Interpretation of Pretrained Language Models

Chenyan Xiong

11-667

Disclaimer

No one really understand why language model works

Very limited theory and very limited empirical observation, especially at large scale

This lecture is to share:

- Observations upon, not causality of, the behavior of LLMs
- Early attempts to interpret their ability
- Useful intuitions and interesting thought experiments

Outline

What is captured in BERT?

Why pretrained models generalize?

What does in-context learning do?

Outline

What is captured in BERT?

- Attention patterns
- Probing capture capabilities in representations

Why pretrained models generalize?

What does in-context learning do?

BERT Attention Patterns

Restate Transformer's attention mechanism:

Attention from $i \to j$: $$\alpha_{ij} = \frac{\exp(q_i \cdot k_j / \sqrt{d_k})}{\sum_t \exp(q_i \cdot k_t / \sqrt{d_k})}$$

New representation of i: $$o_i = \sum_j \alpha_{ij} v_j$$

The new representation of position i is the attention-weighted combination of other positions' value

- Higher α_{ij} $\rightarrow$bigger contribution of position j to position i

BERT Attention Patterns: Stats

Average Entropy of α_{ij}

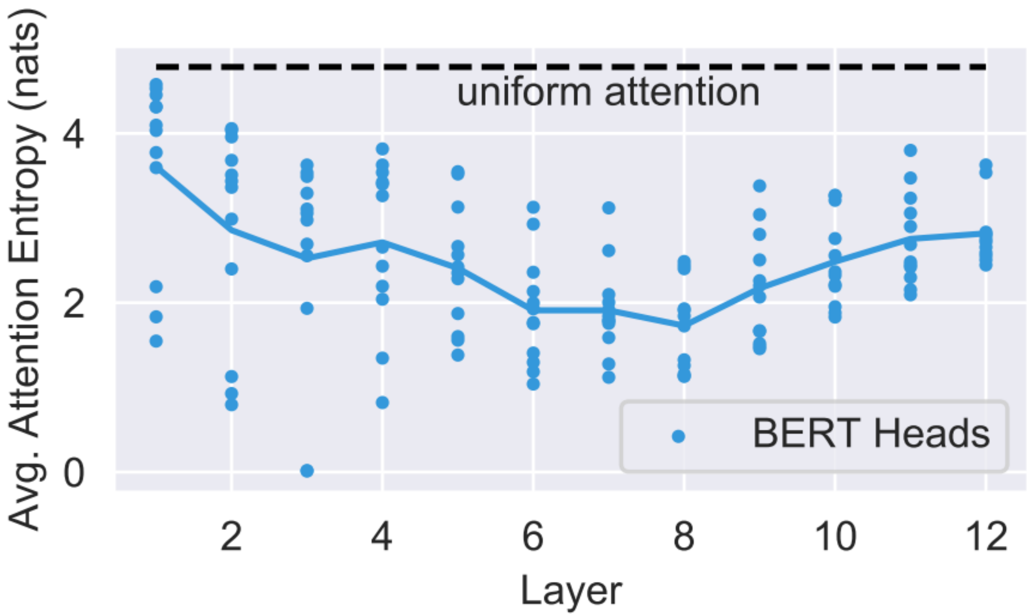

Figure 1: Entropy of BERT Attention Distributions [1]

BERT Attention Patterns: Stats

High entropy heads in lower layers:

- Bag-of-words alike mechanism

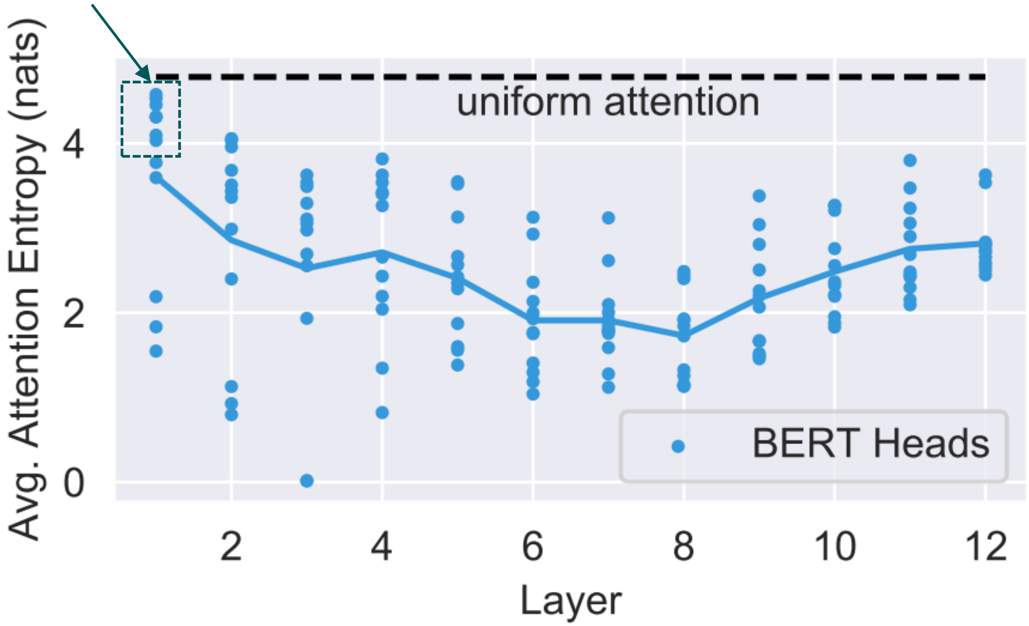

Figure 1: Entropy of BERT Attention Distributions [1]

[1] Clark Et al. "What Does BERT Look At? An Analysis of BERT's Attention." BlackBoxNLP 2019

BERT Attention Patterns: Stats

Lower entropy in middle layers:
- Start forming certain patterns?

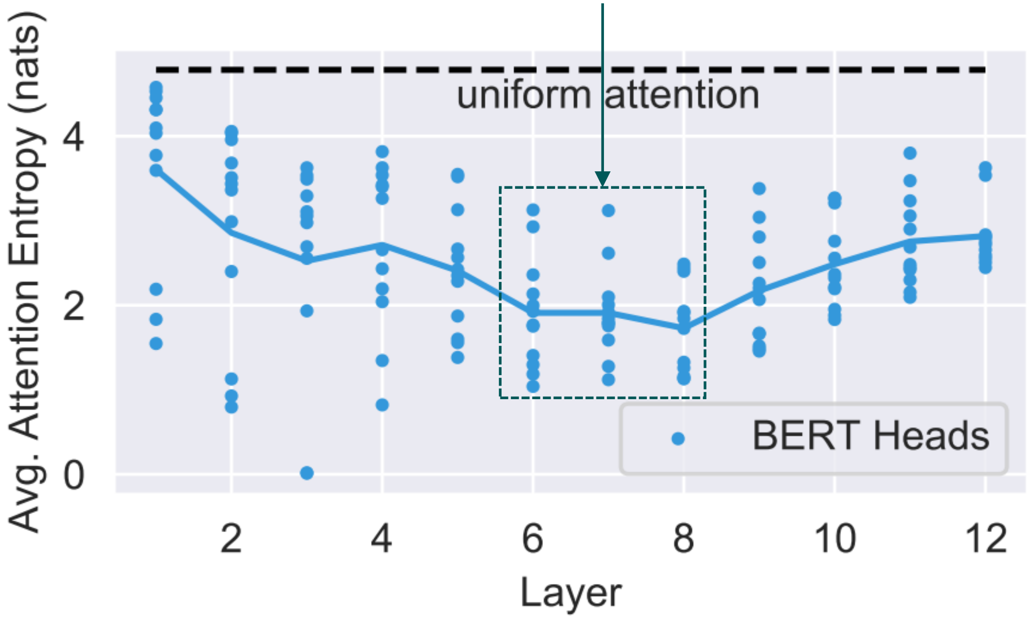

Figure 1: Entropy of BERT Attention Distributions [1]

[1] Clark Et al. “What Does BERT Look At? An Analysis of BERT’s Attention.” BlackBoxNLP 2019

BERT Attention Patterns: Stats

Rising entropy in deep layers:
- More global information?

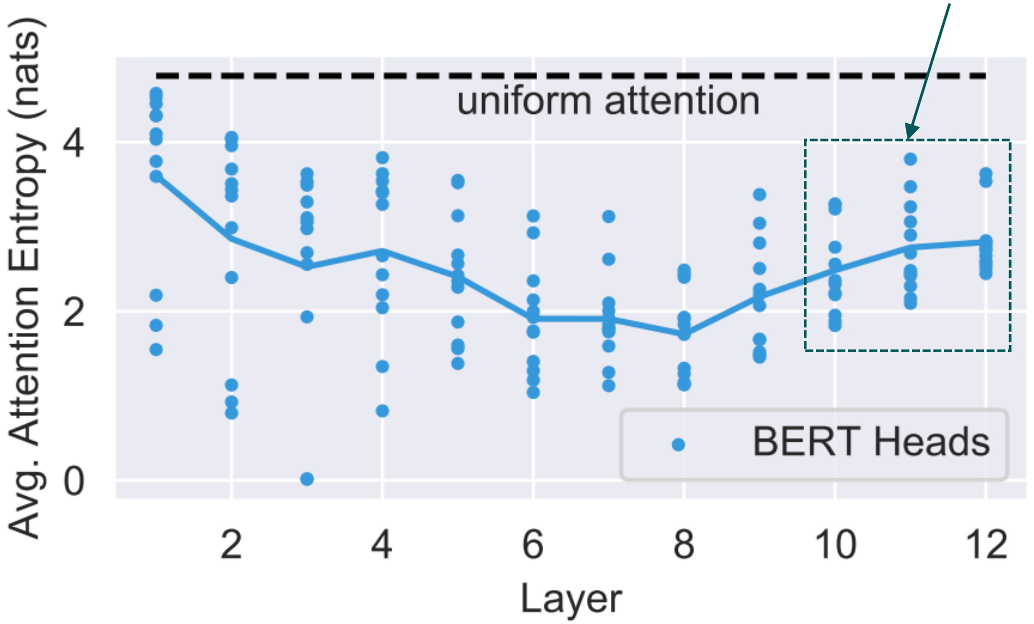

Figure 1: Entropy of BERT Attention Distributions [1]

BERT Attention Patterns: Common Patterns

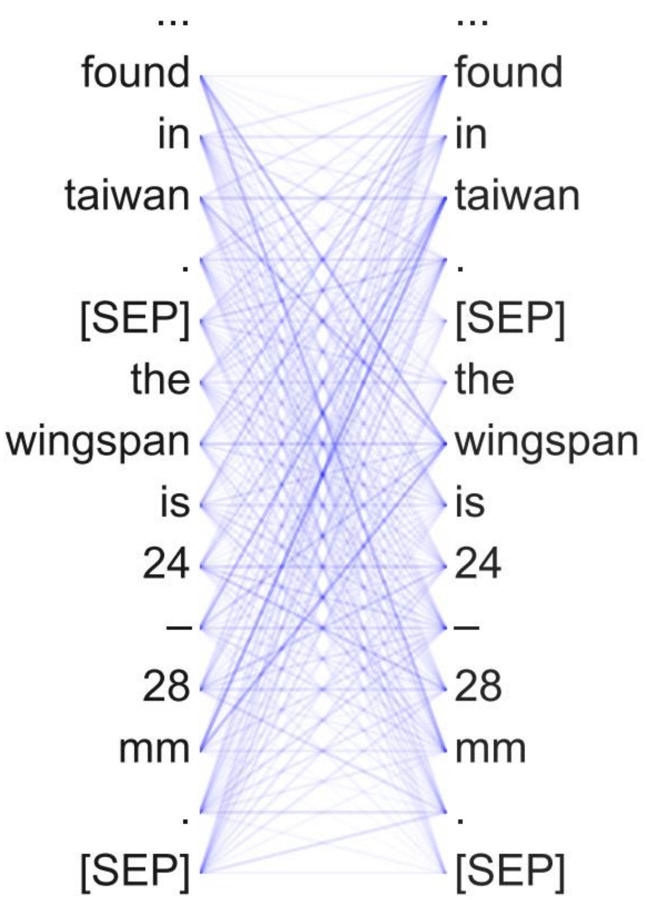

Figure 2: Attend Broadly (Left→Right) [1]

Common Pattern 1: Broad attention

- Neural networks are hard to interpret
- Various stuffs mixed together, hard to tell

[1] Clark Et al. "What Does BERT Look At? An Analysis of BERT's Attention." BlackBoxNLP 2019

BERT Attention Patterns: Common Patterns

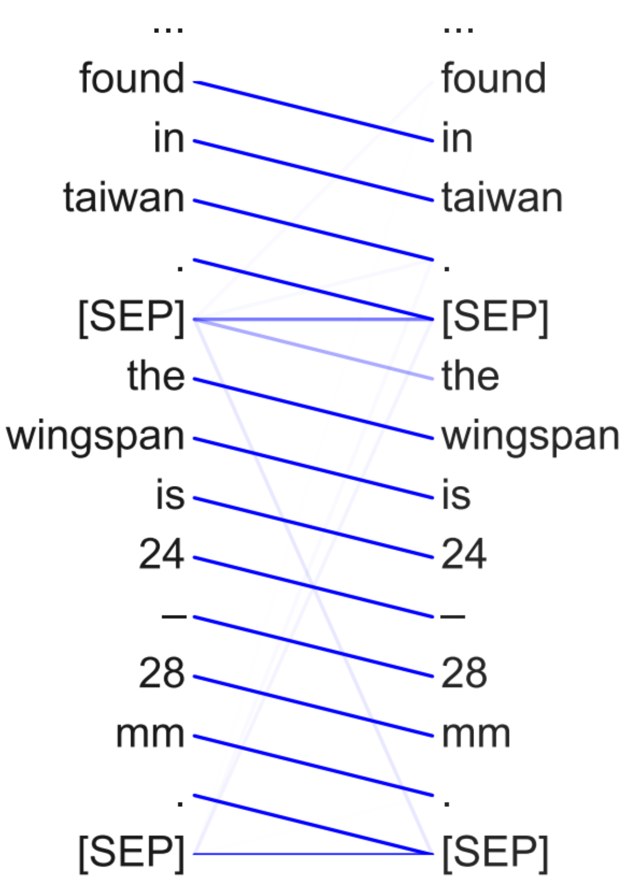

Figure 3: Attend to Next (Left→Right) [1]

Common Pattern 2: Attend to next token

- Reverse RNN style
- Learned positional relation in pretraining

BERT Attention Patterns: Common Patterns

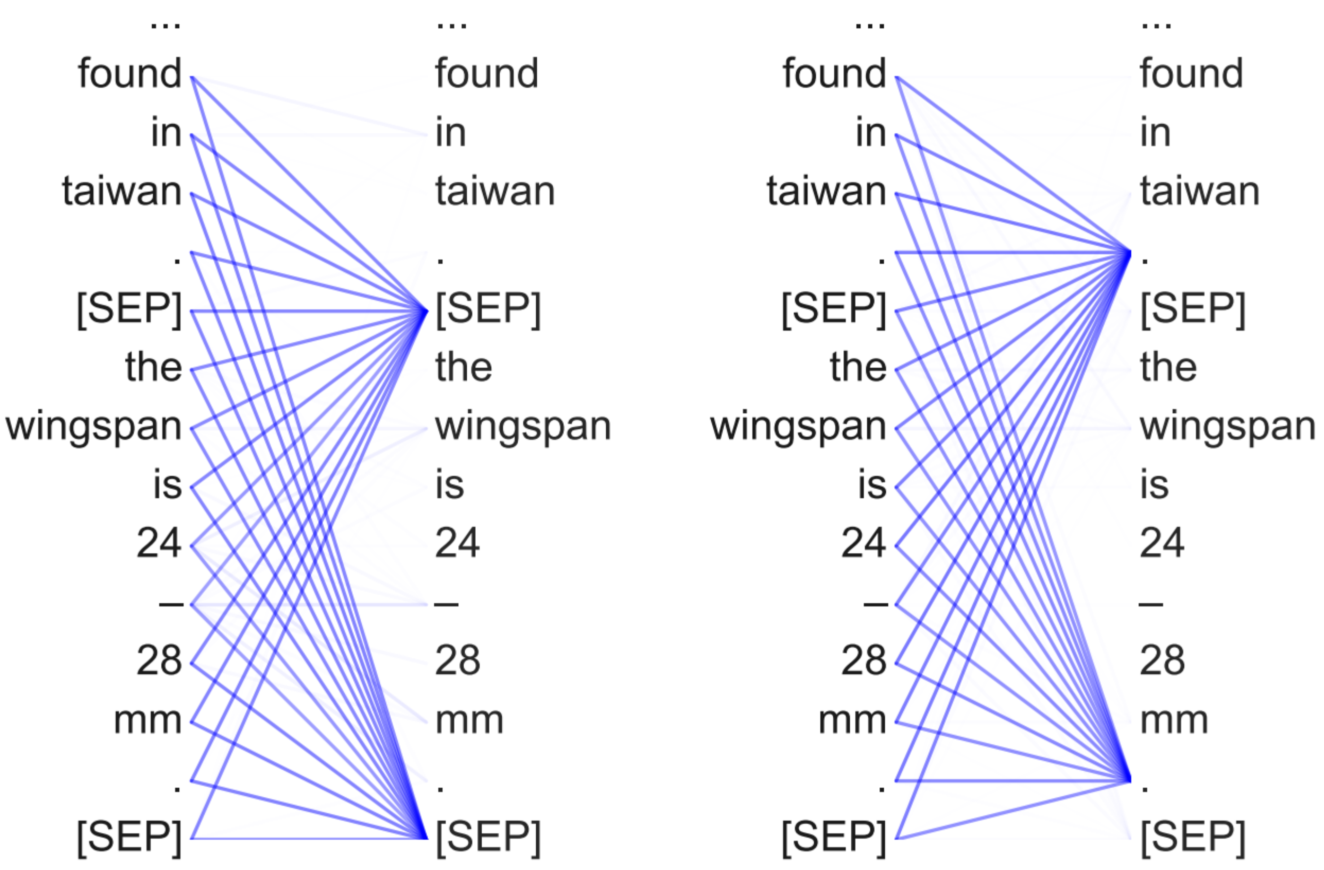

Figure 4: Attend to [SEP] and punctuations (Left→Right) [1]

Common Pattern 3: Attend to [SEP] and "."

