Designing and Interpreting Probes with Control Tasks

John Hewitt

Percy Liang

Overview

An emerging body of NLP work asks

“Does my neural network implicitly learn task Y?”

Overview

parts-of-speech
syntax
semantics

An emerging body of NLP work asks

“Does my neural network implicitly learn task Y?”

Overview

parts-of-speech
syntax
semantics

An emerging body of NLP work asks
“Does my neural network implicitly learn task Y?”

If a neural network hasn’t learned some task,
our methods shouldn’t tell us it has.

Overview

parts-of-speech
syntax
semantics

An emerging body of NLP work asks
“Does my neural network implicitly learn task Y?”

If a neural network hasn’t learned some task,
our methods shouldn’t tell us it has.

(Avoid false positives -- This is hard!)

Probing: supervised analysis of representations

[Shi et al., 2016, Peters et al., 2018, Tenney et al., 2019, Liu et al., 2019, Hewitt and Manning 2019, Kim et al., 2019...]

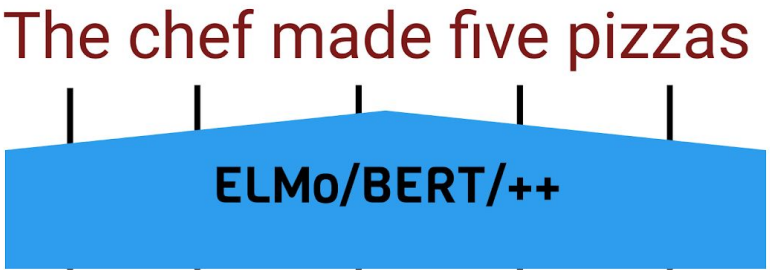
Probing: supervised analysis of representations

Does my network make task (e.g., part-of-speech) labels accessible?

[Shi et al., 2016, Peters et al., 2018, Tenney et al., 2019, Liu et al., 2019, Hewitt and Manning 2019, Kim et al., 2019...]

Probing: supervised analysis of representations

Does my network make task (e.g., part-of-speech) labels accessible?

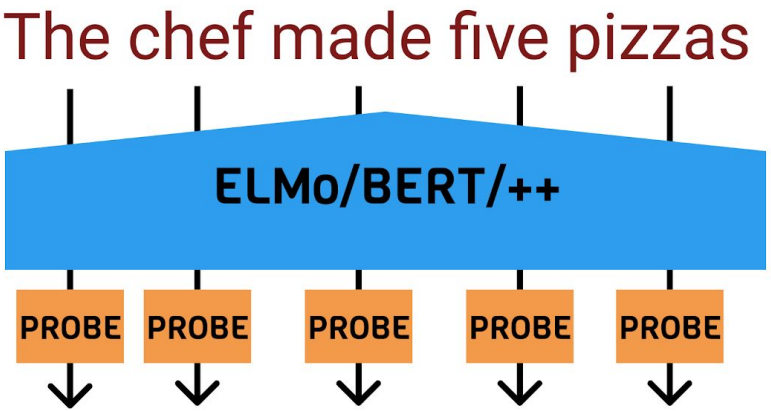


[Shi et al., 2016, Peters et al., 2018, Tenney et al., 2019, Liu et al., 2019, Hewitt and Manning 2019, Kim et al., 2019...]

Probing: supervised analysis of representations

Does my network make task (e.g., part-of-speech) labels accessible?

Choose a function family
to decode the task**.** (e.g., linear)



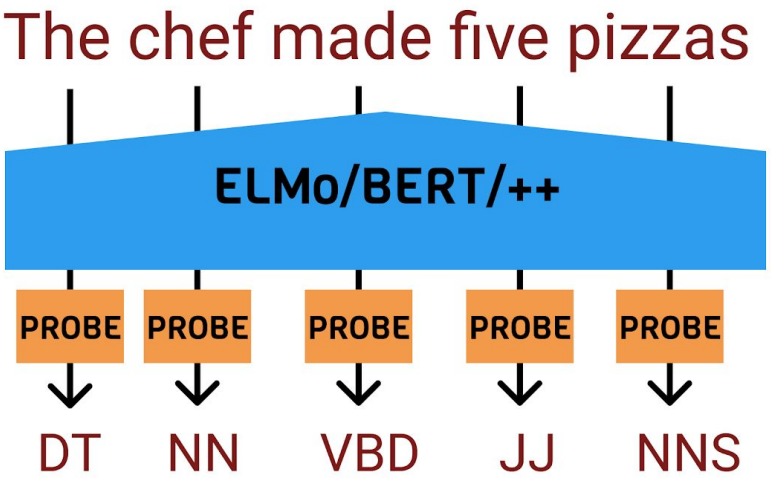
[Shi et al., 2016, Peters et al., 2018, Tenney et al., 2019, Liu et al., 2019, Hewitt and Manning 2019, Kim et al., 2019...]

Probing: supervised analysis of representations

Does my network make task (e.g., part-of-speech) labels accessible?

Choose a function family
to decode the task**.** (e.g., linear)

Train a function
representations --> task



[Shi et al., 2016, Peters et al., 2018, Tenney et al., 2019, Liu et al., 2019, Hewitt and Manning 2019, Kim et al., 2019...]

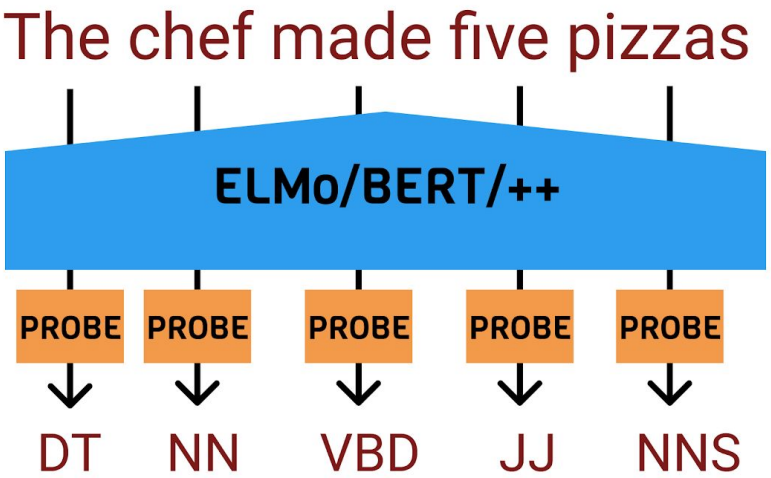
Probing: supervised analysis of representations

Does my network make task (e.g., part-of-speech) labels accessible?

Choose a function family
to decode the task**.** (e.g., linear)

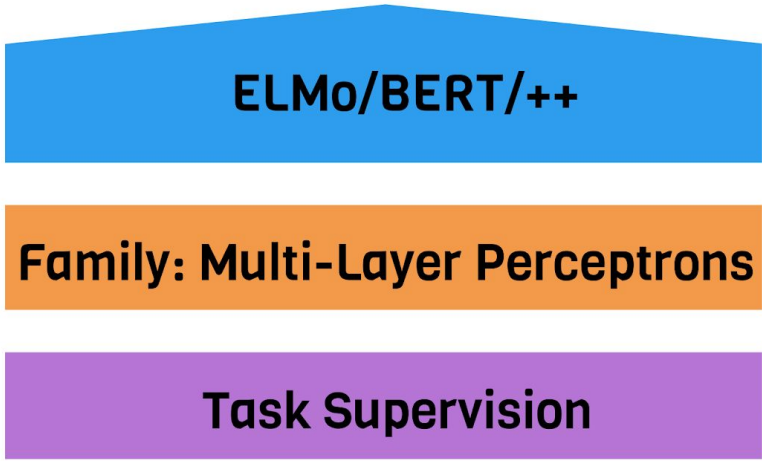
Train a function
representations --> task

