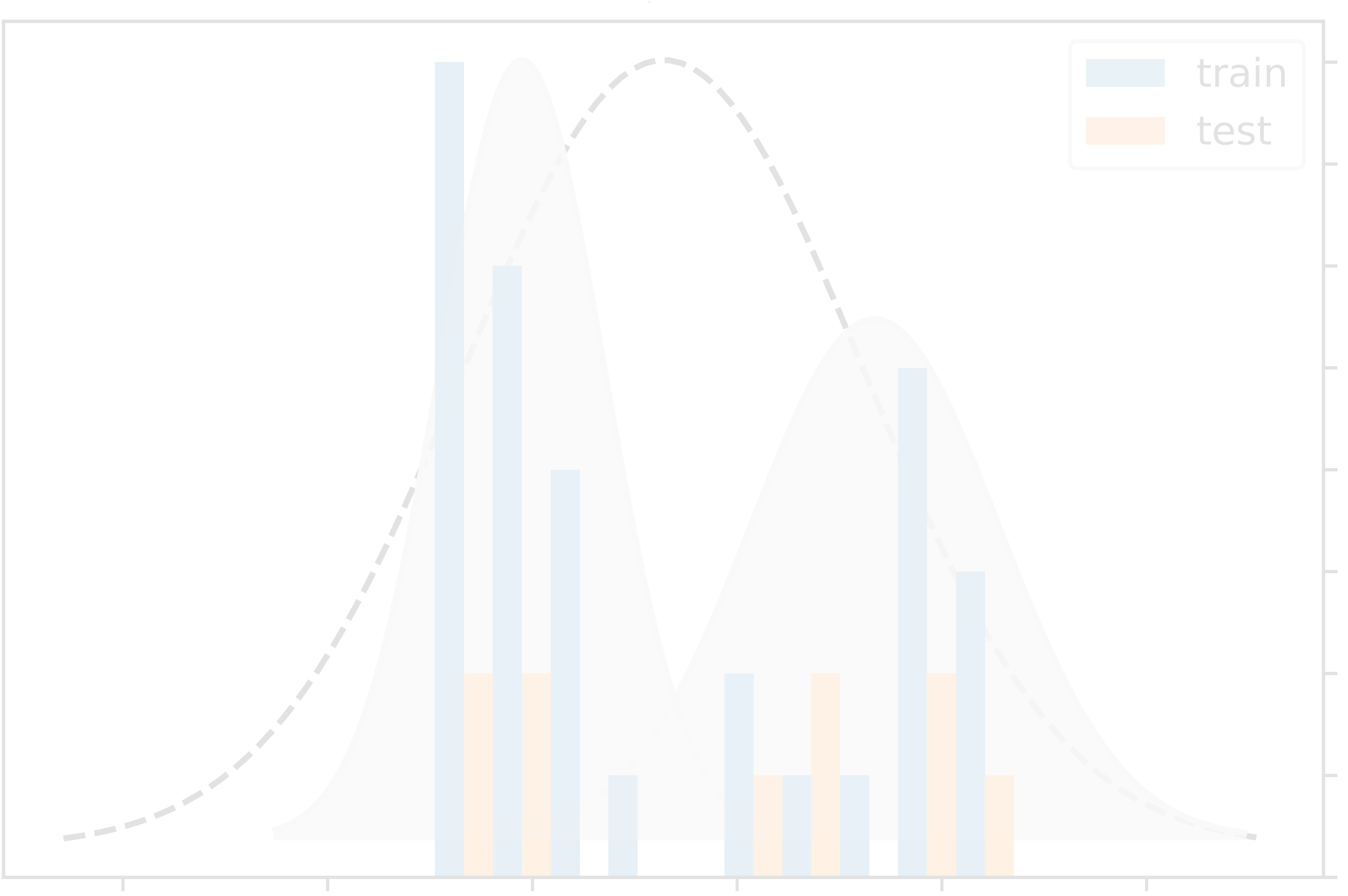


Simple Gaussian Mixture Models

The capital of Slovenia is Ljubljana, with 270,000 inhabitants.
Slovenia has 270,000 inhabitants.