- Centralizing attention to specific tokens
- Effect unclear
  - Some consider it a "none" operation
  - Some consider it as an information hub
  - Maybe a mix of both, at different heads

[1] Clark Et al. "What Does BERT Look At? An Analysis of BERT's Attention." BlackBoxNLP 2019

BERT Attention Patterns: Linguistic Examples

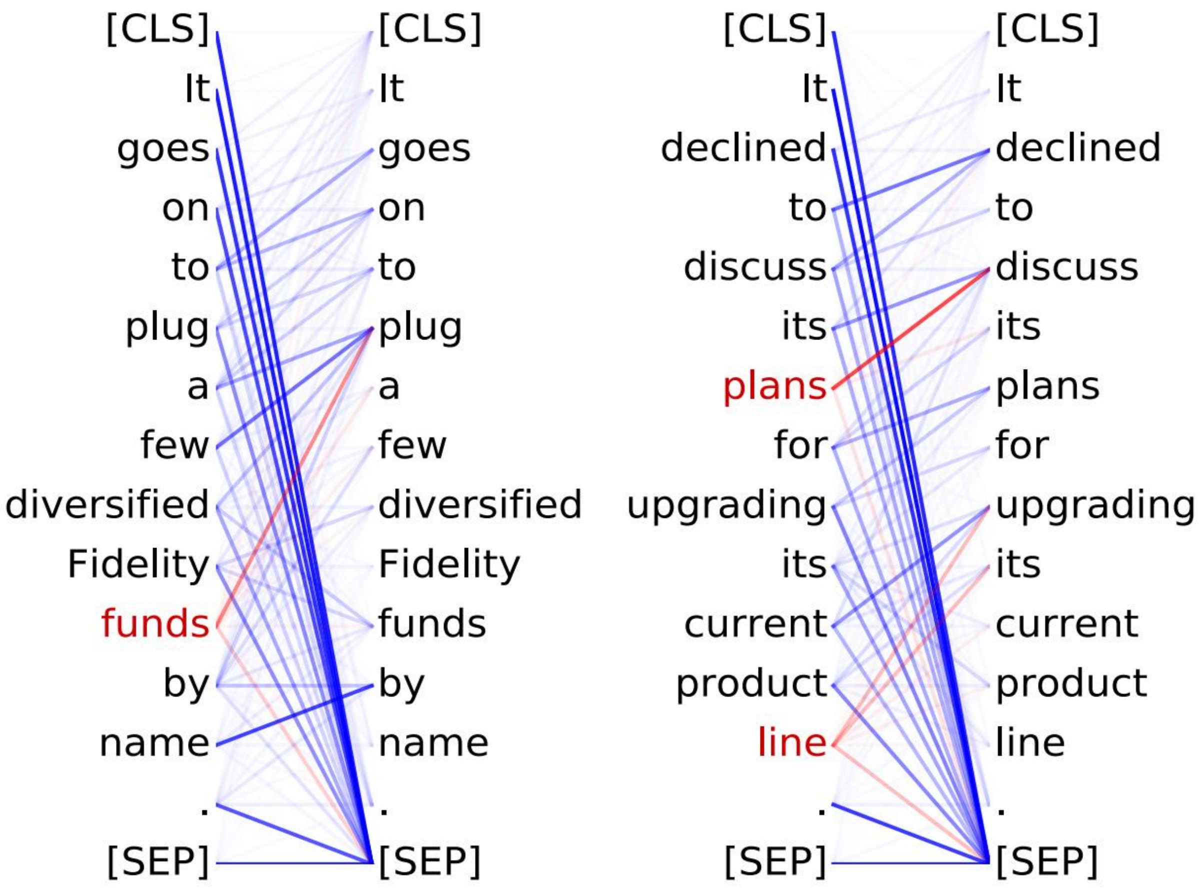

Figure 5: Objects Attend to their Verbs (Left→Right) [1]

BERT Attention Patterns: Linguistic Examples

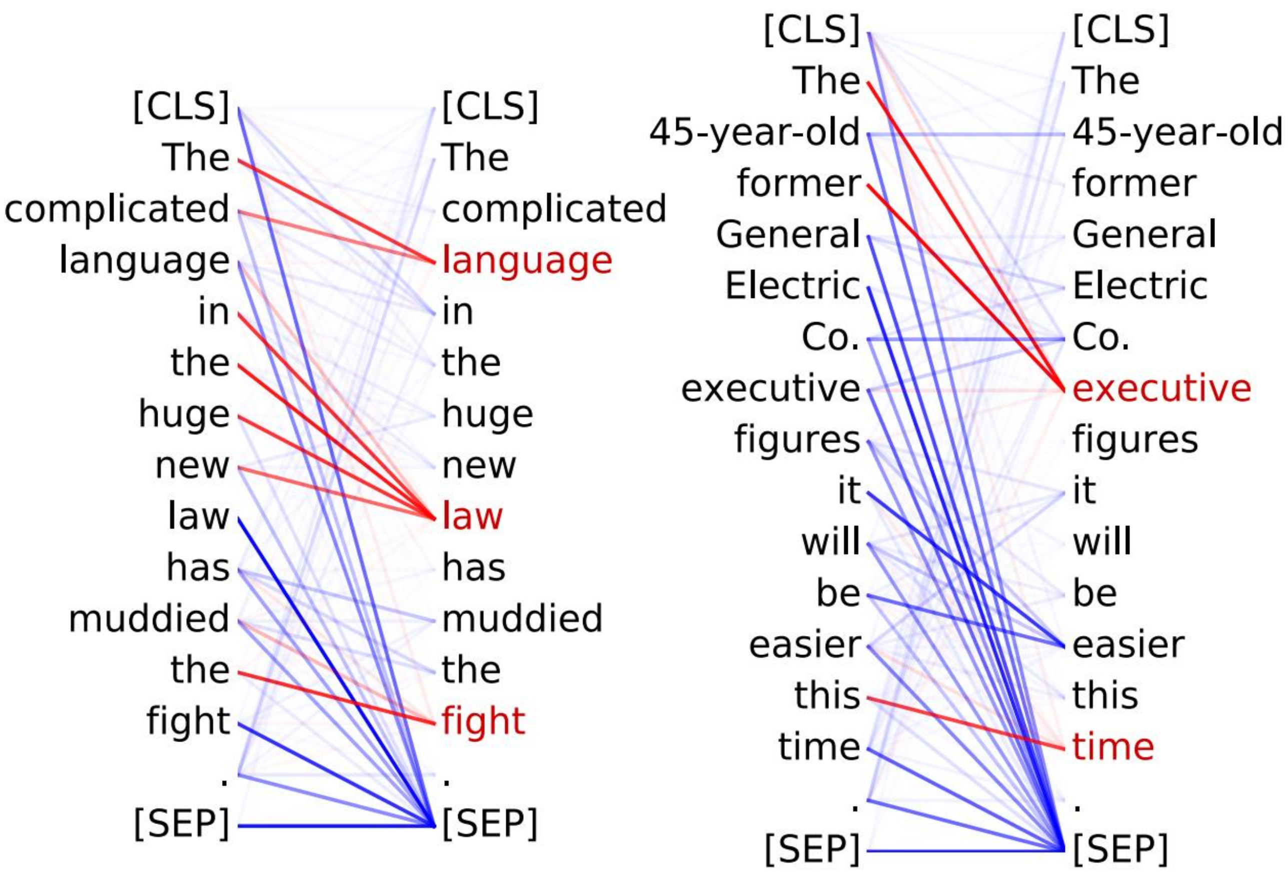

Figure 6: Noun Modifiers Attend to their Noun (Left→Right) [1]

[1] Clark Et al. “What Does BERT Look At? An Analysis of BERT’s Attention.” BlackBoxNLP 2019

BERT Attention Patterns: Summaries

Many language phenomena are captured somewhere in the pretrained parameters

- Some attention head corresponds to linguistic relations
- More captured in pretraining, may not change much in fine-tuning

BERT Attention Patterns: Summaries

Many language phenomena are captured somewhere in the pretrained parameters

- Some attention head corresponds to linguistic relations
- More captured in pretraining, may not change much in fine-tuning

Practical Implications:

- Attention weights reflect the importance perceived by language models
- An effective way to gather feedback from LLMs (handy in later lectures)

Outline

What is captured in BERT?

- Attention patterns
- **Probing capture capabilities in representations**

Why pretrained models generalize?

What does in-context learning do?

Probing Pretraining Representations

Probing what is stored in the representations of pretrained models

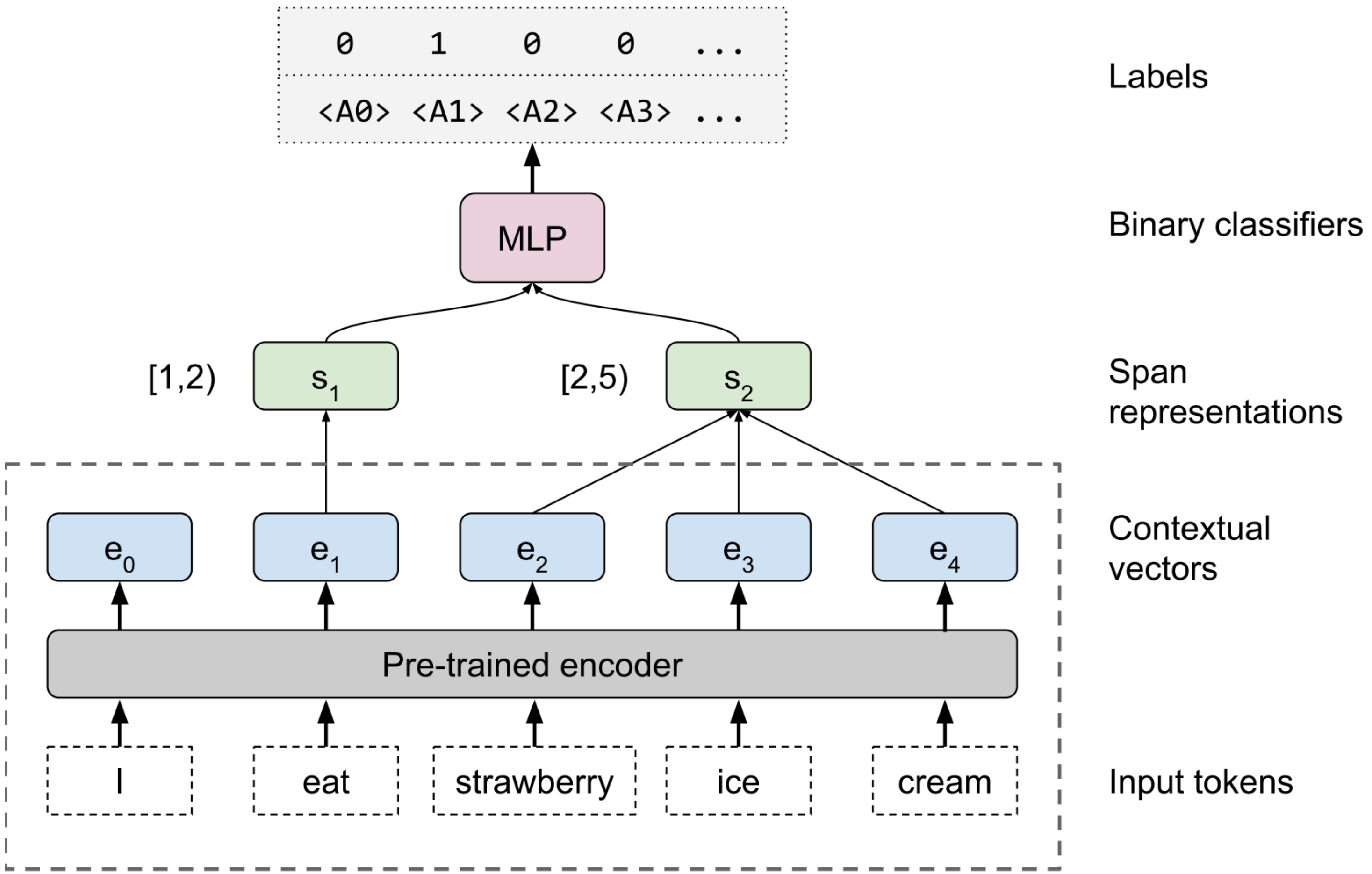

Figure 7: Edge Probing Technique [2]

[2] Tenney, Ian, et al. "What do you learn from context? probing for sentence structure in contextualized word representations." ICLR 2019

Probing Pretraining Representations

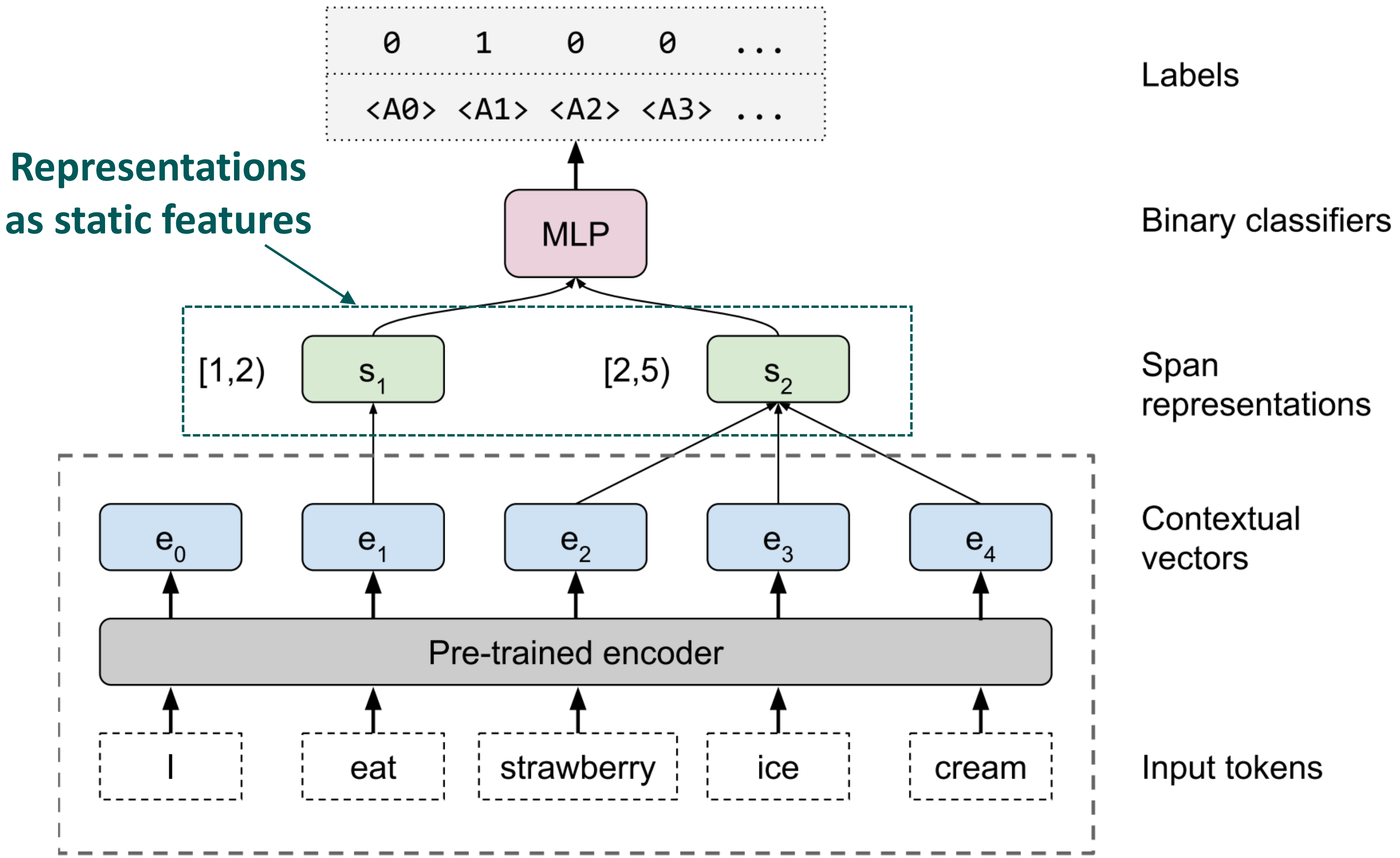

Figure 7: Edge Probing Technique [2]

Mixing representations from layers:

$$\boldsymbol{h}_t^{\text{mix}} = \sum_l w^l \boldsymbol{h}_t^l \,; w^l = \text{softmax}(a^l)$$

- Weighted combination of layers (l)
- Combination weights (a^l) is trained per task with the classification layer