Interpret accuracy
on held-out data

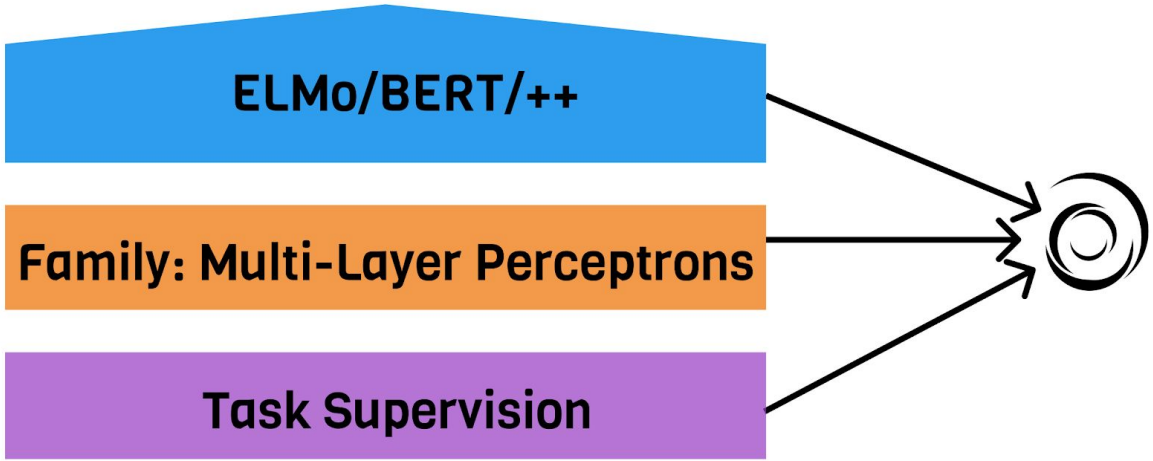


[Shi et al., 2016, Peters et al., 2018, Tenney et al., 2019, Liu et al., 2019, Hewitt and Manning 2019, Kim et al., 2019...]

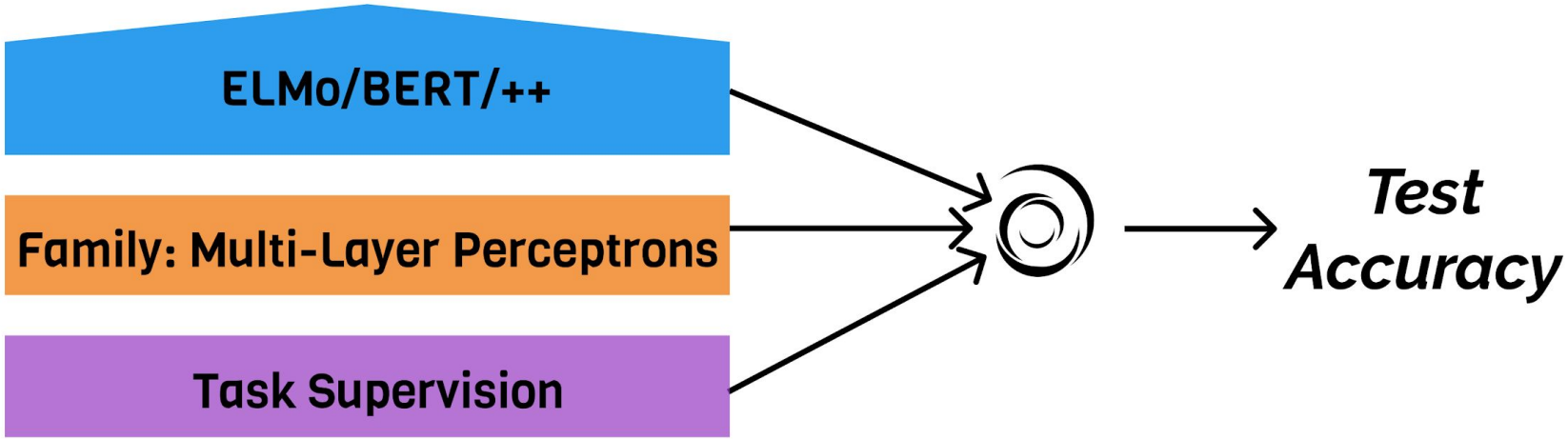
The probe confounder problem



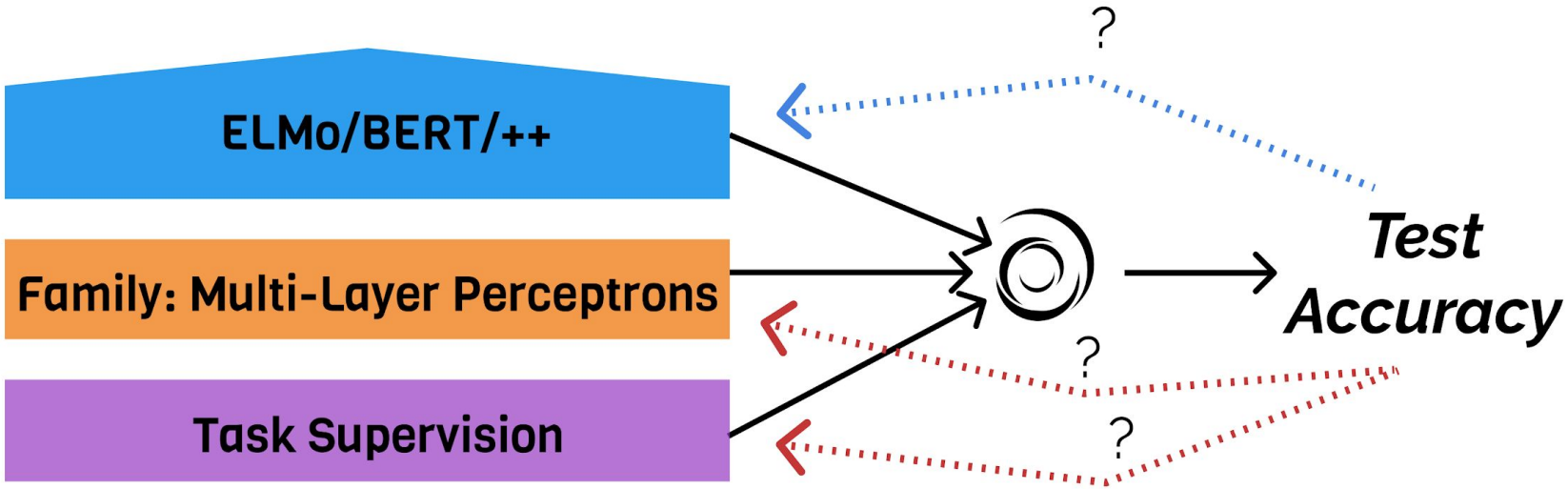
The probe confounder problem



The probe confounder problem



The probe confounder problem



Should we give credit to the **representation**?
(and/or) the **probe** and the **task supervision**?

This work:

This work:

1. Does high probe test accuracy mean the representation learned the task?

This work:

1. Does high probe test accuracy mean the representation learned the task?

2. How does the design of probes affect probing results?

This work:

1. Does high probe test accuracy mean the representation learned the task?

2. How does the design of probes affect probing results?

3. Can the probe confounder problem affect probing conclusions in practice?

Question 1

Does high probe test accuracy mean the representation learned a task?

No. Our *control tasks* are learned by probes but not encoded by representations.

Question 1

Does high probe test accuracy mean the representation learned a task?

No. Our *control tasks* are learned by probes but not encoded by representations.

Probing: Does ELMo learn part-of-speech?

Probing: Does ELMo learn part-of-speech?

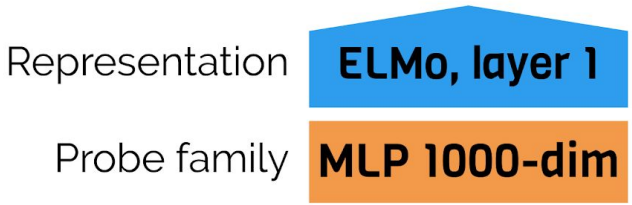
Train and test probes on ELMo representations on the Penn Treebank

Probing: Does ELMo learn part-of-speech?

Train and test probes on ELMo representations on the Penn Treebank

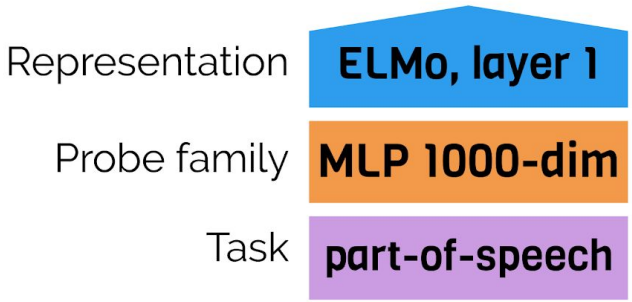
Probing: Does ELMo learn part-of-speech?