Takeaways



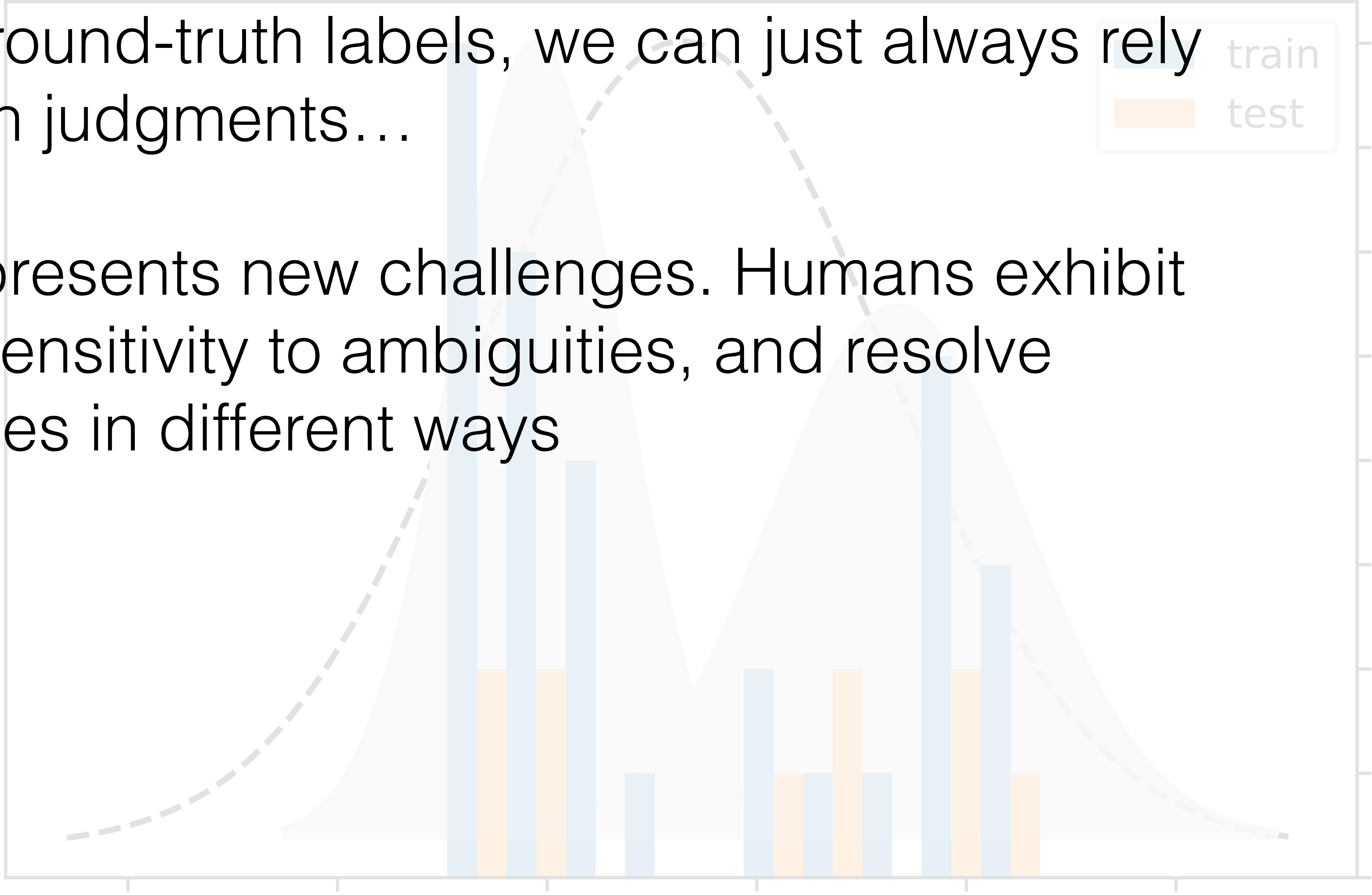
Takeaways

- Its tempting to say that rather than using theories to assign ground-truth labels, we can just always rely on human judgments…