Probing Pretraining Representations

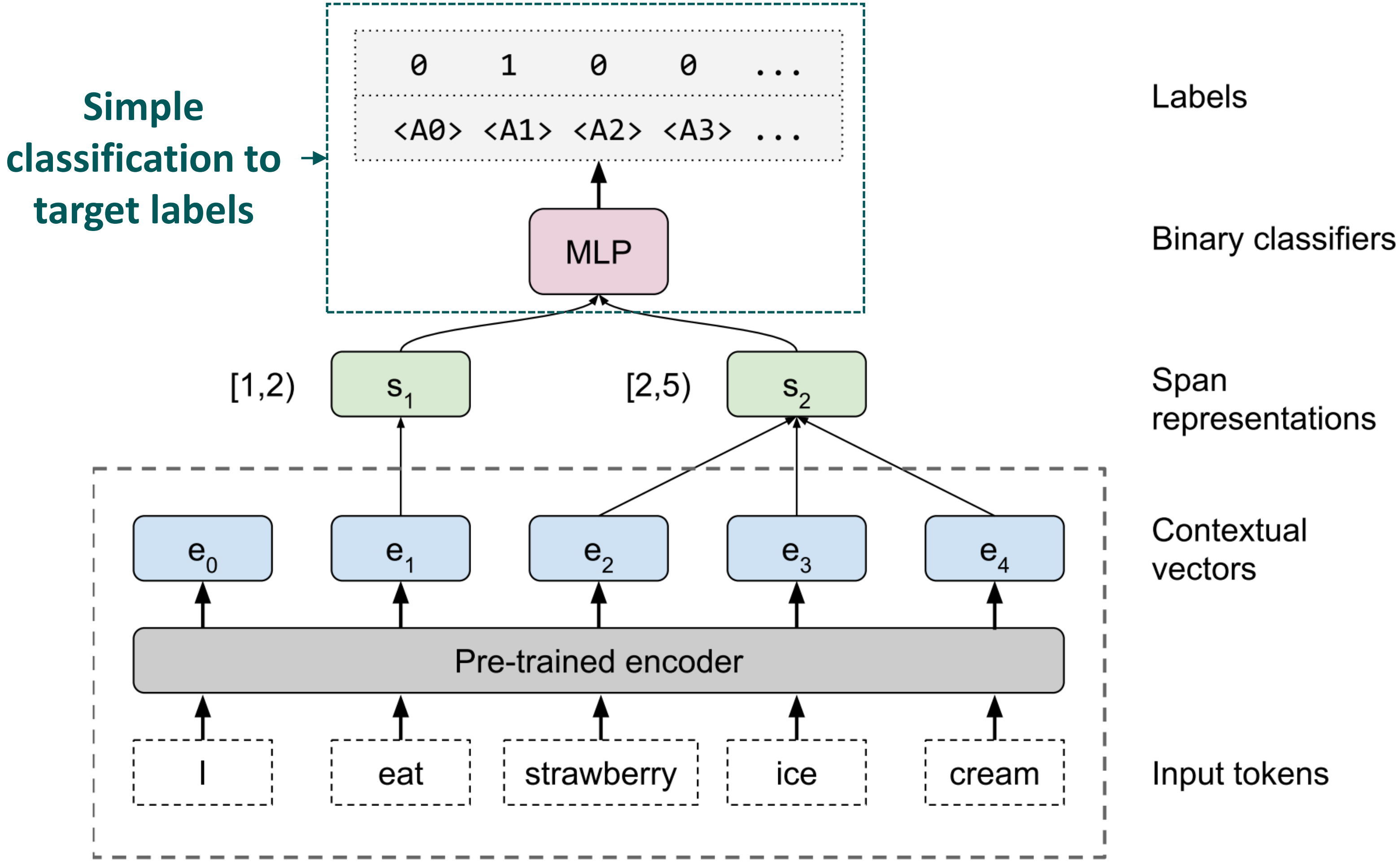

Figure 7: Edge Probing Technique [2]

Mixing representations from layers:

$$\boldsymbol{h}_t^{\text{mix}} = \sum_l w^l \boldsymbol{h}_t^l \, ; w^l = \text{softmax}(a^l)$$

- Weighted combination of layers (l)
- Combination weights (a^l) is trained per task with the classification layer

If the representation perform well

- as static features
- for simple MLP classifier
- In a language task

Then it encodes information required by that task

Probing Pretraining Representations

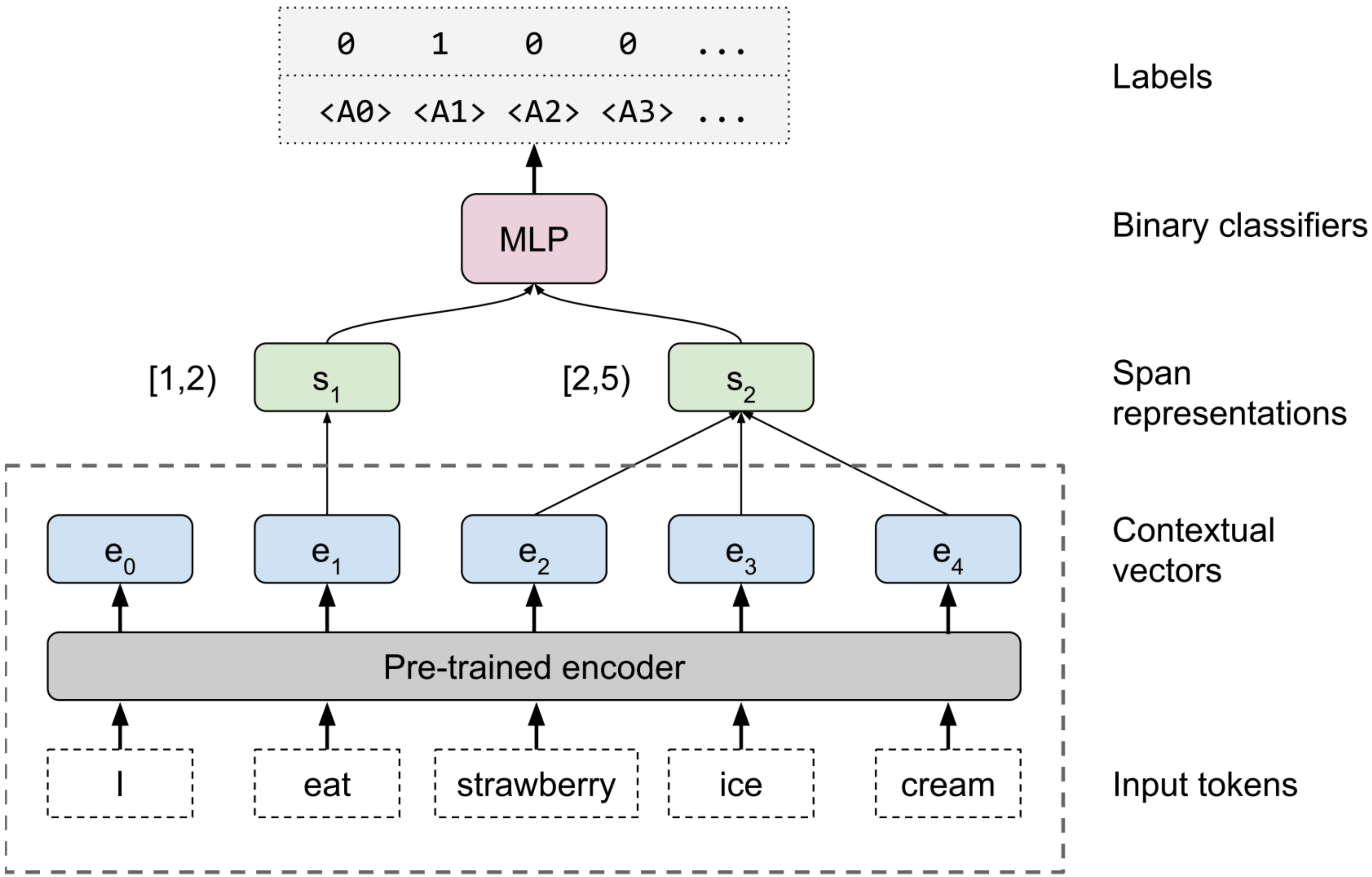

Figure 7: Edge Probing Technique [2]

Mixing representations from layers:

$$\boldsymbol{h}_t^{\text{mix}} = \sum_l w^l \boldsymbol{h}_t^l \,;\, w^l = \text{softmax}(a^l)$$

Center-of-Gravity:

$$E[l] = \sum_l l \cdot w^l$$

- Expected layer to convey the information needed by the probe task
- Larger → information at higher layers

[2] Tenney, Ian, et al. "What do you learn from context? probing for sentence structure in contextualized word representations." ICLR 2019

Probing Pretraining Representations

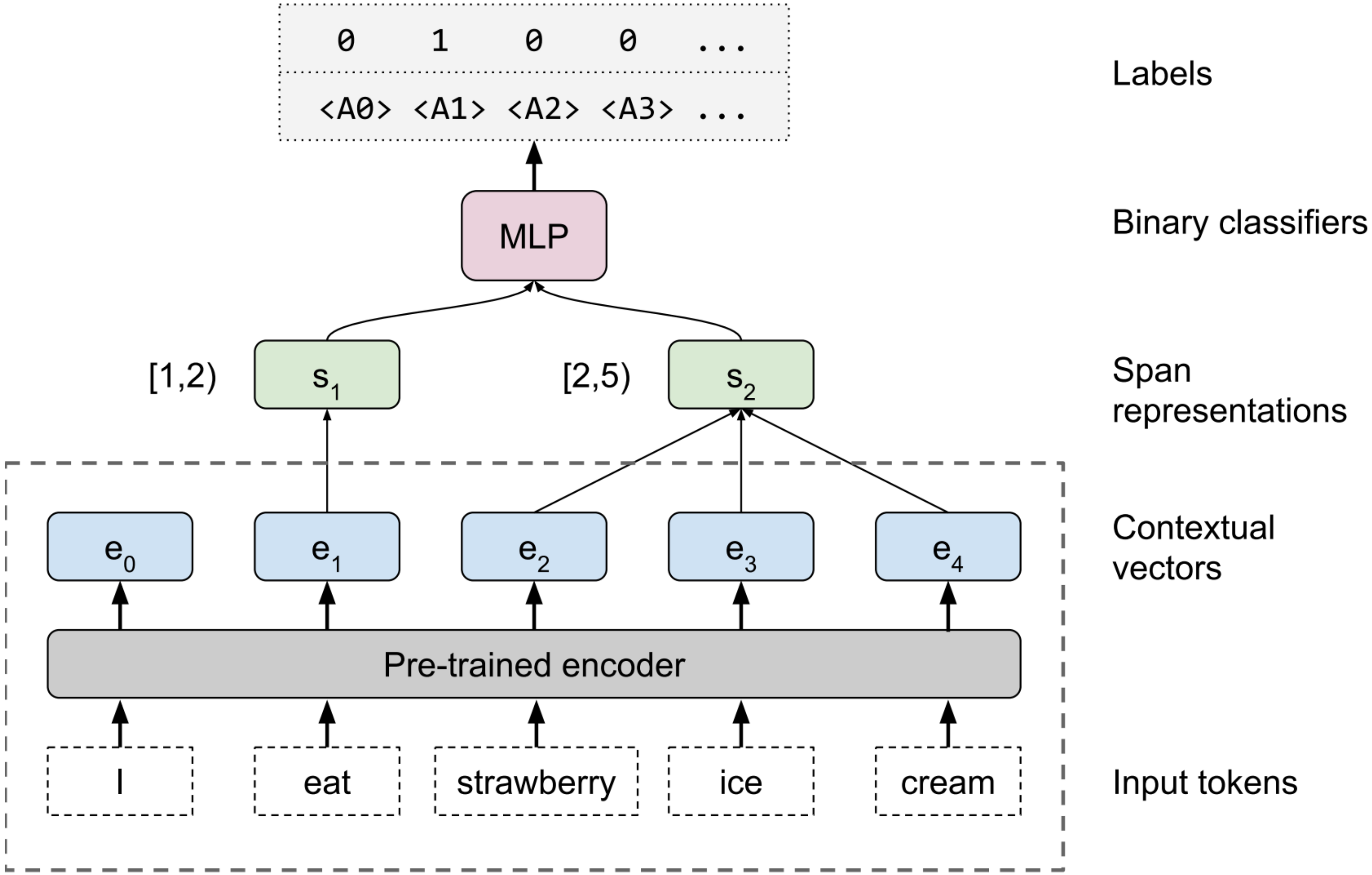

Figure 7: Edge Probing Technique [2]

Mixing representations from layers:

$$\boldsymbol{h}_t^{\text{mix}} = \sum_l w^l \boldsymbol{h}_t^l \, ; w^l = \text{softmax}(a^l)$$

Center-of-Gravity:

$$E[l] = \sum_l l \cdot w^l$$

- Expected layer to convey the information

Expected Layer:

$$\Delta^l = \text{ProbeAcc}(0{:}\, l) - \text{ProbeAcc}\ (0{:}\, l-1)$$

$$E[\Delta^l] = \frac{\sum_l l \cdot \Delta^l}{\sum_l \Delta^l}$$

- Δ^l : The benefit of adding layer l
- $E[\Delta^l]$: The expected layer to solve the probing task

Probing Pretraining Representations: Probing Tasks

Task	Description	Type
Part-of-Speech	Is the token a verb, noun, adj, etc.	Syntactic
Constituent Labeling	Is the span a noun phrase, verb phrase, etc.	Syntactic
Dependency Labeling	Label the functional relationship between tokens, e.g. subject-object?	Syntactic
Named Entity Labeling	Classify the entity type of a span, e.g., person, location, etc.	Syntactic/Semantic
Semantic Role Labeling	Label the predicate-augment structure of a sentence	Semantic
Coreference	Determine the reference of mentions to entities	Semantic
Semantic Proto-Role	Classifier the detailed role of predicate-augment	Semantic
Relation Classification	Predict real-world relations between entities	Semantic/Knowledge

Table 1: Example Language Tasks to Probe BERT [2]

[2] Tenney, Ian, et al. "What do you learn from context? probing for sentence structure in contextualized word representations." ICLR 2019

Probing Pretraining Representations: Probing Results

Table 2: Overall Probing Results [2]

Probing Task	**GPT-1 (base)**	**BERT (base)**	**BERT (Large)**
Part-of-Speech	95.0	96.7	96.9
Constituent Labeling	84.6	86.7	87.0
Dependency Labeling	94.1	85.1	95.4
Named Entity Labeling	92.5	96.2	96.5
Semantic Role Labeling	89.7	91.3	92.3
Coreference	86.3	90.2	91.4
Semantic Proto-Role	83.1	86.1	85.8
Relation Classification	81.0	82.0	82.4
Macro Average	88.3	89.3	91.0

All very good numbers:

- The pretrained representations convey syntactic and sematic information

[2] Tenney, Ian, Dipanjan Das, and Ellie Pavlick. "BERT Rediscovers the Classical NLP Pipeline." ACL. 2019.

Probing Pretraining Representations: Across Layers

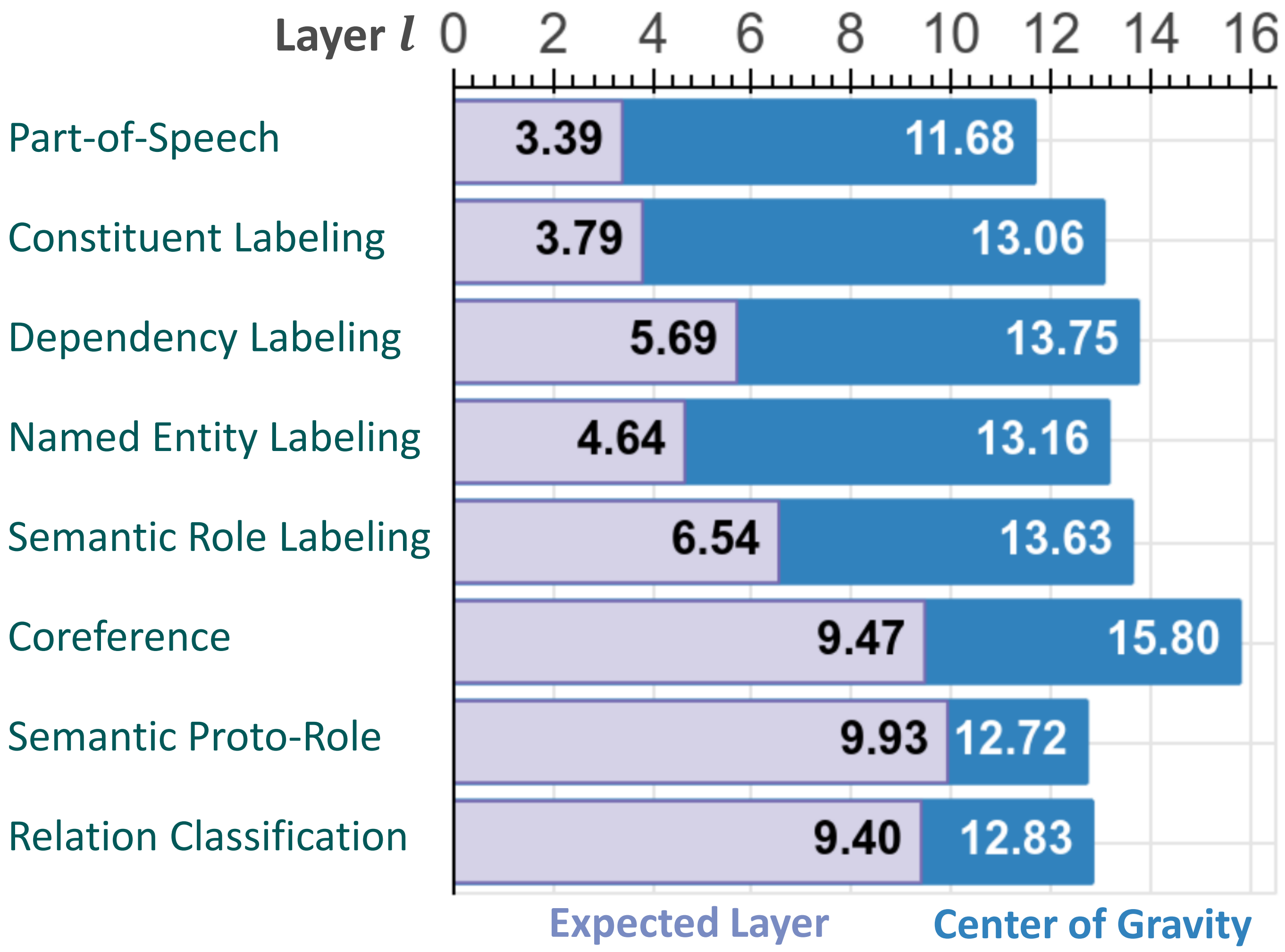

Figure 8: Edge Probing Results of BERT Large [3].