Train and test probes on ELMo representations on the Penn Treebank



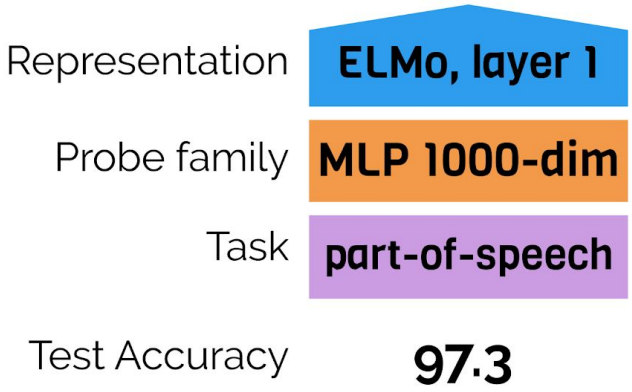
Probing: Does ELMo learn part-of-speech?

Train and test probes on ELMo representations on the Penn Treebank



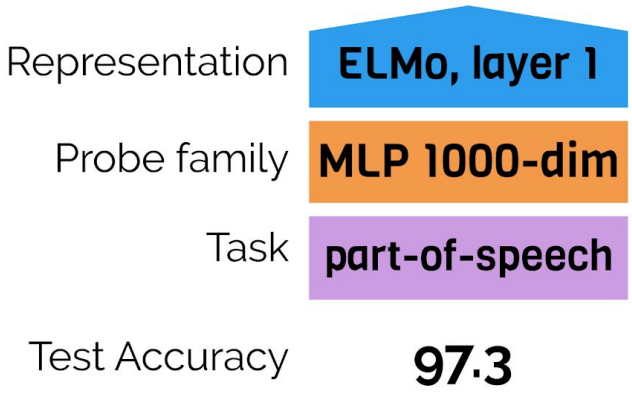
Probing: Does ELMo learn part-of-speech?

Train and test probes on ELMo representations on the Penn Treebank



Probing: Does ELMo learn part-of-speech?

Train and test probes on ELMo representations on the Penn Treebank



Probe achieves high accuracy!

Does the accuracy faithfully reflect the extent to which ELMo has learned part-of-speech tagging?

Defining **control tasks** for linguistic tasks

Defining **control tasks** for linguistic tasks

1. Look at task output space. e.g., 45 parts-of-speech.

Defining **control tasks** for linguistic tasks

1. Look at task output space. e.g., 45 parts-of-speech.
2. **Randomly partition** vocabulary into 45 categories

Defining **control tasks** for linguistic tasks

1. Look at task output space. e.g., 45 parts-of-speech.
2. **Randomly partition** vocabulary into 45 categories

Category1

house,
eat,
...

Defining **control tasks** for linguistic tasks

1. Look at task output space. e.g., 45 parts-of-speech.
2. **Randomly partition** vocabulary into 45 categories

Category1	**Category2**
house, eat, ...	pizza, the, ...

Defining **control tasks** for linguistic tasks

1. Look at task output space. e.g., 45 parts-of-speech.
2. **Randomly partition** vocabulary into 45 categories

Defining **control tasks** for linguistic tasks

1. Look at task output space. e.g., 45 parts-of-speech.
2. **Randomly partition** vocabulary into 45 categories

3. **Deterministically label** sentences in a corpus by looking up category for each word

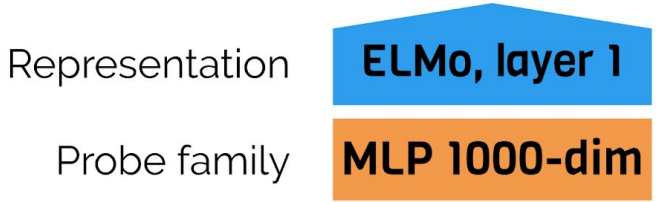
Defining **control tasks** for linguistic tasks

1. Look at task output space. e.g., 45 parts-of-speech.
2. **Randomly partition** vocabulary into 45 categories

3. **Deterministically label** sentences in a corpus by looking up category for each word

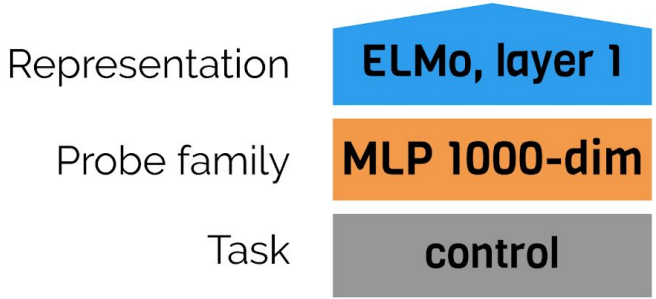
Do probes learn control tasks?

Train and test probes on ELMo representations on the Penn Treebank



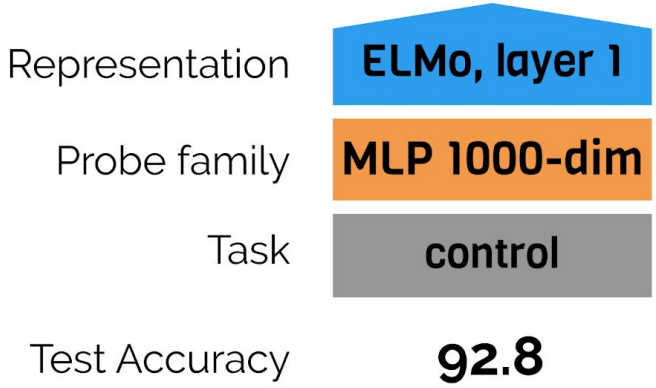
Do probes learn control tasks?

Train and test probes on ELMo representations on the Penn Treebank



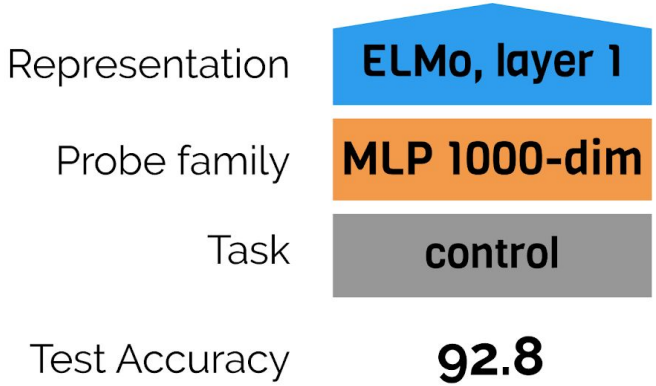
Do probes learn control tasks?

Train and test probes on ELMo representations on the Penn Treebank



Do probes learn control tasks?

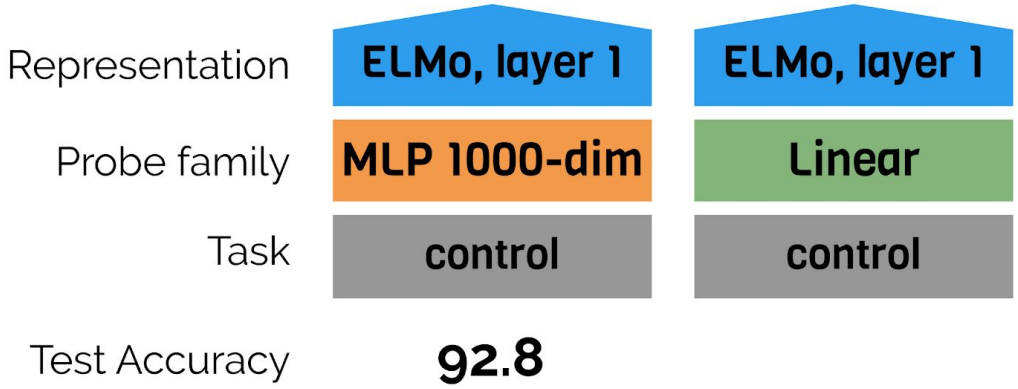
Train and test probes on ELMo representations on the Penn Treebank



MLP probe: high accuracy on control tasks; does not reflect representation!