Takeaways

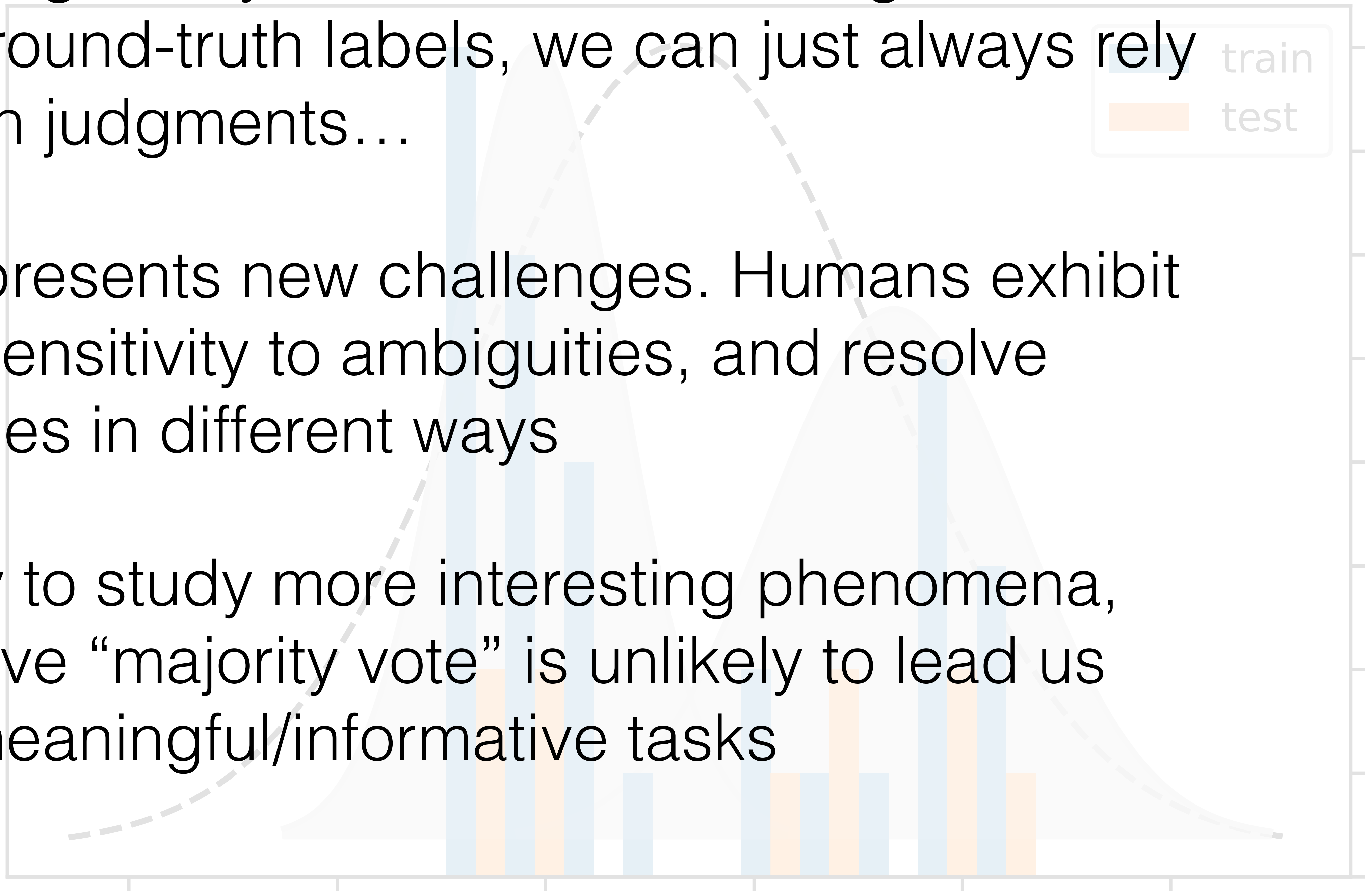
- Its tempting to say that rather than using theories to assign ground-truth labels, we can just always rely on human judgments…

- But this presents new challenges. Humans exhibit varying sensitivity to ambiguities, and resolve ambiguities in different ways



Takeaways

- Its tempting to say that rather than using theories to assign ground-truth labels, we can just always rely on human judgments…
- But this presents new challenges. Humans exhibit varying sensitivity to ambiguities, and resolve ambiguities in different ways
- As we try to study more interesting phenomena, using naive “majority vote” is unlikely to lead us toward meaningful/informative tasks



Conclusion

Conclusion

- Hot Take: Text-Only evals are dead. Maybe we just need to be working with situated language.

Conclusion

- Hot Take: Text-Only evals are dead. Maybe we just need to be working with situated language.

- Cooler Take: We need new eval tools. Many of the interesting phenomena we care about don't manifest neatly as inference or acceptability tasks.

Conclusion

- Hot Take: Text-Only evals are dead. Maybe we just need to be working with situated language.

- Cooler Take: We need new eval tools. Many of the interesting phenomena we care about don’t manifest neatly as inference or acceptability tasks.

- Theories of semantic representations in humans are not cut-and-dry, which makes it hard to establish meaningful eval standards. We should be engaging more with (and contributing to!) psych/ling research on these topics.