Mixing representations from layers:

$$\boldsymbol{h}_t^{\text{mix}} = \sum_l w^l \boldsymbol{h}_t^l \,; w^l = \text{softmax}(a^l)$$

Center-of-Gravity:

$$E[l] = \sum_l l \cdot w^l$$

- Expected layer to convey the information

Expected Layer:

$$\Delta^l = \text{ProbeAcc}(0{:}\,l) - \text{ProbeAcc}\,(0{:}\,l-1)$$

$$E[\Delta^l] = \frac{\sum_l l \cdot \Delta^l}{\sum_l \Delta^l}$$

- Δ^l : The benefit of adding layer l
- $E[\Delta^l]$: The expected layer to solve the probing task

Probing Pretraining Representations: Across Layers

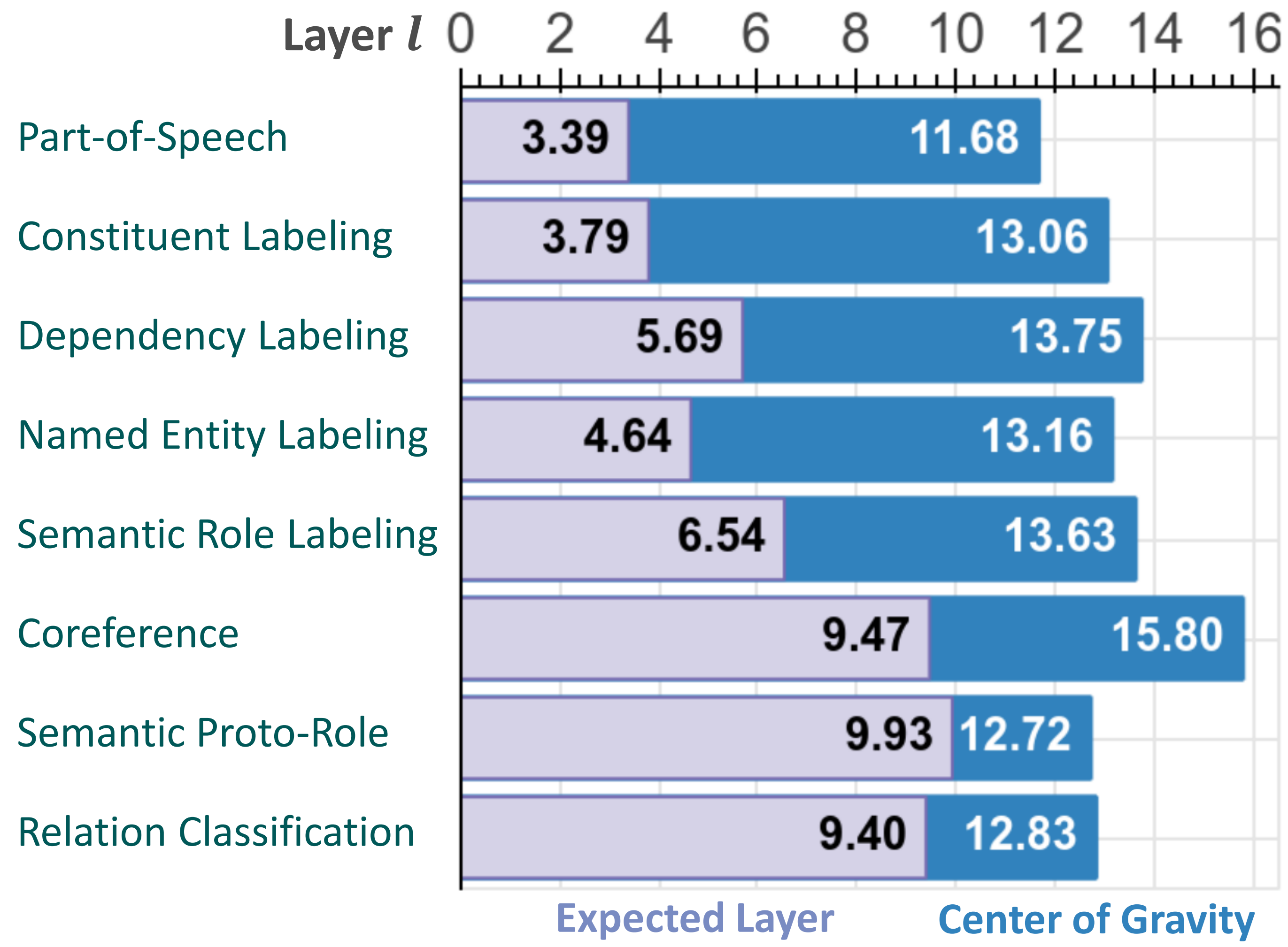

Figure 8: Edge Probing Results of BERT Large [3].

Different tasks are tackled at different layers

- Syntactic tasks at lower layers
- Semantic/Knowledge tasks at higher ones

[3] Tenney, Ian, Dipanjan Das, and Ellie Pavlick. "BERT Rediscovers the Classical NLP Pipeline." ACL. 2019.

Probing Pretraining Representations: Across Training Steps

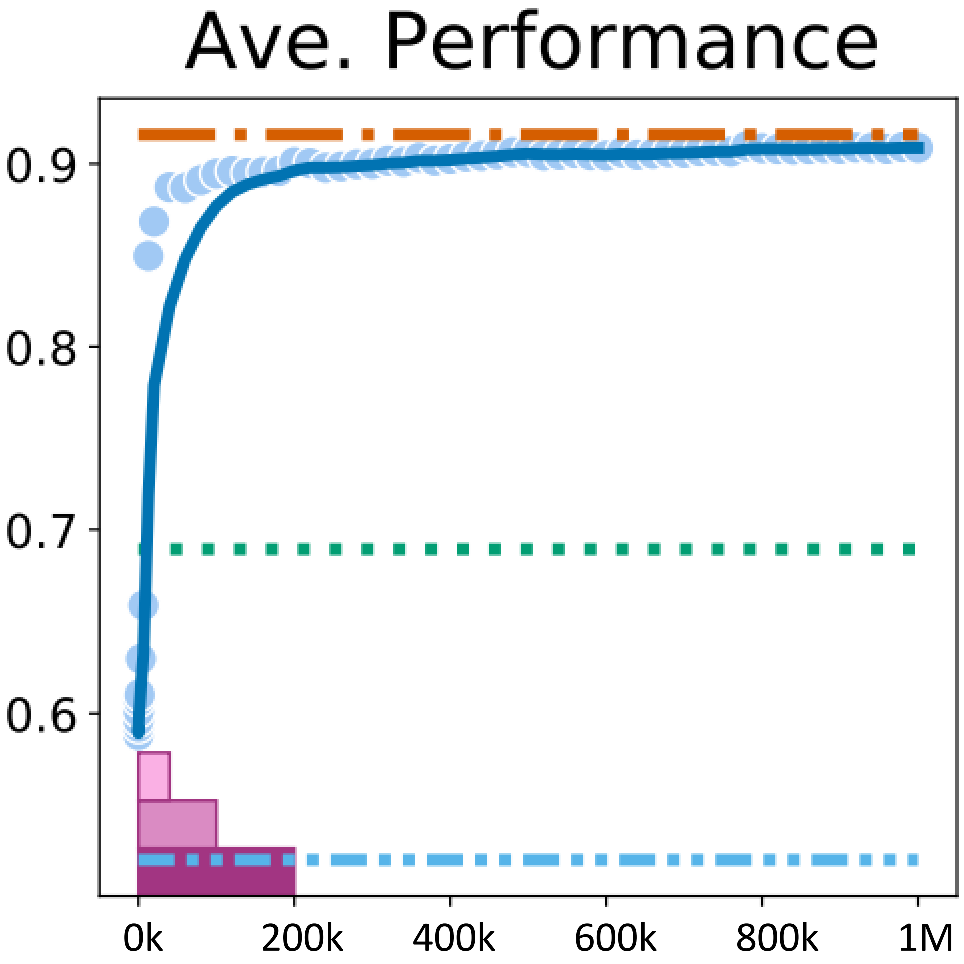

Figure 9: Linguistics Task Probing at RoBERTa Pretraining Steps [4].

Example Linguistic Tasks:

- Part-of-Speech
- Named Entity Labeling
- Syntactic Chunking

Probing Pretraining Representations: Across Training Steps

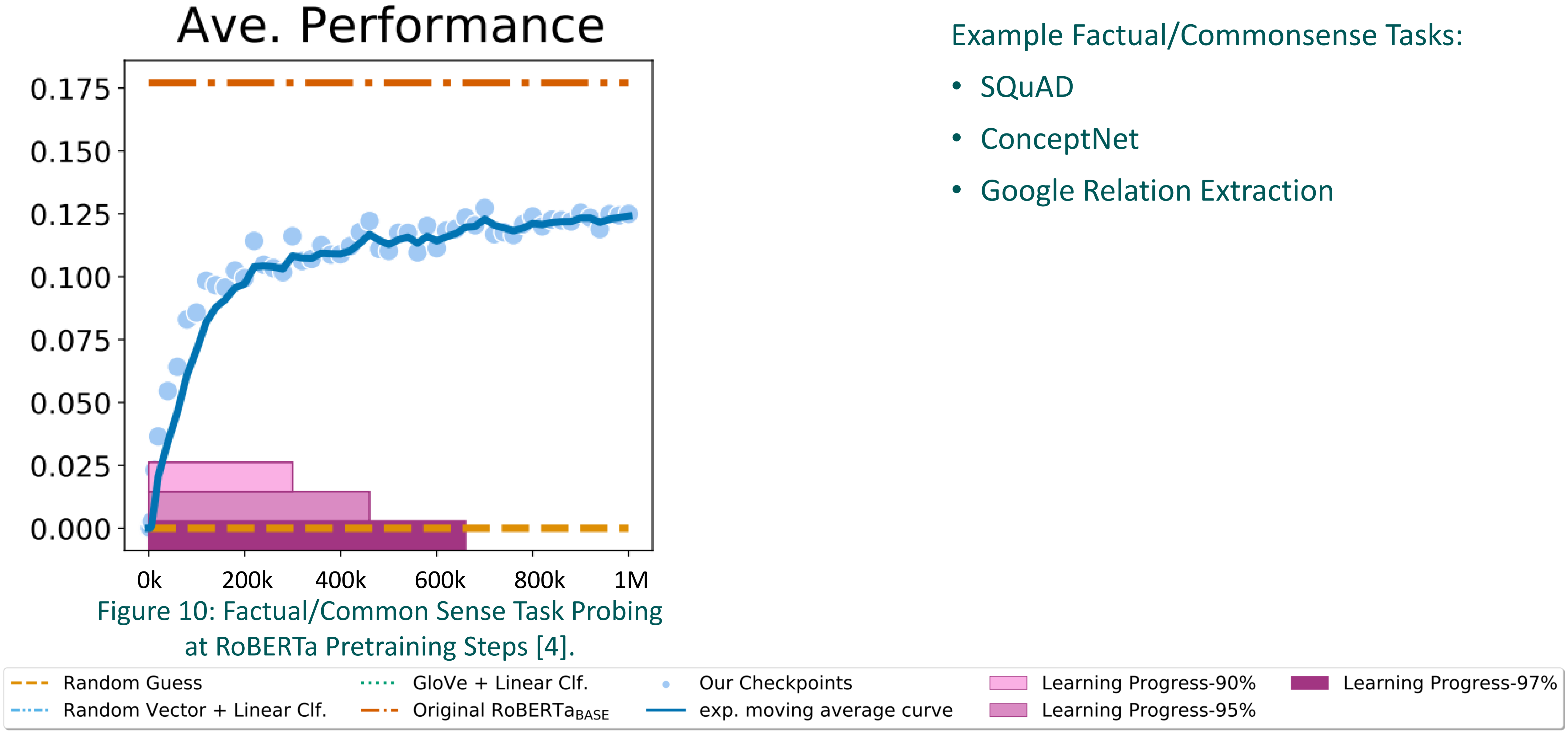

Figure 10: Factual/Common Sense Task Probing at RoBERTa Pretraining Steps [4].

Example Factual/Commonsense Tasks:

- SQuAD
- ConceptNet
- Google Relation Extraction

[4] Liu, et al. "Probing Across Time: What Does RoBERTa Know and When?." EMNLP 2021.

Probing Pretraining Representations: Across Training Steps

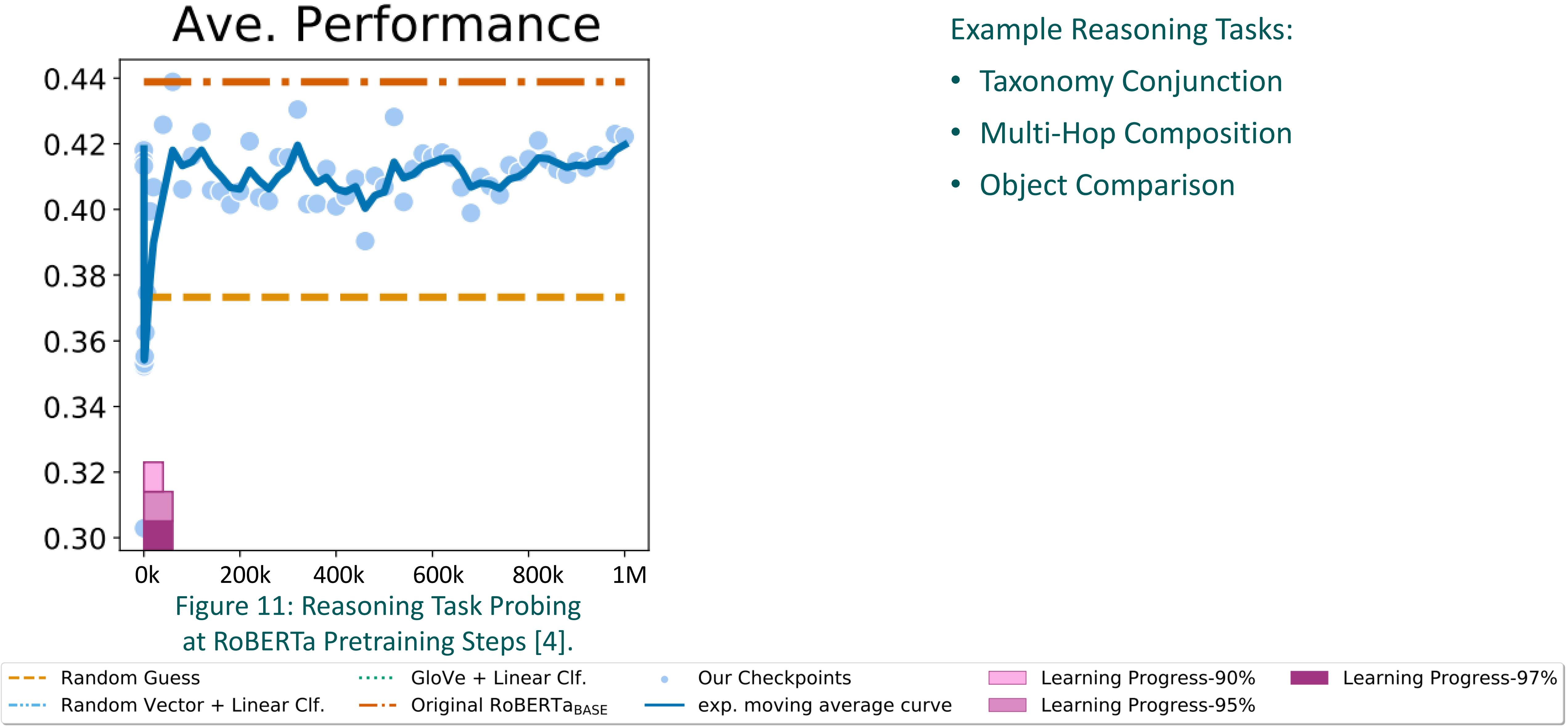

Figure 11: Reasoning Task Probing at RoBERTa Pretraining Steps [4].

Example Reasoning Tasks:

- Taxonomy Conjunction
- Multi-Hop Composition
- Object Comparison

[4] Liu, et al. "Probing Across Time: What Does RoBERTa Know and When?." EMNLP 2021.

Probing Pretraining Representations: Across Training Steps

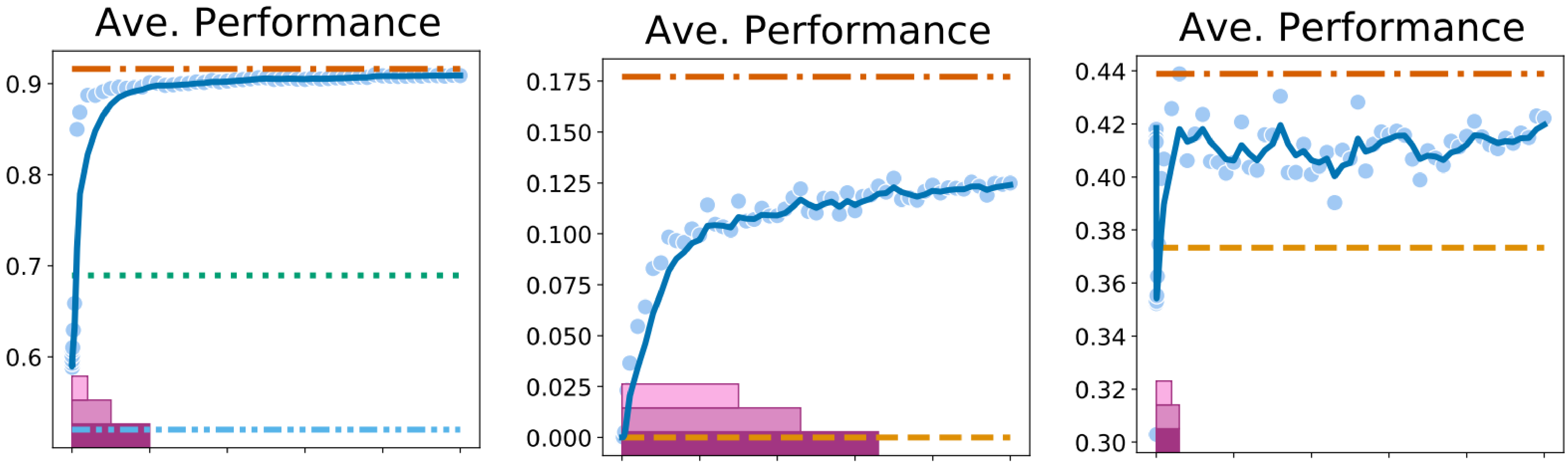

Figure 11: Probing at Pretraining steps in Linguistic (left), Factual/Commonsense (middle), and Reasoning (right) tasks [4]

- Capturing tasks at different conceptual difficulty at different rate
- Emergent improvements
- Certain tasks require certain scale

Probing Pretraining Representations: Summary

From the observatory point of view:

- Some attention patterns are intuitive
- Pretrained representations convey strong language information
- Different tasks are captured at different layers and different steps
- And the conceptual difficulty of tasks aligns with where & when they are captured

Probing Pretraining Representations: Summary

From the observatory point of view:

- Some attention patterns are intuitive
- Pretrained representations convey strong language information
- Different tasks are captured at different layers and different steps
- And the conceptual difficulty of tasks aligns with where & when they are captured

It is tempting to think language models capture language semantics from a ground up way:

Syntactic →Semantic → Factual → Reasoning →General Intelligence

- Like a classic NLP pipeline
- Like how human brains learn natural language

Probing Pretraining Representations: Summary

From the observatory point of view:

- Some attention patterns are intuitive
- Pretrained representations convey strong language information
- Different tasks are captured at different layers and different steps
- And the conceptual difficulty of tasks aligns with where & when they are captured

It is tempting to think language models capture language semantics from a ground up way:

Syntactic →Semantic → Factual → Reasoning →General Intelligence

- Like a classic NLP pipeline
- Like how human brains learn natural language

But:

- Classic NLP tasks are not really ground up, best systems are often more direct & straightforward
- We really do not know how human brains work, perhaps less than we know how LLM works

Practical implications:

- Efficient inference by only using what is needed: early exist, sparsity, distillation, etc.