Do probes learn control tasks?

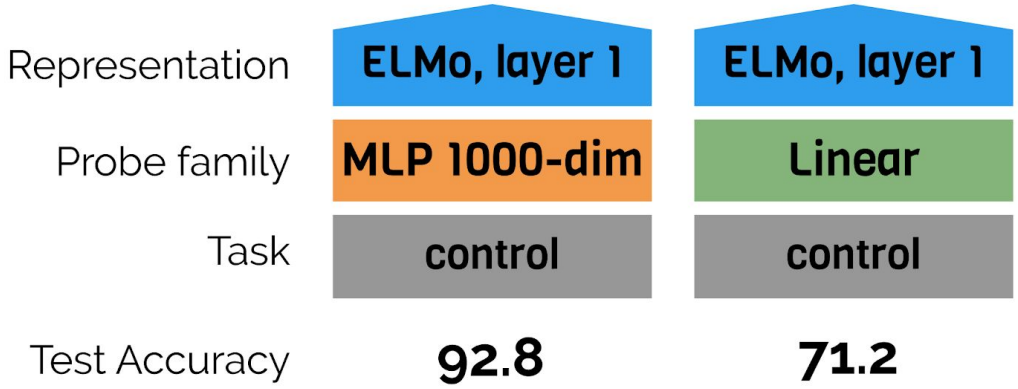
Train and test probes on ELMo representations on the Penn Treebank



MLP probe: high accuracy on control tasks; does not reflect representation!

Do probes learn control tasks?

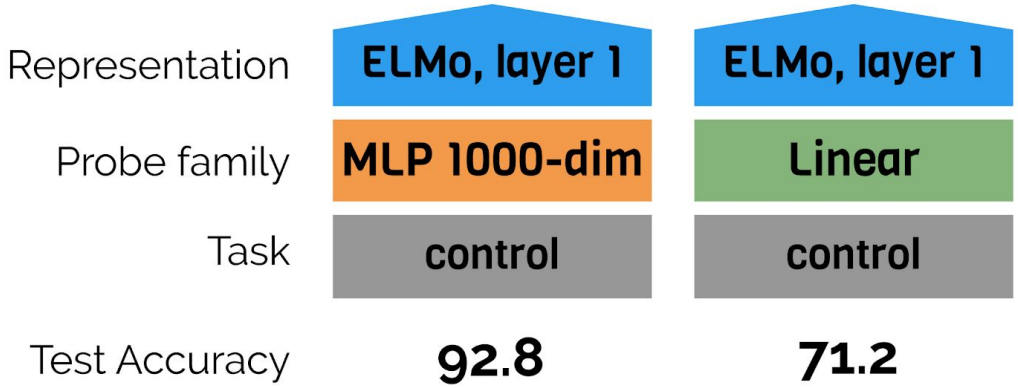
Train and test probes on ELMo representations on the Penn Treebank



MLP probe: high accuracy on control tasks; does not reflect representation!

Do probes learn control tasks?

Train and test probes on ELMo representations on the Penn Treebank



MLP probe: high accuracy on control tasks; does not reflect representation!

Linear probe: lower accuracy on control tasks

Selectivity for interpreting probing results

Idea: get a rough measure of how linguistic task accuracy may derive from probe expressivity and supervision.

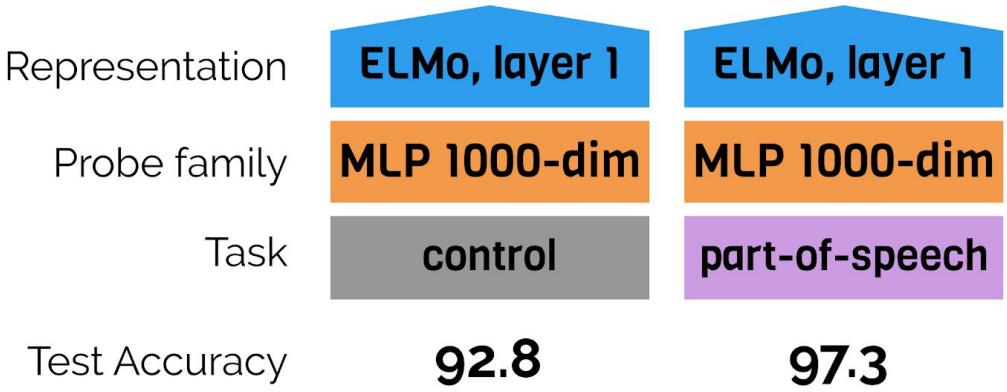
Selectivity for interpreting probing results

Idea: get a rough measure of how linguistic task accuracy may derive from probe expressivity and supervision.

We define ***selectivity*** as a probe's accuracy on the linguistic task minus its accuracy on the control task

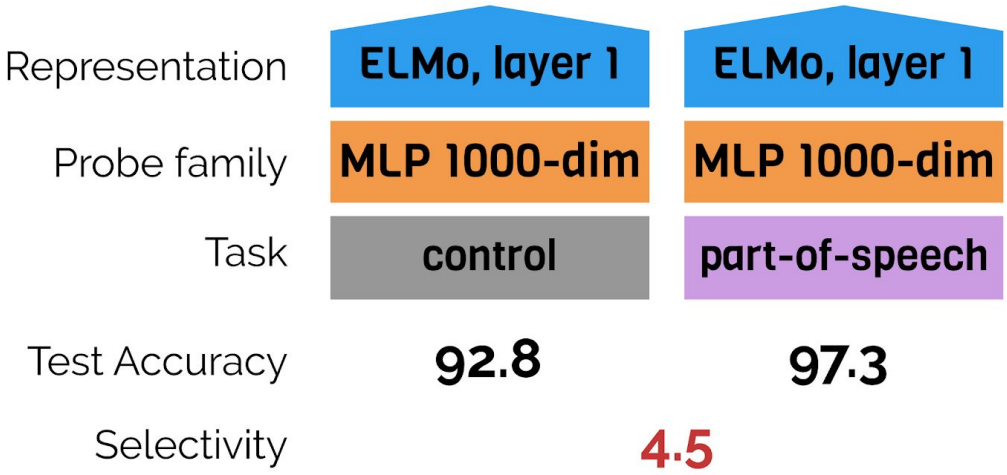
Probing part-of-speech vs control task

Can control tasks and selectivity help put probing accuracies in context?



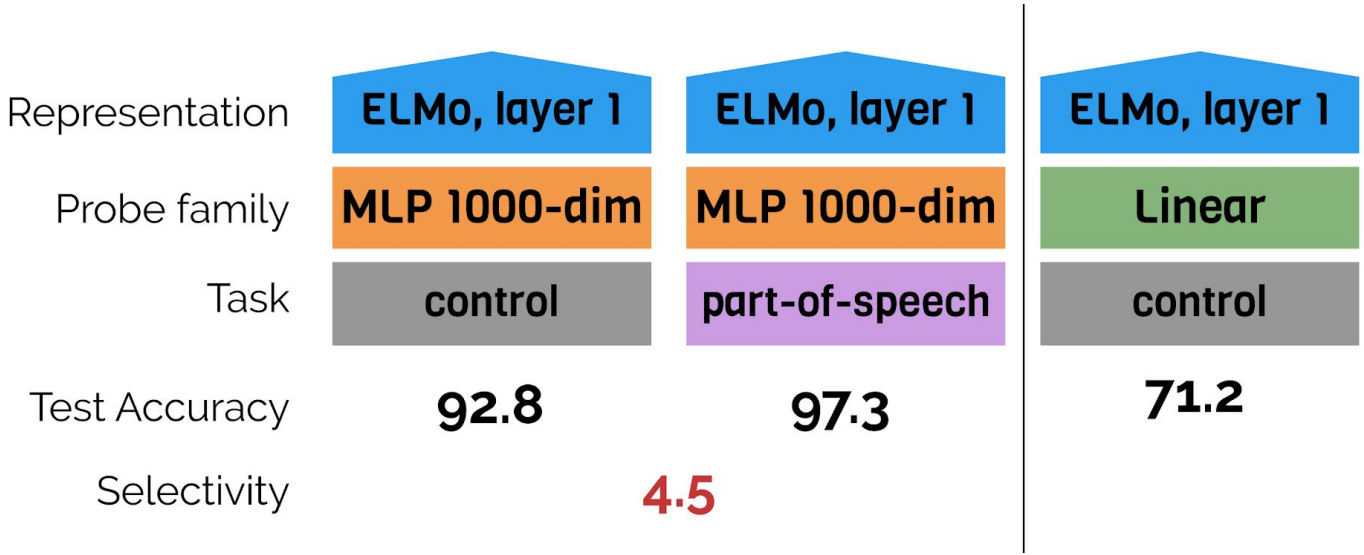
Probing part-of-speech vs control task

Can control tasks and selectivity help put probing accuracies in context?



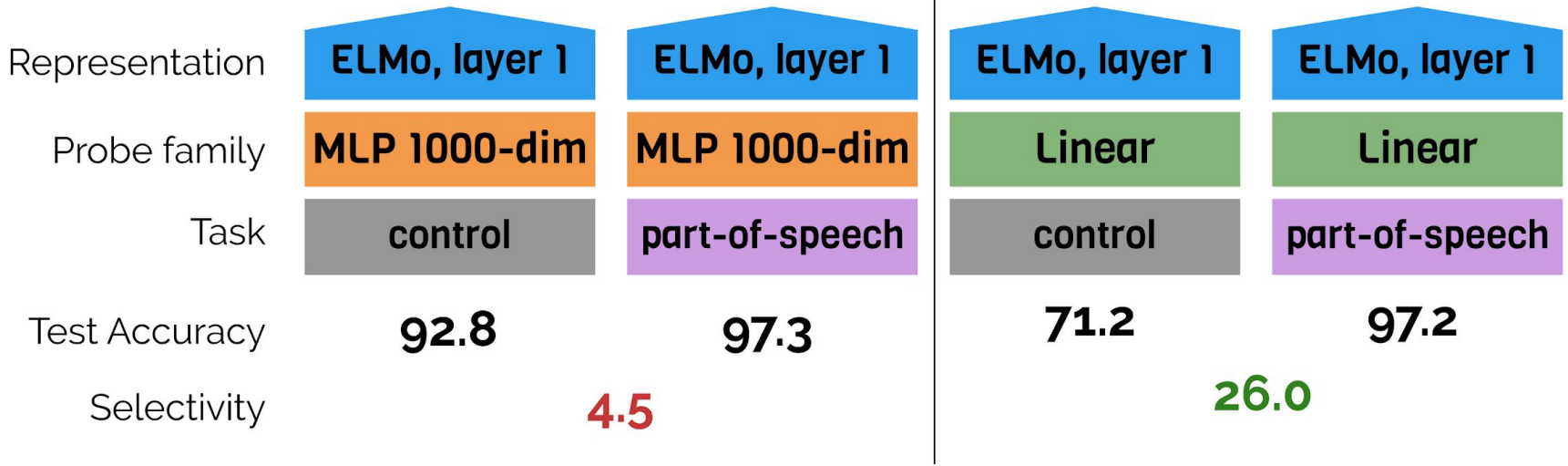
Probing part-of-speech vs control task

Can control tasks and selectivity help put probing accuracies in context?



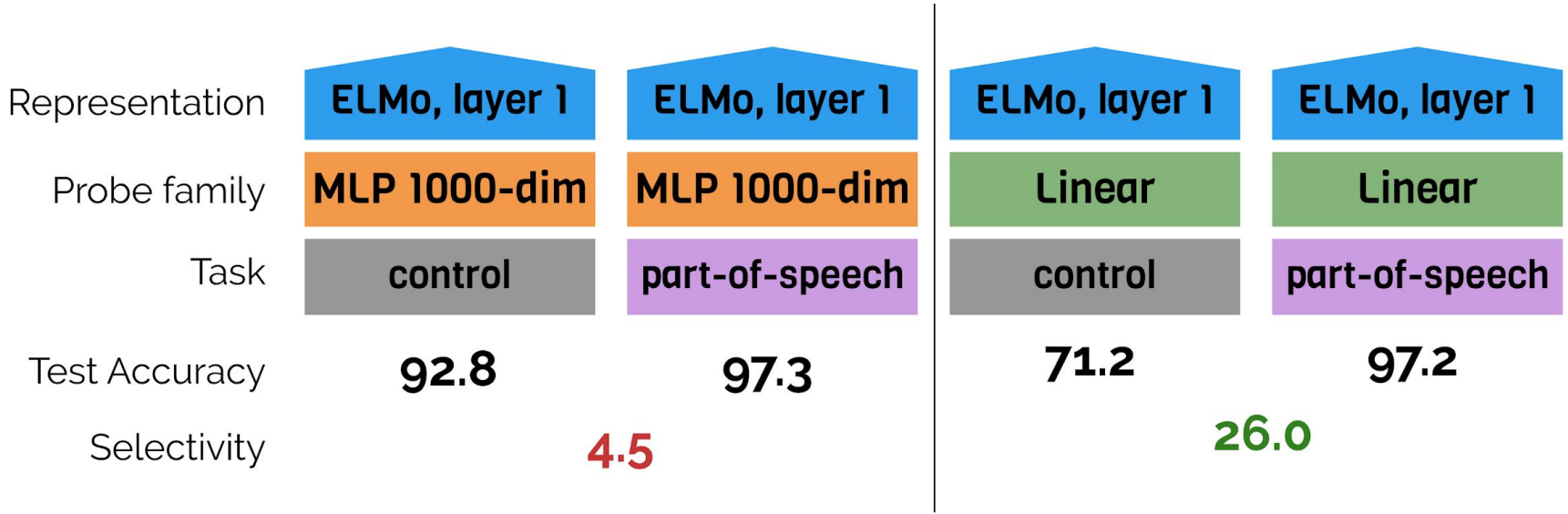
Probing part-of-speech vs control task

Can control tasks and selectivity help put probing accuracies in context?



Probing part-of-speech vs control task

Can control tasks and selectivity help put probing accuracies in context?



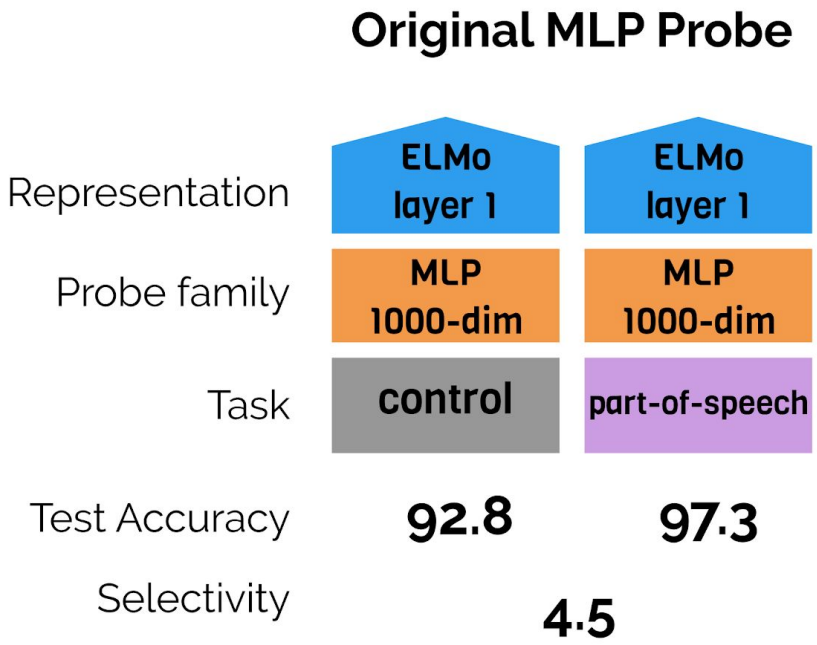
Probes with similar linguistic task accuracy may have very different selectivity