Outline

What is captured in BERT?

Why pretrained models generalize?

- Loss landscapes
- Implicit bias of language models

What does in-context learning do?

Understand Generation Ability: Overview

Why pretrained models generalize to many fine-tuning tasks?

- Even on tasks with sufficient supervised label

Why larger models and longer pretraining steps improve generalization?

- In statistical machine learning: more complicated model + exhaustive training is recipe for overfitting
- But they indeed are the core advantages of pretraining models

Visualization of Loss Landscape

Plot the loss function around a model parameter θ

- Challenge: θ is super high dimension

Approximation: plot the loss landscape of θ towards two other parameters θ_1 and θ_2 [5]

$$f(\alpha, \beta) = \mathrm{loss}(\theta + \alpha(\theta_1 - \theta) + \beta(\theta_2 - \theta))$$

- A plot along the axes of α and β the linear interpolation

[5] Li, et al. "Visualizing the loss landscape of neural nets.“ NeurIPS 2018.

Visualization of Loss Landscape

Plot the loss function around a model parameter θ

- Challenge: θ is super high dimension

Approximation: plot the loss landscape of θ towards two other parameters θ_1 and θ_2 [5]

$$f(\alpha, \beta) = \text{loss}(\theta + \alpha(\theta_1 - \theta) + \beta(\theta_2 - \theta))$$

- A plot along the axes of α and β the linear interpolation

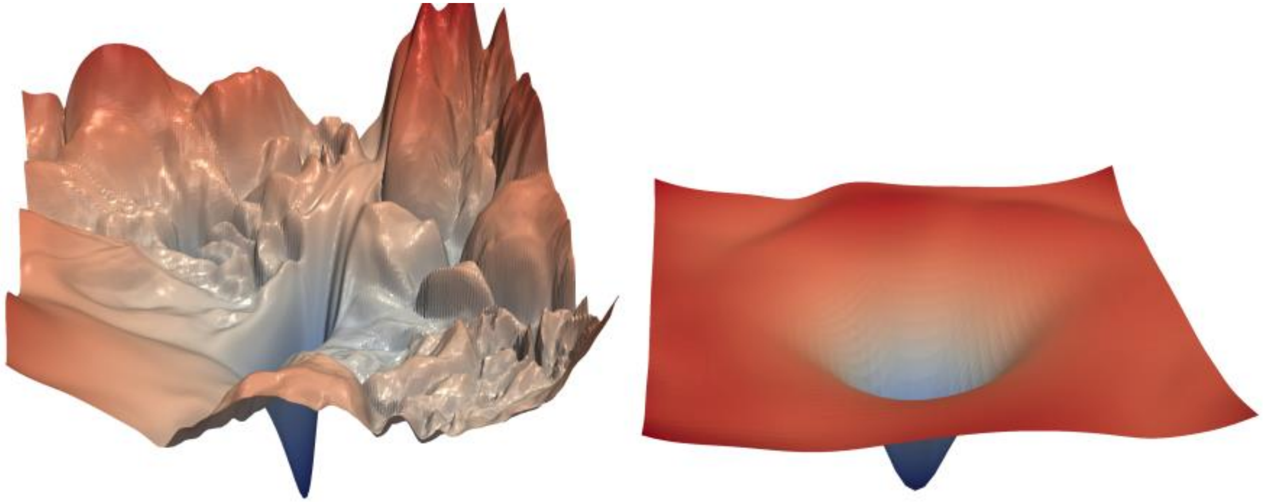

Figure 12: A sharp loss landscape and a smooth loss landscape [5]

[5] Li, et al. "Visualizing the loss landscape of neural nets." NeurIPS 2018.

Visualization of Loss Landscape: BERT

BERT landscape in finetuning [6]

$$f(\alpha, \beta) = \text{loss}(\theta + \alpha(\theta_1 - \theta) + \beta(\theta_2 - \theta))$$

- θ starting parameter of fine-tuning: pretrained or random initialized
- θ_1 the finetuned parameter of this task
- θ_2 the finetuned parameter of another task, which is meaningful

[6] Hao, Yaru, et al. "Visualizing and Understanding the Effectiveness of BERT." EMNLP 2019.

Visualization of Loss Landscape: BERT

BERT landscape in finetuning [6]

$$f(\alpha, \beta) = \text{loss}(\theta + \alpha(\theta_1 - \theta) + \beta(\theta_2 - \theta))$$

- θ starting parameter of fine-tuning: pretrained or random initialized
- θ_1 the finetuned parameter of this task
- θ_2 the finetuned parameter of another task, which is meaningful

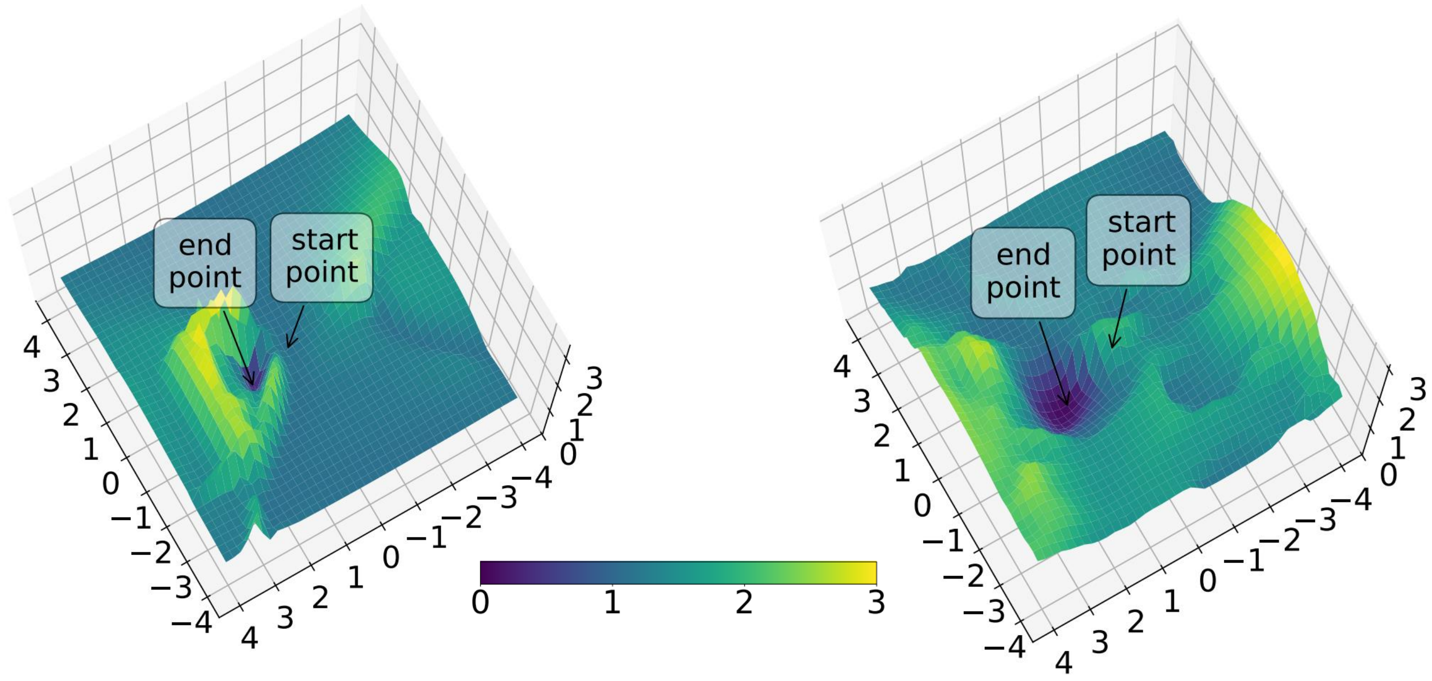

Figure 13: Loss landscape of finetuning MNLI from random or pretrained BERT [6]

Visualization of Loss Landscape: BERT

BERT landscape in finetuning [6]

$$f(\alpha, \beta) = \text{loss}(\theta + \alpha(\theta_1 - \theta) + \beta(\theta_2 - \theta))$$

- θ starting parameter of fine-tuning: pretrained or random initialized
- θ_1 the finetuned parameter of this task
- θ_2 the finetuned parameter of another task, which is meaningful

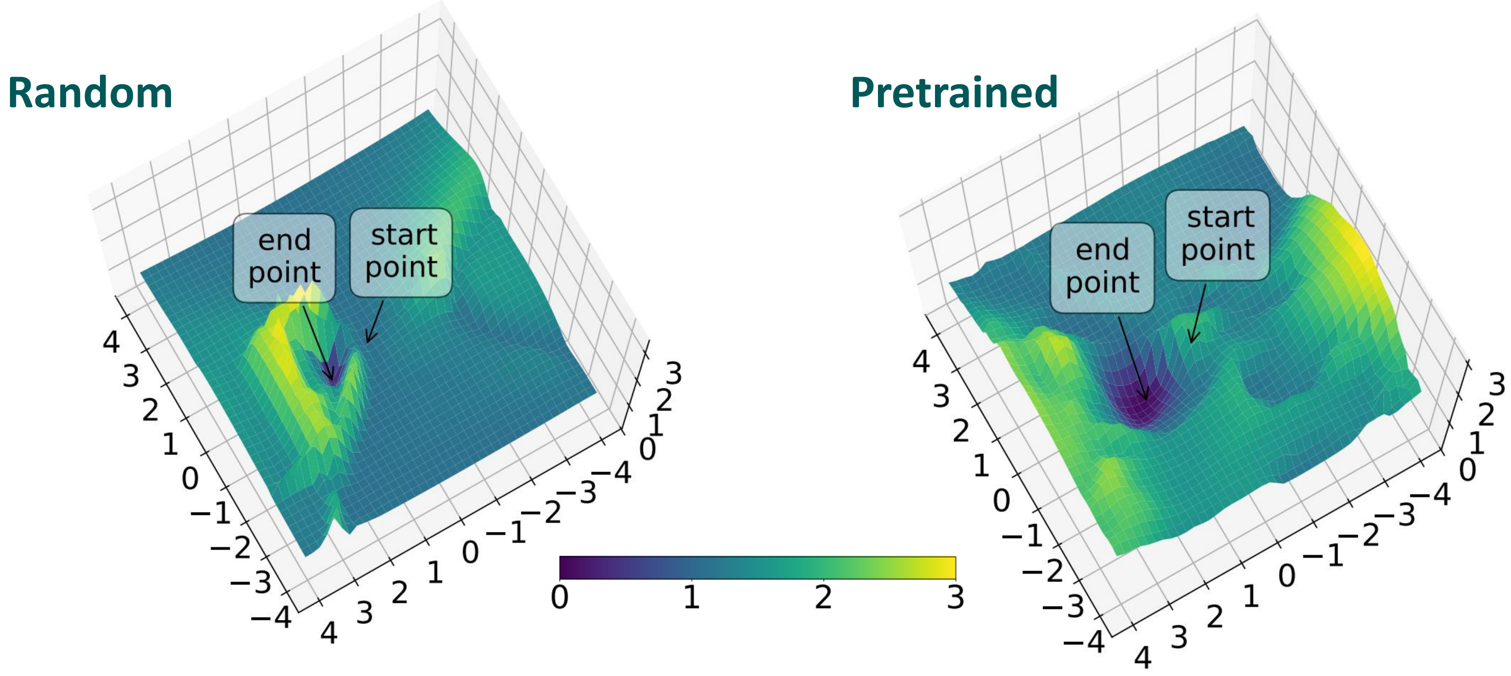

Figure 13: Loss landscape of finetuning MNLI from random or pretrained BERT [6]

Visualization of Loss Landscape: BERT

Plot the optimization path: project the checkpoint θ' at different steps to the loss landscape

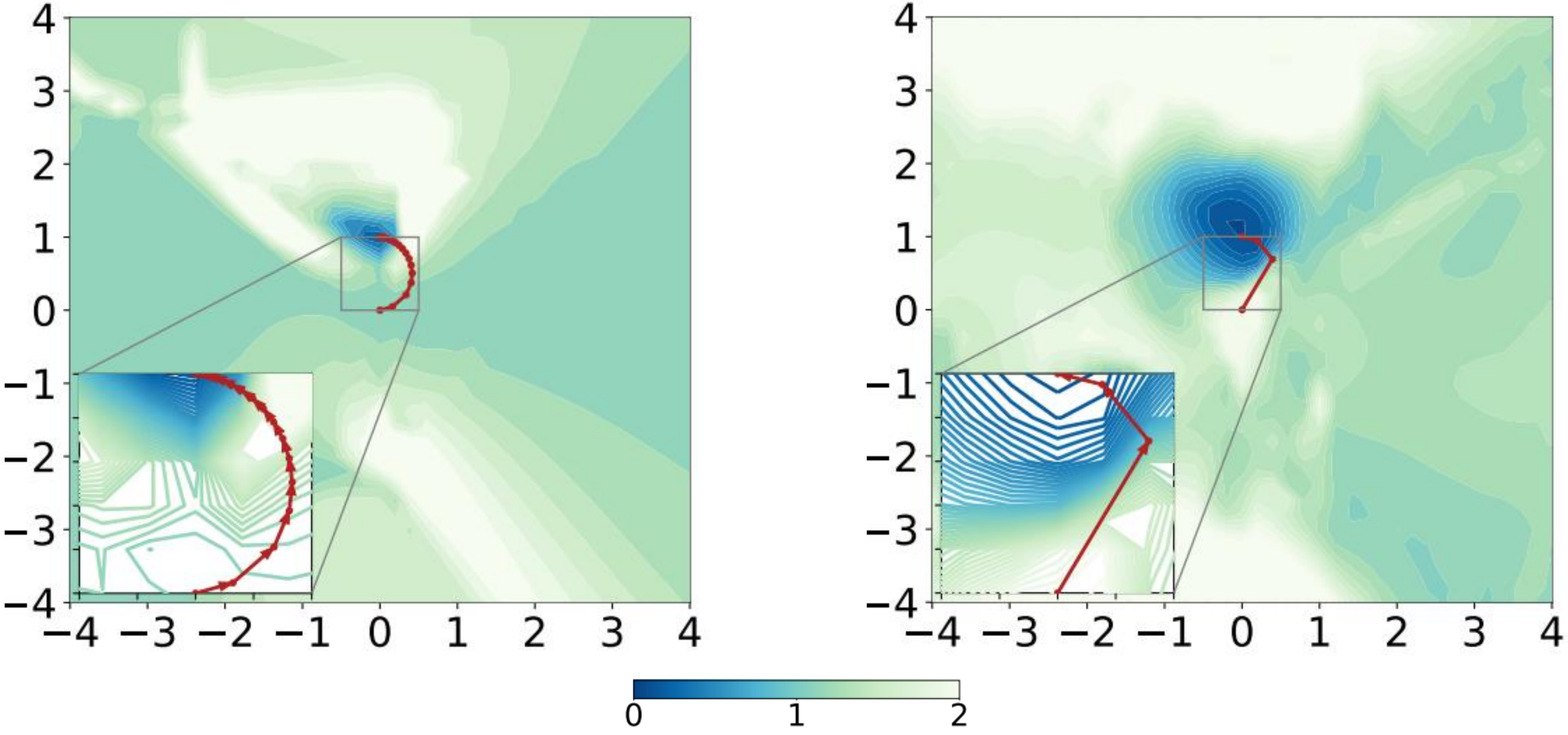

Figure 14: Optimization Trajectory when finetuning MNLI from random (left) and pretrained (right) BERT [6]

Outline

What is captured in BERT?

Why pretrained models generalize?

- Loss landscapes
- **Implicit bias of language models**

What does in-context learning do?