Question 2

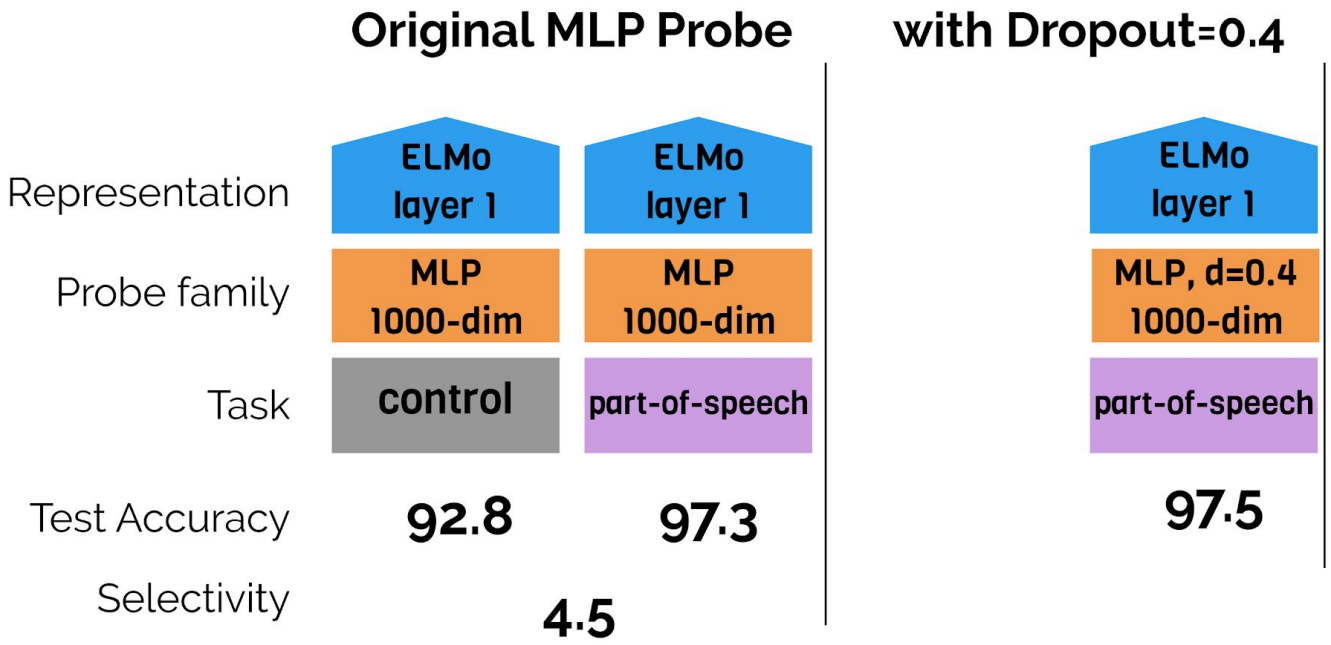
How does the design of probes affect probing results?

Designing for good linguistic task generalization does not necessarily lead to selective probes

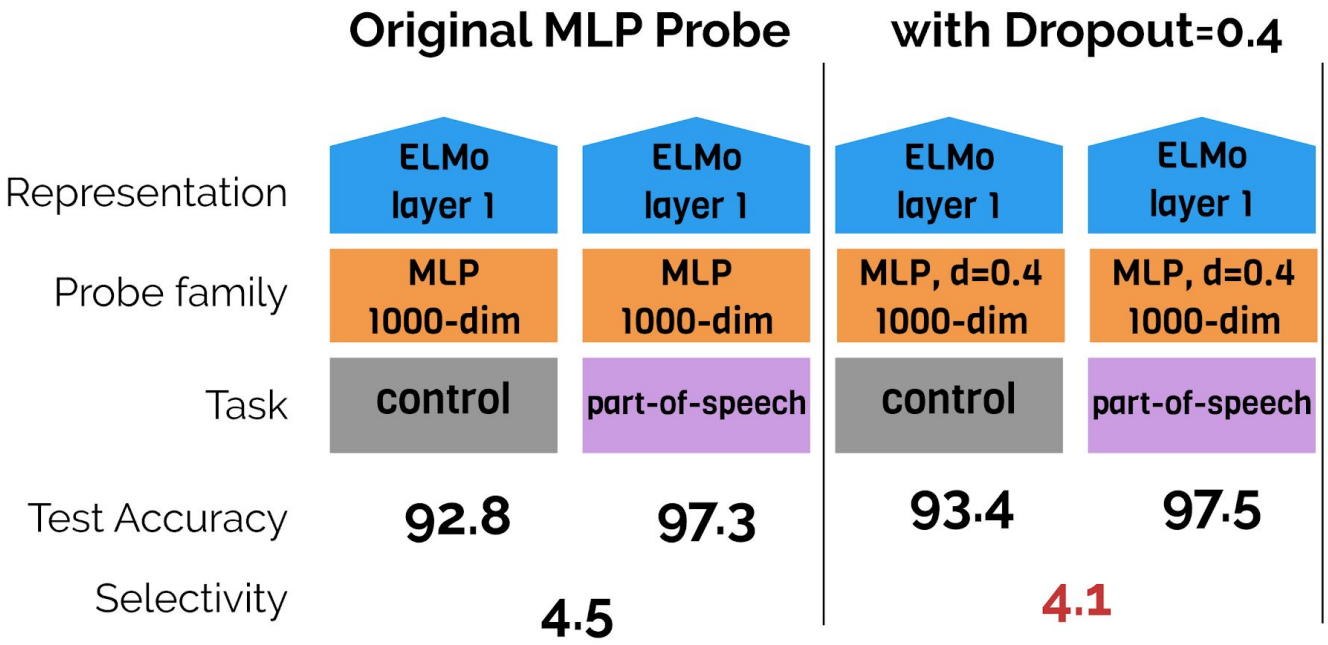
Designing probes with control tasks



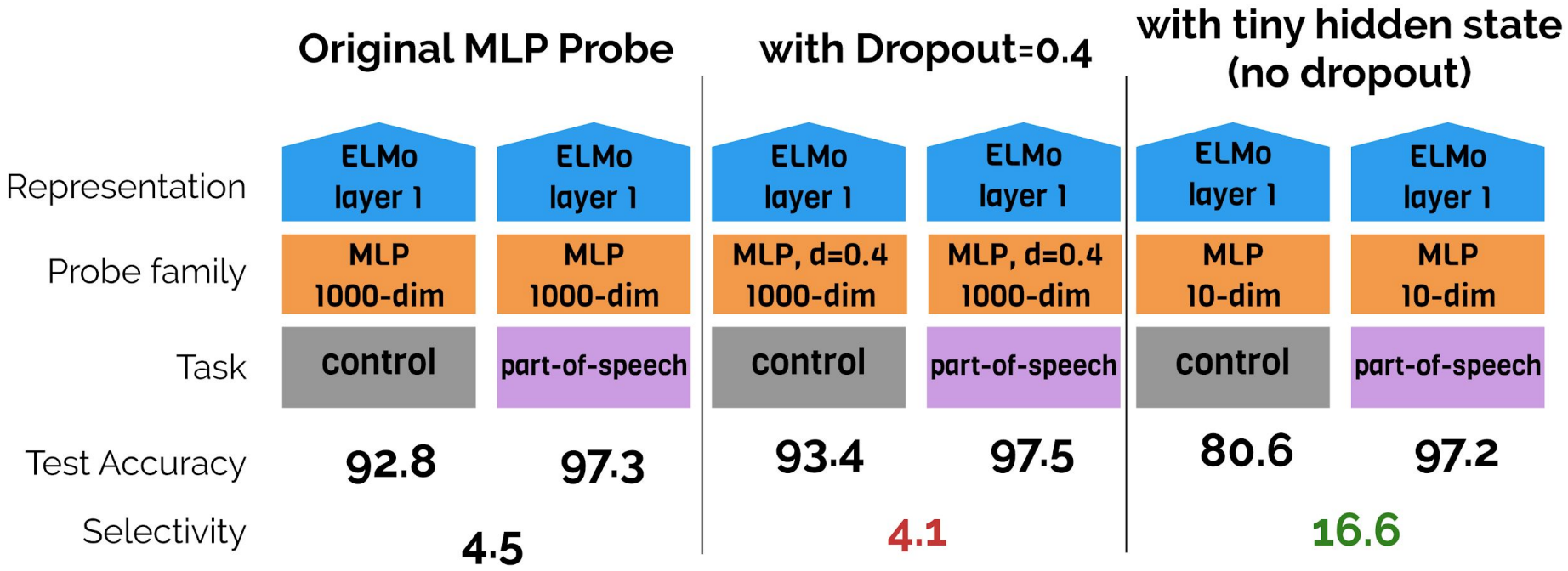
Designing probes with control tasks



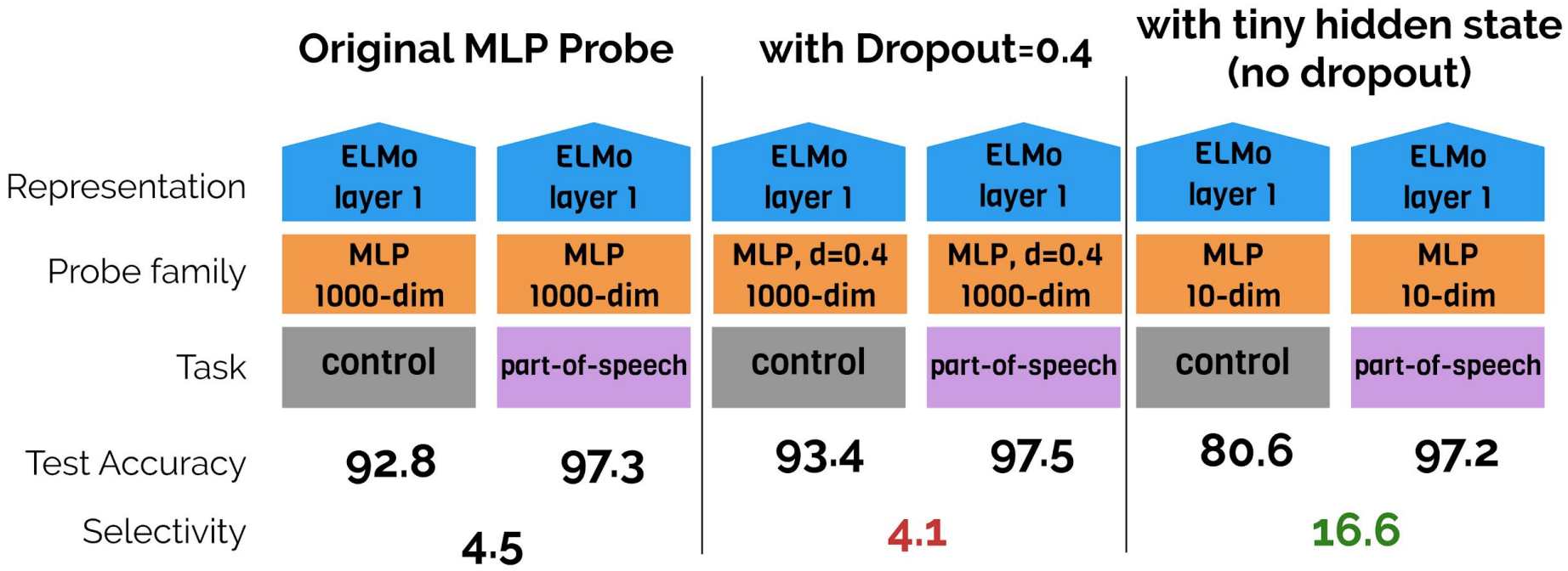
Designing probes with control tasks



Designing probes with control tasks



Designing probes with control tasks



Simply regularizing — to minimize generalization gap — doesn't necessarily lead to selectivity!

Question 3

Can the probe confounder problem affect probing conclusions in practice?

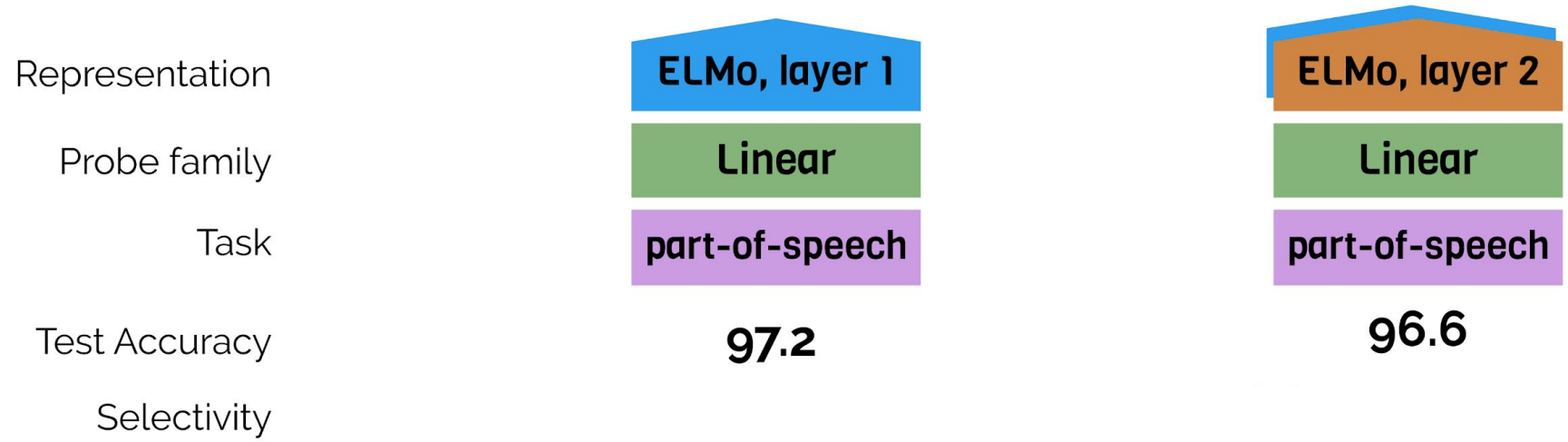
Yes — probes may be picking up on spurious signals

Re-examining probes on ELMo's layers

Is ELMo1 better at part-of-speech than ELMo2?

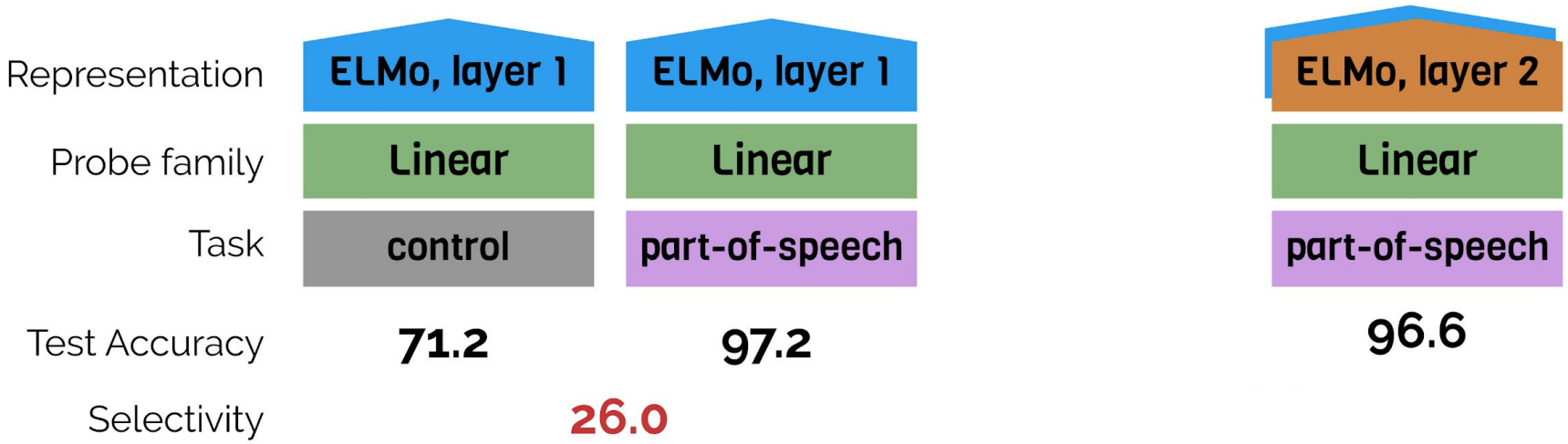
Re-examining probes on ELMo's layers

Is ELMo1 better at part-of-speech than ELMo2?