Inductive Bias of Language Models: Pretraining Longer

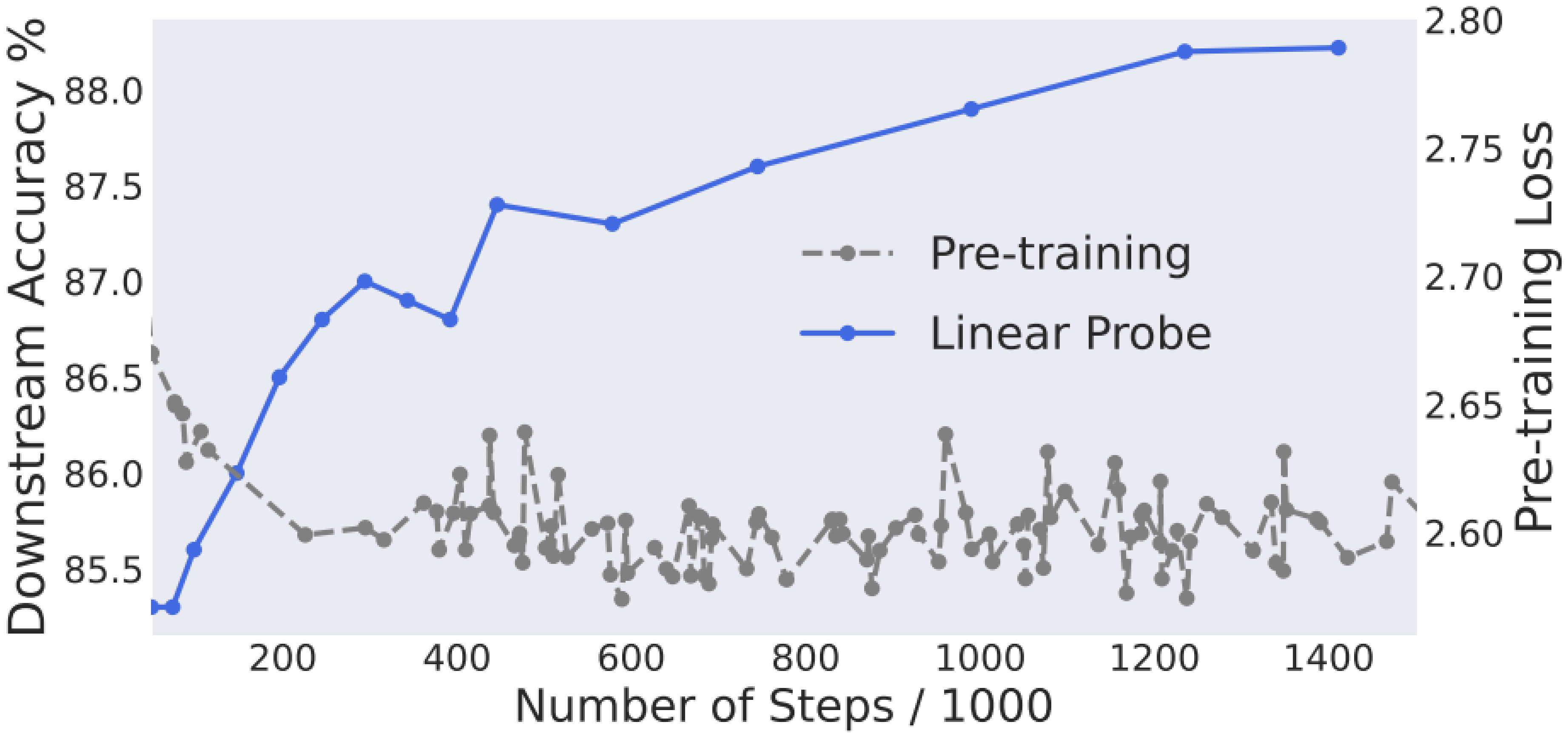

Figure 15: Probing Performances versus Pretraining Loss of a 25M Parameter BERT [7]

[7] Liu, Hong, et al. "Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models." ICML 2023.

Inductive Bias of Language Models: Pretraining Longer

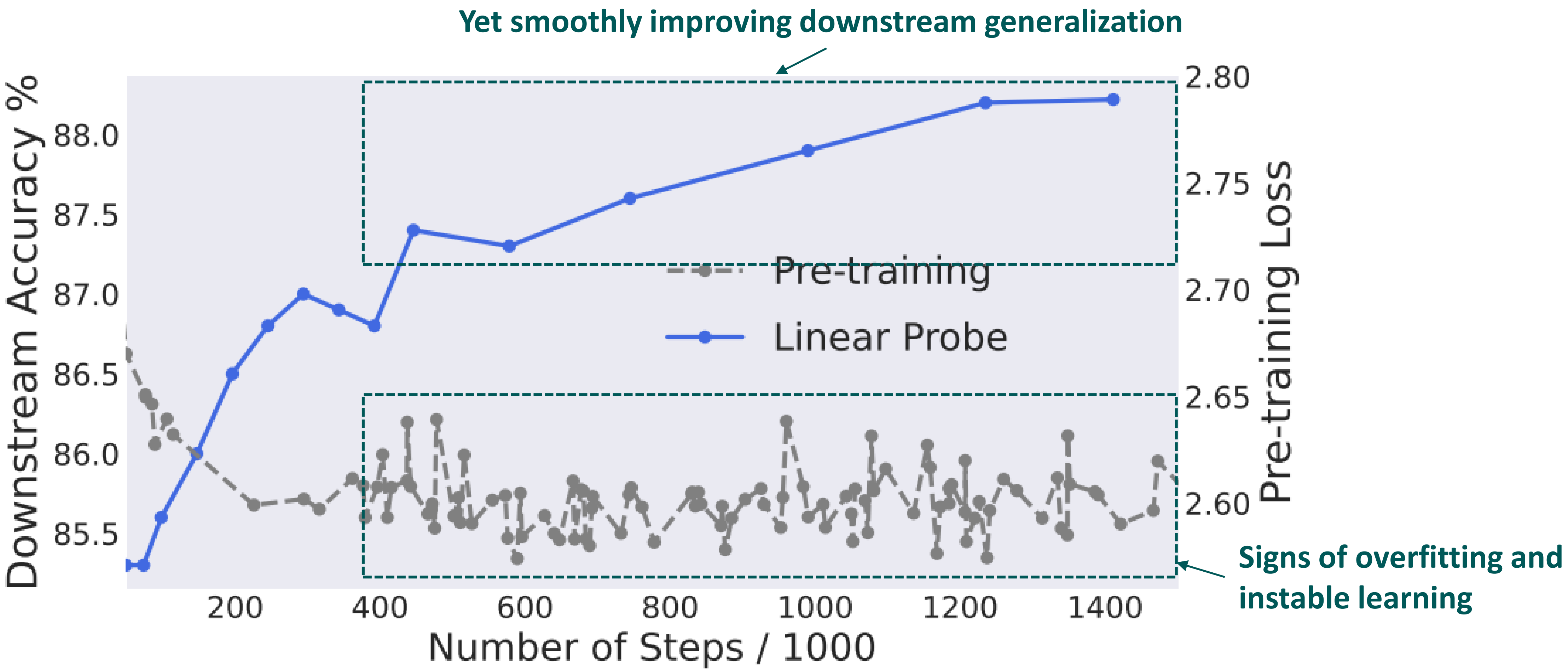

Figure 15: Probing Performances versus Pretraining Loss of a 25M Parameter BERT [7]

[7] Liu, Hong, et al. "Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models." ICML 2023.

Inductive Bias of Language Models: Pretraining Longer

Same pretraining loss but flattener loss shape

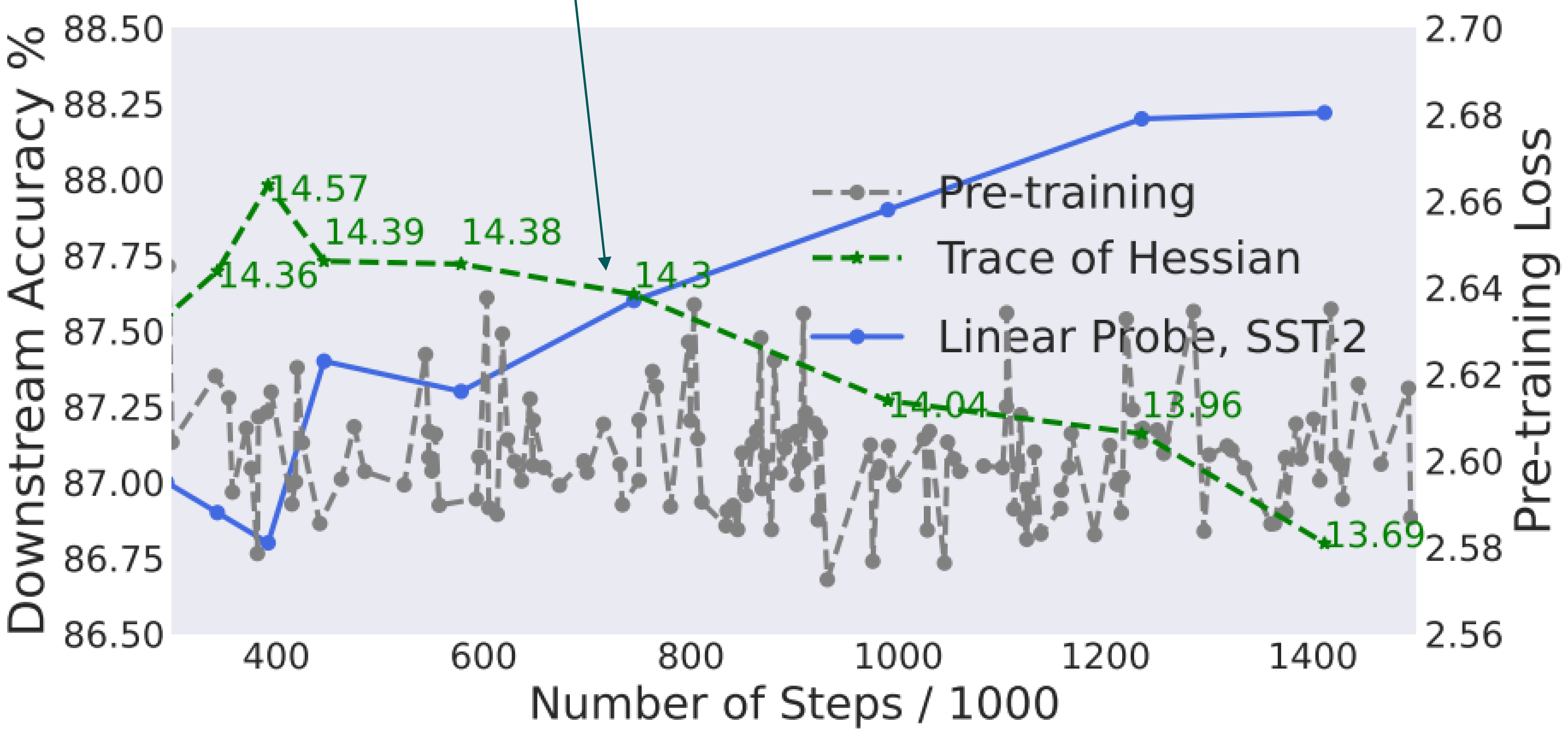

Figure 15: Probing Performances versus Pretraining Loss of a 25M Parameter BERT [7]

Trace of (Loss) Hessian: A reflection of the loss flatness

Inductive Bias of Language Models: Larger Models

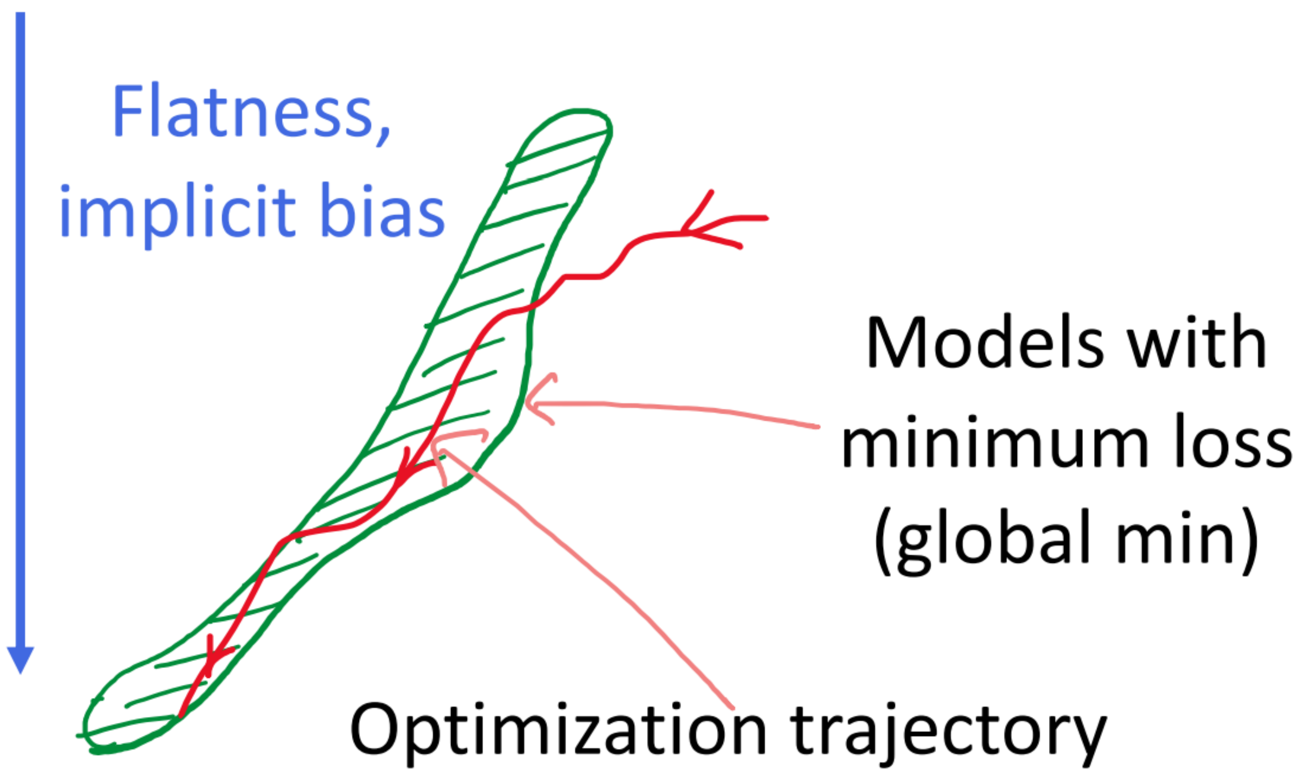

Figure 16: Illustration of Optimization Trajectory [7]

[7] Liu, Hong, et al. "Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models." ICML 2023.

Inductive Bias of Language Models: Larger Models

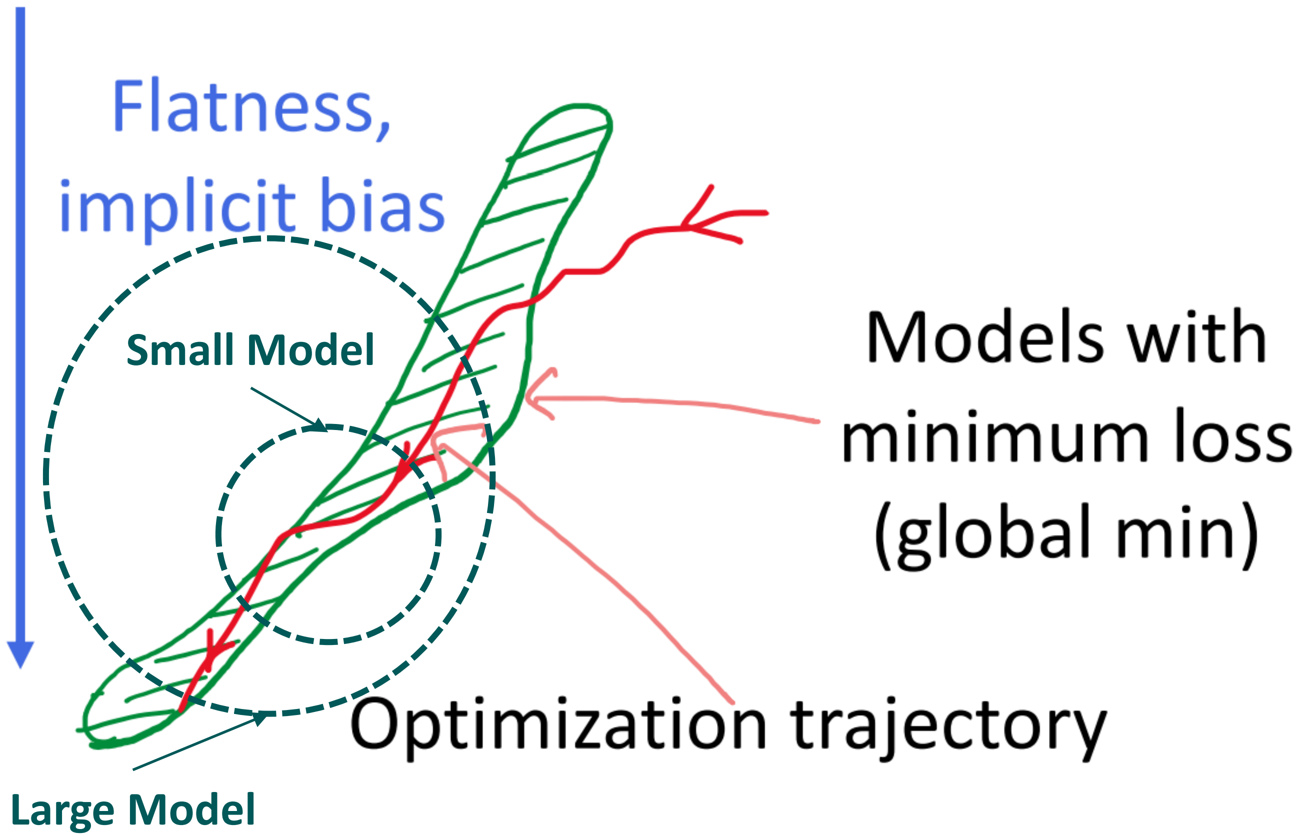

Figure 16: Illustration of Optimization Trajectory [7]

Larger models can reach a flattener optima:

1. Larger transformers have bigger solution space
2. They cover smaller transformers
3. Optimizer keep seeking for flattener optima, even reached same loss

[7] Liu, Hong, et al. "Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models." ICML 2023.

Why Pretrained Models Generalize: Summary

Many observations on pretrained models lead to flatter optima

- Better starting point
- Better loss shape
- Pretraining longer and larger Transformers lead to more flatness

Why Pretrained Models Generalize: Summary

Many observations on pretrained models lead to flatter optima

- Better starting point
- Better loss shape
- Pretraining longer and larger Transformers lead to more flatness

Why flatness matters?

- Many empirical evidences showing its connection to generalization ability
- Intuitively, more robust to data variations/noises
- Theoretically, argued that it leads to simpler network solutions
    - Hochreiter, S. and Schmidhuber, J. Flat minima. Neural Computing 1997

Why Pretrained Models Generalize: Summary

Many observations on pretrained models lead to flatter optima

- Better starting point
- Better loss shape
- Pretraining longer and larger Transformers lead to more flatness

Why flatness matters?

- Many empirical evidences showing its connection to generalization ability
- Intuitively, more robust to data variations/noises
- Theoretically, argued that it leads to simpler network solutions
    - Hochreiter, S. and Schmidhuber, J. Flat minima. Neural Computing 1997

Why pretrained models prefer flatter optima?

- A inductive bias of the optimizer, the architecturer, the pretraining loss, or the combination of them?
- Much more research required

Outline

What is captured in BERT?

Why pretrained models generalize?

What does in-context learning do?

- Semantic Prior or Input-Label Mapping
- Connection with Gradient Decent

In-Context Learning Interpretation: Observations

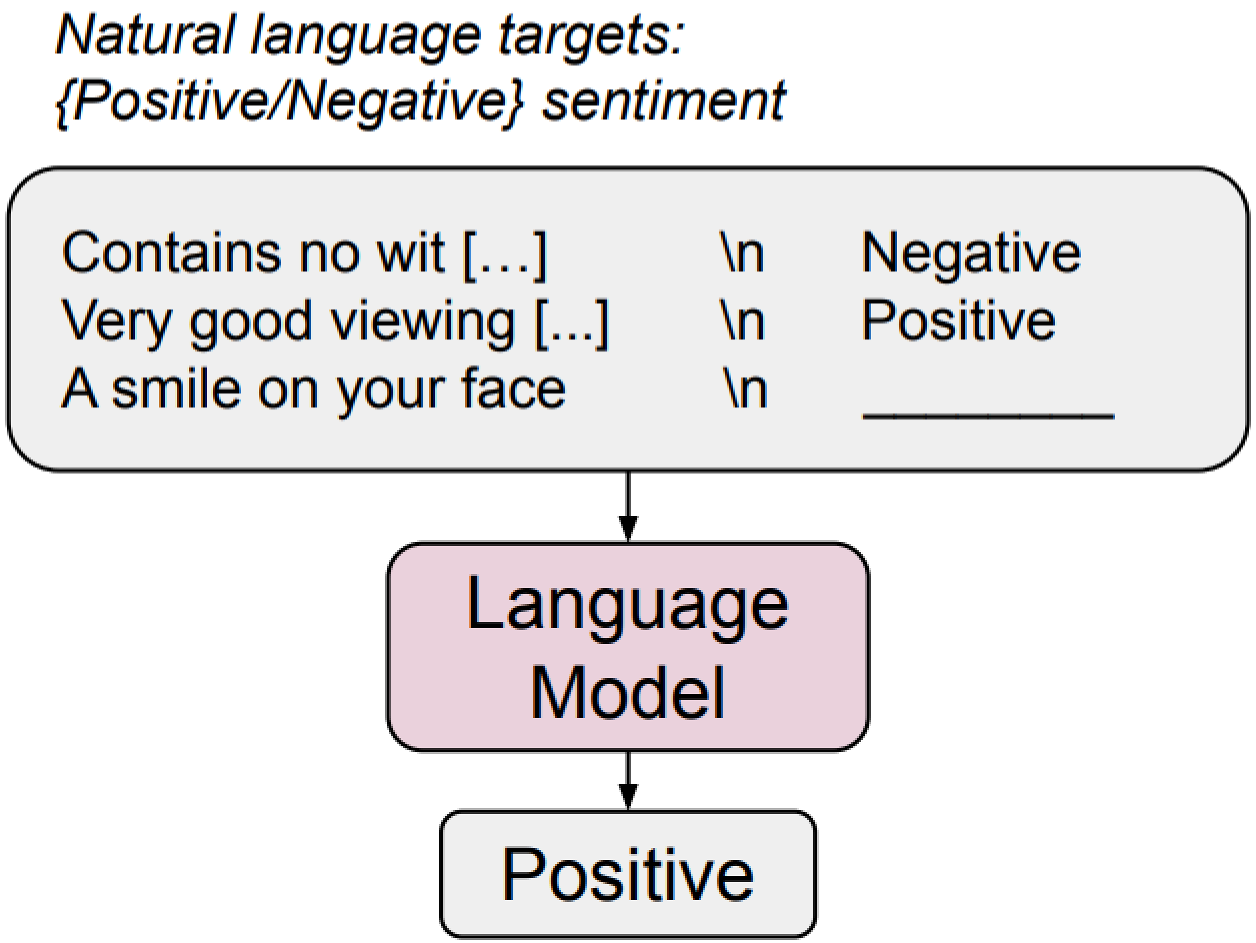

Figure 17: Regular In-Context Learning [8]

Two sources of information:

- Semantic knowledge captured in LLM
- In-context training signals (input-label mapping)

[8] Wei, et al. "Larger language models do in-context learning differently." arXiv 2023.

In-Context Learning Interpretation: Observations

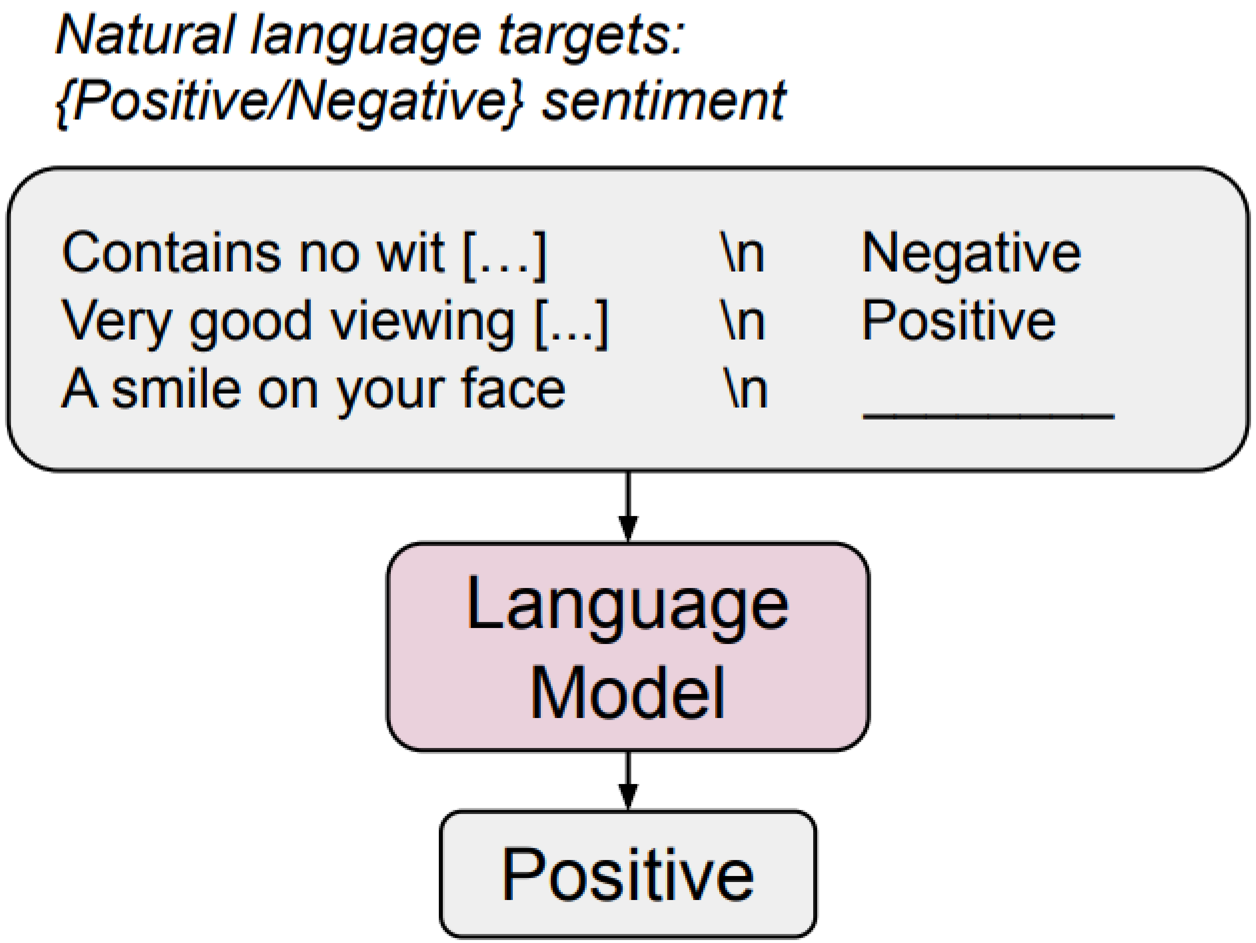

Figure 17: Regular In-Context Learning [8]

Two sources of information:

- Semantic knowledge captured in LLM
- In-context training signals (input-label mapping)

Which one works? Mixed observations:

- Random in-context labels work

→ Existing semantic knowledge

- Order of in-context data matter

→ In-context training signals

In-Context Learning Interpretation: Random Label Test

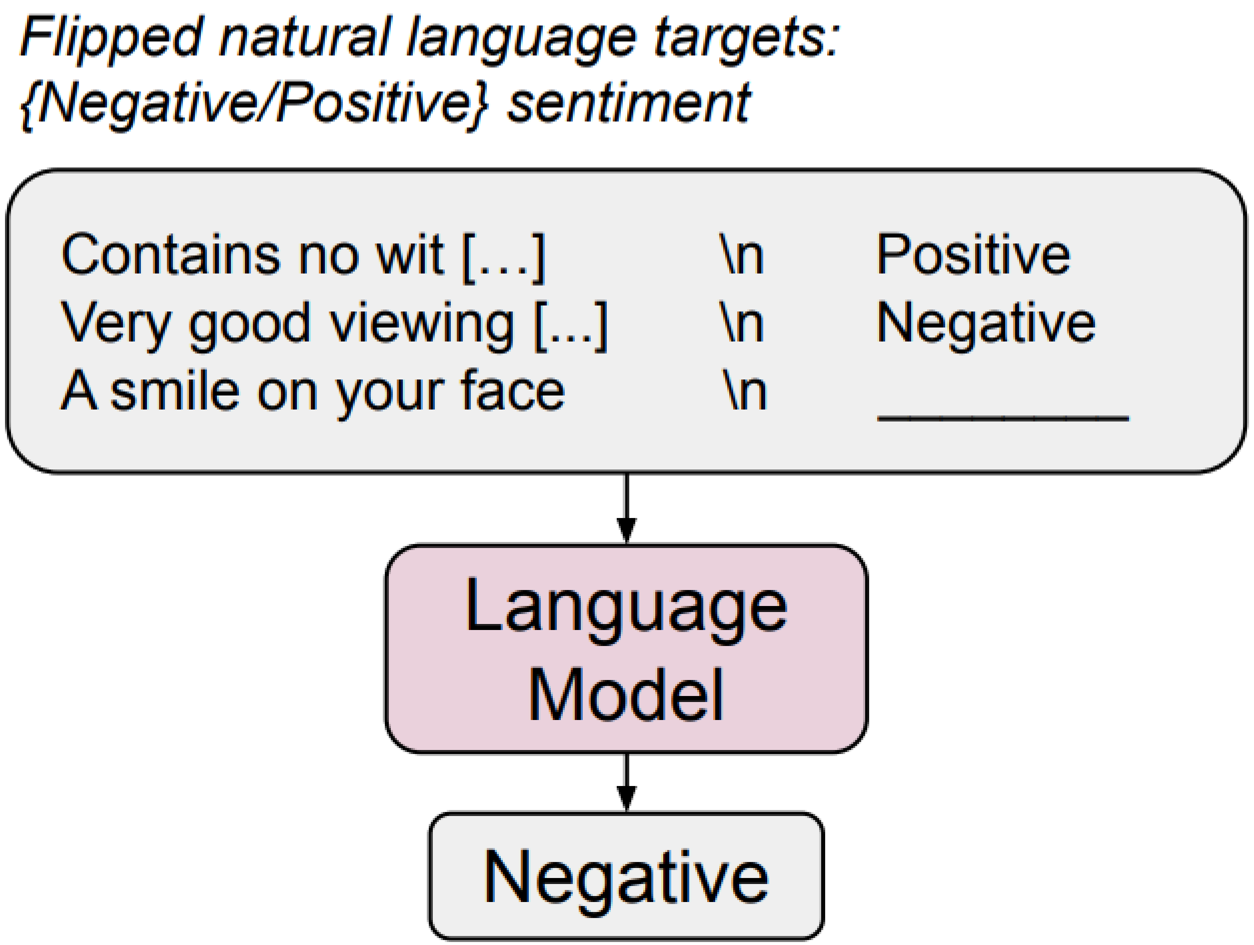

Figure 18: Flipped-Label In-Context Learning [8]

Randomly flip X% of binary labels

- More flips (X↑), more requirement of existing knowledge to make correct prediction

Behavior of models with bigger X%

- Those care less use more inner knowledge
- Those impacted more learn more in-context

[8] Wei, et al. "Larger language models do in-context learning differently." arXiv 2023.

In-Context Learning Interpretation: Random Label Test

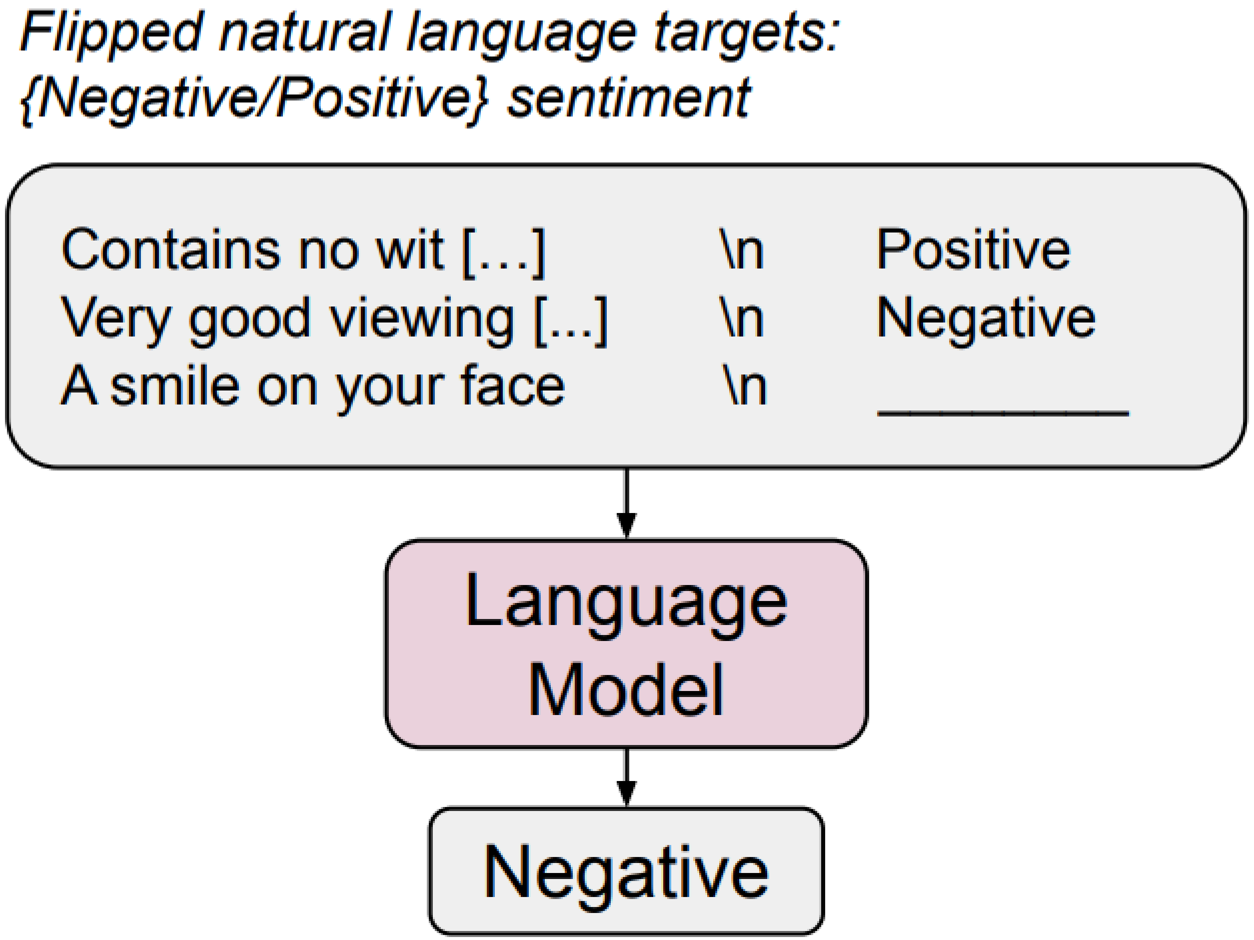

Figure 18: Flipped-Label In-Context Learning [8]

Randomly flip X% of binary labels

- More flips (X↑), more requirement of existing knowledge to make correct prediction

Behavior of models with bigger X%

- Those care less use more inner knowledge
- Those impacted more learn more in-context

Question:

- Does larger LM care more, or less about bigger X?