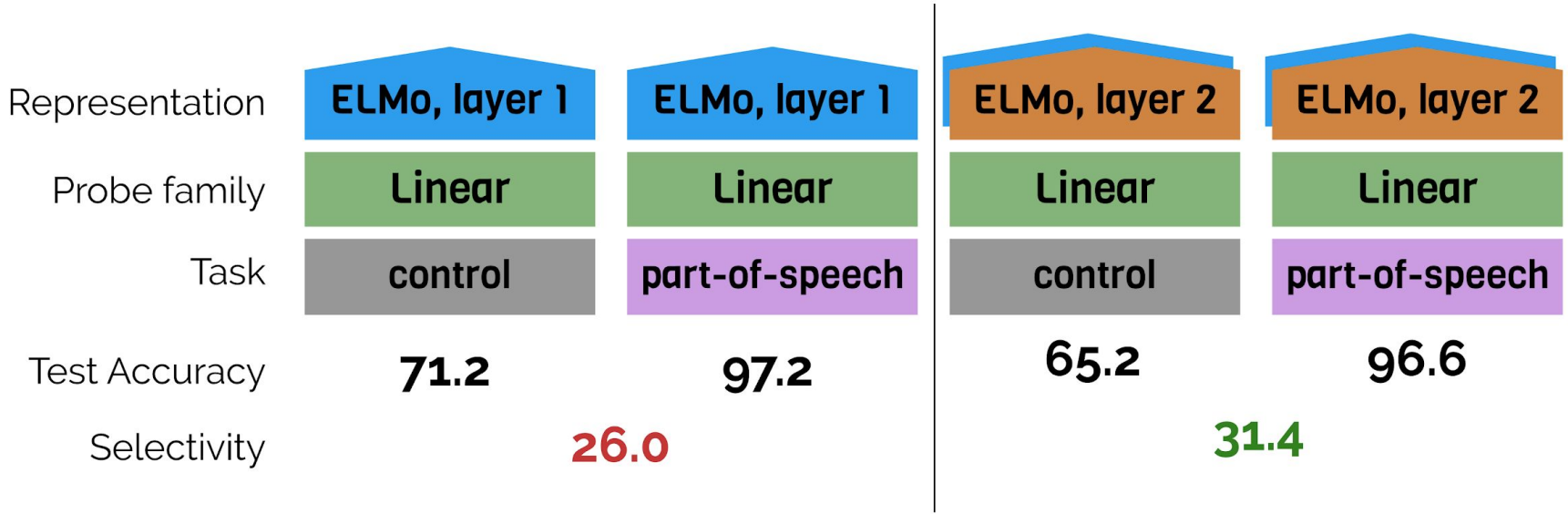
Re-examining probes on ELMo's layers

Is ELMo1 better at part-of-speech than ELMo2?



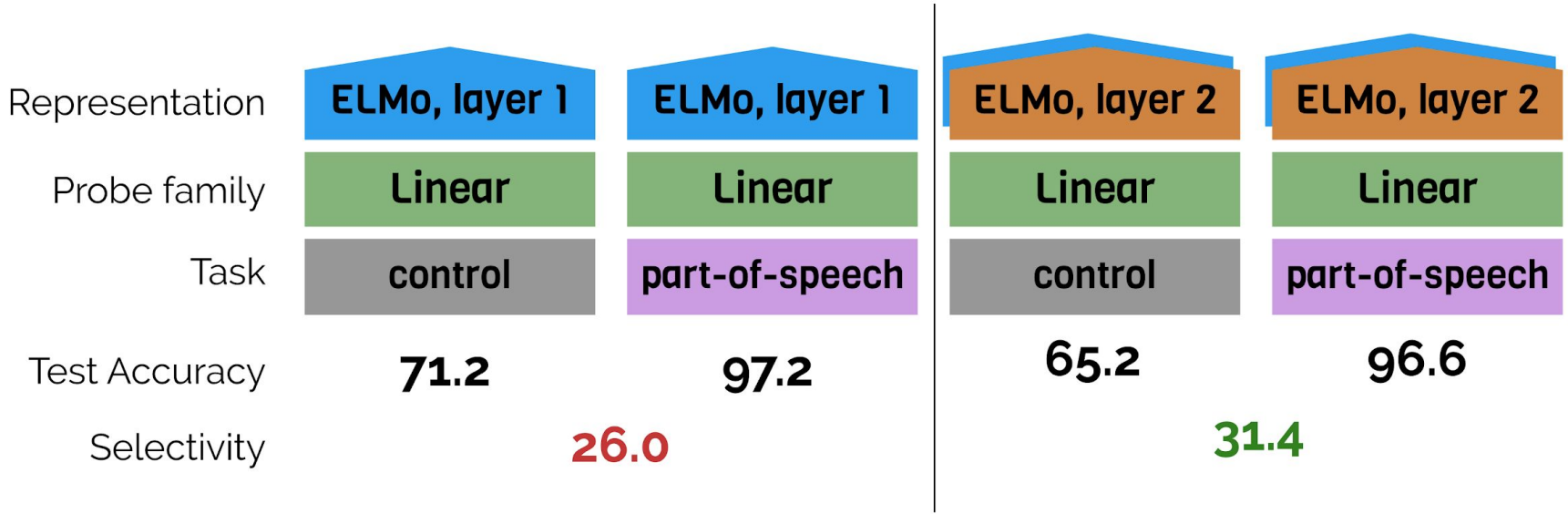
Re-examining probes on ELMo's layers

Is ELMo1 better at part-of-speech than ELMo2?



Re-examining probes on ELMo's layers

Is ELMo1 better at part-of-speech than ELMo2?



ELMo1 part-of-speech gains over ELMo2 may be explained by easier access to a **spurious signal: word identity**

Limitations

Limitations

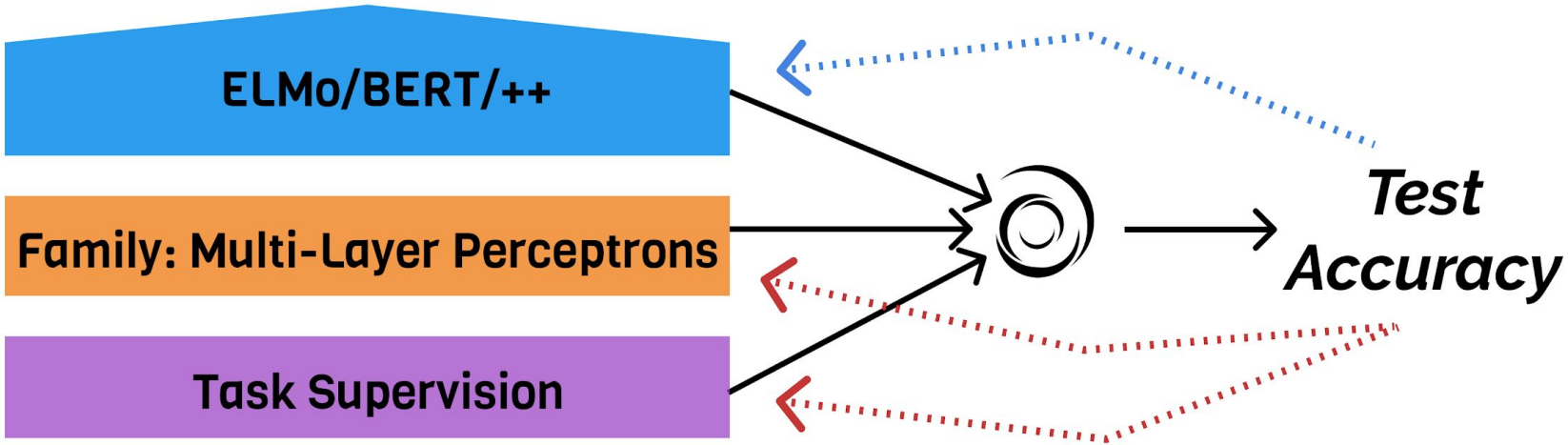
Our control tasks only use **word identity**; there are many possible spurious signals in probing

Limitations

Our control tasks only use **word identity**; there are many possible spurious signals in probing

Selectivity **builds intuition** but does not permit fine-grained claims, like “my model got *this* selectivity, so it learned the task.”

Thanks!


ELMo/BERT/++
Family: Multi-Layer Perceptrons
Task Supervision
Test Accuracy