In-Context Learning Interpretation: Random Label Test

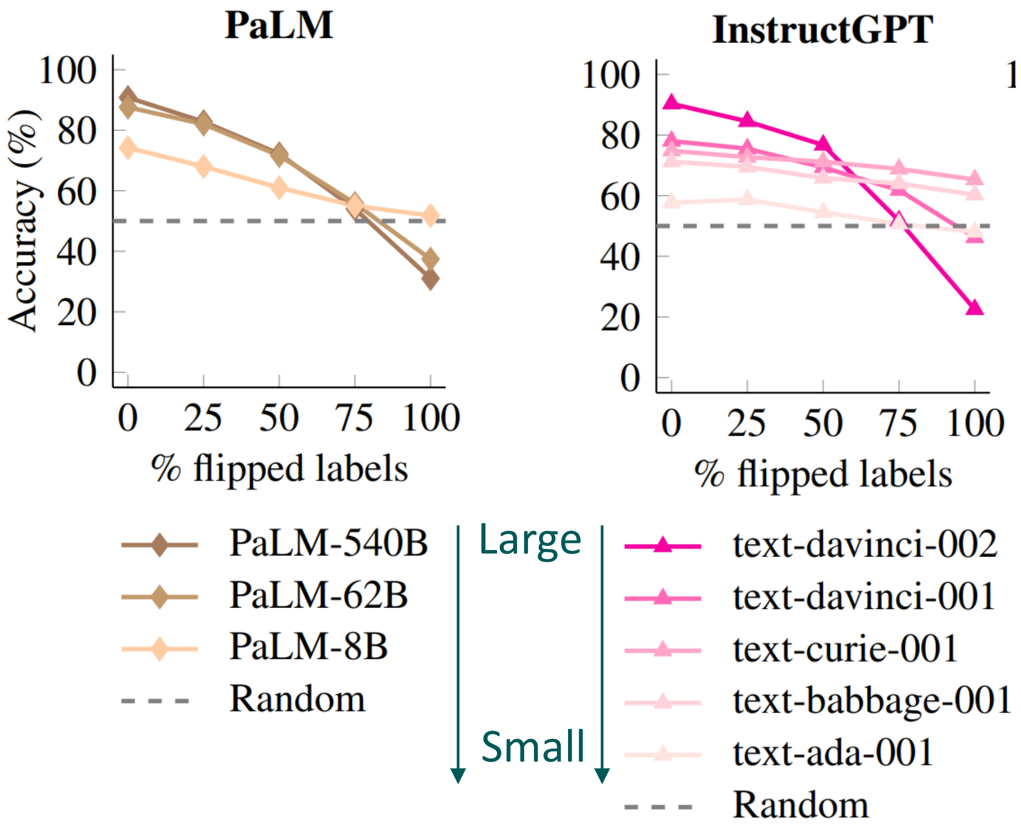

Figure 19: PaLM and GPT in Flipped-Label In-Context Learning, binary classification with 16 examples per class [8]

Larger models perform better with 0% flipped label

- But are much more sensitive to label flips

In-Context Learning Interpretation: Random Label Test

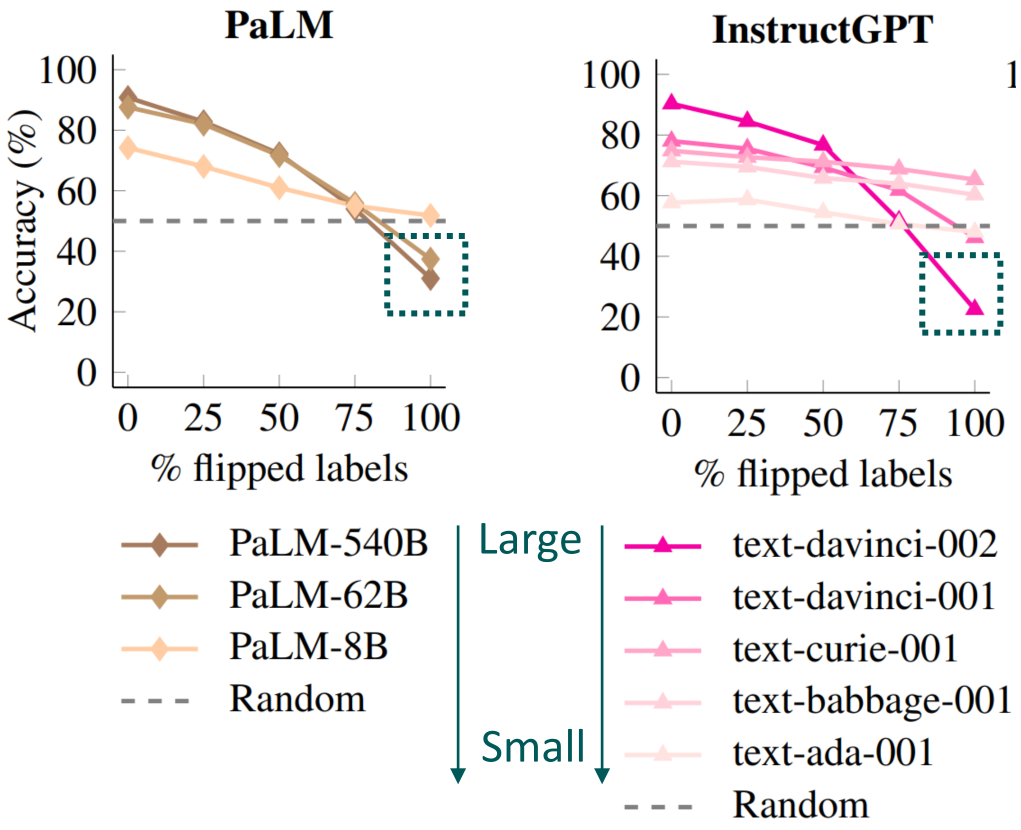

Figure 19: PaLM and GPT in Flipped-Label In-Context Learning, binary classification with 16 examples per class [8]

Larger models perform better with 0% flipped label

- But are much more sensitive to label flips

The strongest models can even over-correct

- With merely 32 in-context labels

There must be some learning in in-context learning

- Especially in larger LMs

In-Context Learning Interpretation: No Semantic Test

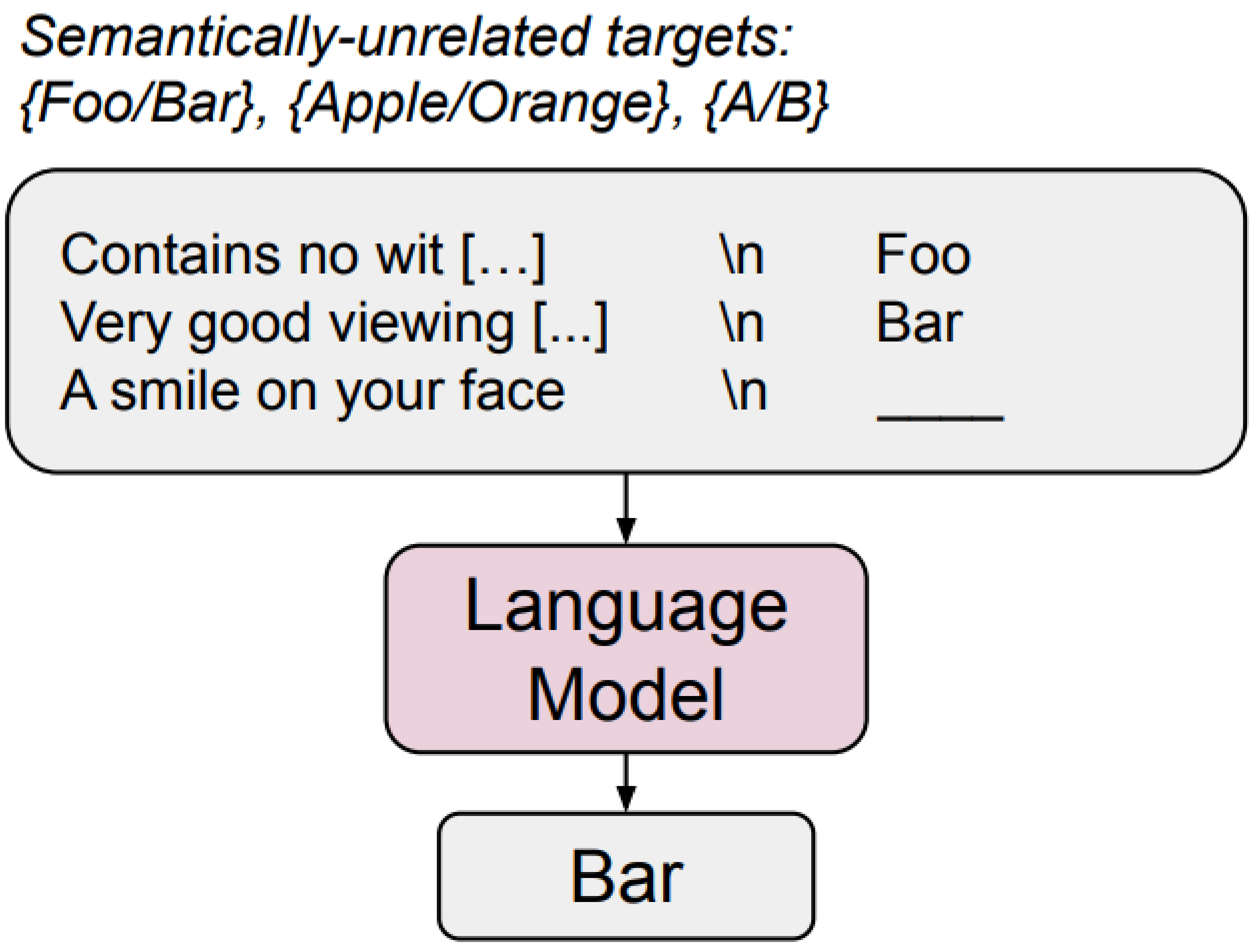

Figure 20: In-Context Learning with Semantically-Unrelated Label Terms [8]

Use semantically-unrelated label terms

- E.g., foo / bar instead of positive / negative
- Models have to learn more from in-context

Behavior of models with unrelated labels

- Those perform well learns more in-context
- Those impacted rely more in existing knowledge

In-Context Learning Interpretation: No Semantic Test

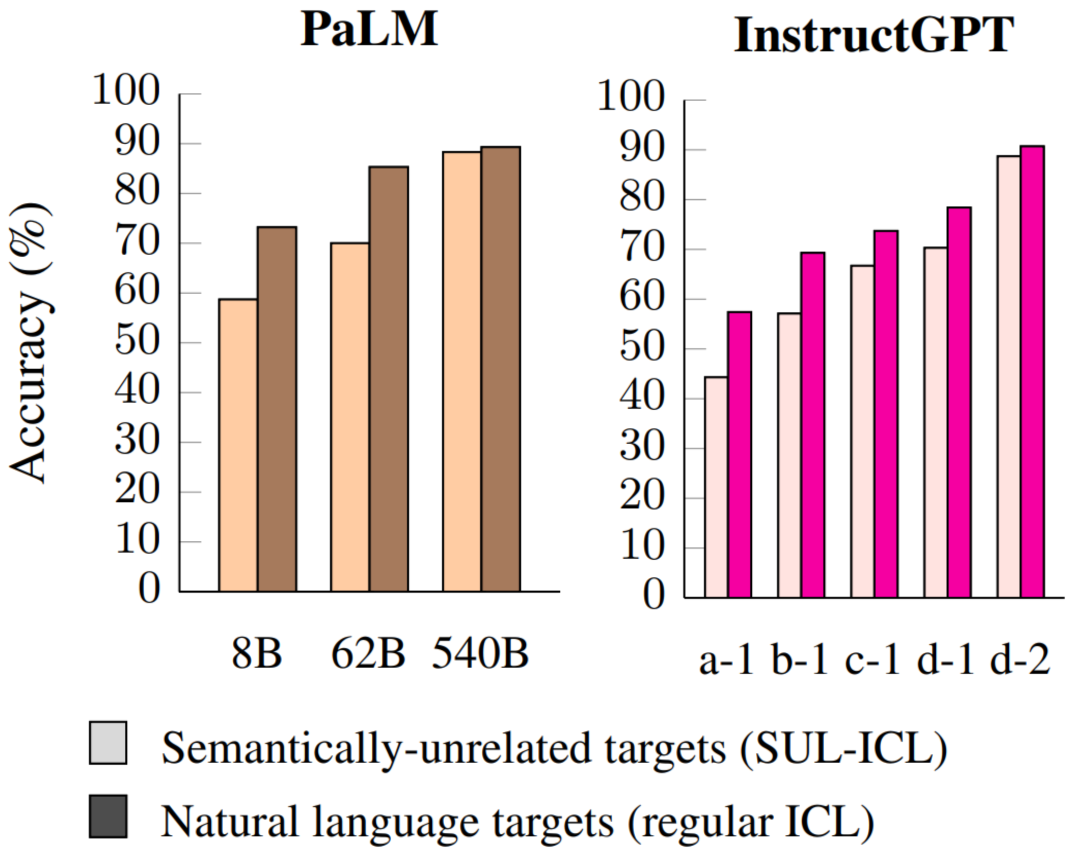

Figure 21: In-Context Learning Accuracy with Semantically-Unrelated Labels versus Related Labels [8]

Larger models work better with unrelated labels

- They learn in-context label mappings better

Smaller models are more prune to unrelated labels

- They rely more on their prior-knowledge

[8] Wei, et al. "Larger language models do in-context learning differently." arXiv 2023.

In-Context Learning Interpretation: No Semantic Test

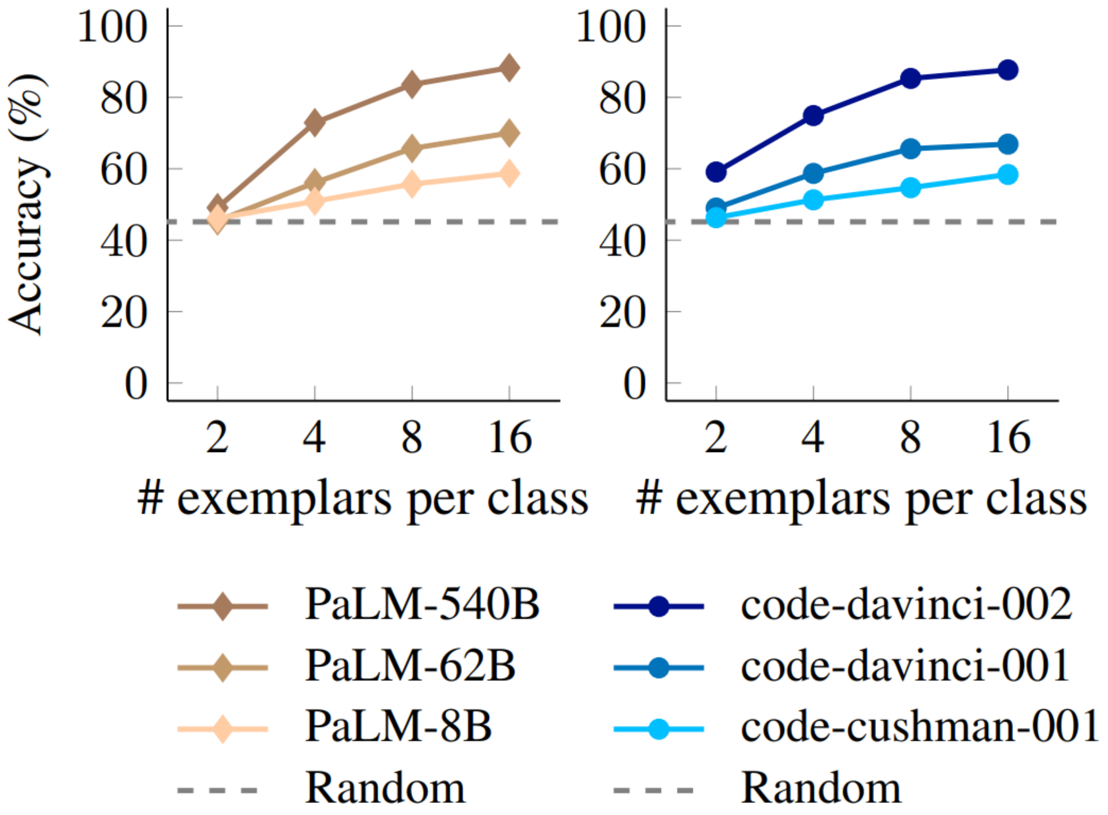

Figure 22: In-Context Learning with Different Number of Semantically-Unrelated Labels [8]

Larger models better leverages in-context examples

- Advantages more pronounces with more labels

Not much better than random with two examples

- Confirms unrelated labels are not aligned with existing semantic knowledge

In-Context Learning Interpretation: Observations

Smaller LMs rely more on existing knowledge and are less effective in learning from in-context

- Less sensitive to flipped labels
- Hard to capture semantically-unrelated input-label mappings
- Random labels unlikely to change output of small LMs

Larger LMs are more effectively in learning from in-context examples

- Can reverse their semantic prior to predict flipped labels
- Can learn semantic-unrelated label mappings
- Better utilizes more in-context examples

In-Context Learning Interpretation: Observations

Smaller LMs rely more on existing knowledge and are less effective in learning from in-context

- Less sensitive to flipped labels
- Hard to capture semantically-unrelated input-label mappings
- Random labels unlikely to change output of small LMs

Larger LMs are more effectively in learning from in-context examples

- Can reverse their semantic prior to predict flipped labels
- Can learn semantic-unrelated label mappings
- Better utilizes more in-context examples

Why? How can LLMs learn from in-context examples?

Outline

What is captured in BERT?

Why pretrained models generalize?

What does in-context learning do?

- Semantic Prior or Input-Label Mapping
- **Connection with Gradient Decent**

Learning in In-Context Learning: Gradient Construction

One can *manually* construct a Transformer (TF_{GD}) that does gradient operation in in-context learning

- Its prediction given in-context learning examples (X_k, Y_k)

  == a reference model after performing SGD on (X_k, Y_k)
- The predict change of adding a new (x, y) is similar with reference model after an SGD step with (x, y)

[9] Oswald, et al. "Transformers Learn In-Context by Gradient Descent." ICML 2023.

Learning in In-Context Learning: Gradient Construction

One can *manually* construct a Transformer (TF_{GD}) that does gradient operation in in-context learning

- Its prediction given in-context learning examples (X_k, Y_k)
  == a reference model after performing SGD on (X_k, Y_k)
- The predict change of adding a new (x, y) is similar with reference model after an SGD step with (x, y)

Currently it can be done in these conditions [9]:

- Linear self-attention, no SoftMax
- Reference model is a simple regression model such as linear regression
- Can stack linear self-attention with MLP but nothing more, i.e. no layer norm etc.

[9] Oswald, et al. “Transformers Learn In-Context by Gradient Descent." ICML 2023.

Learning in In-Context Learning: Gradient Construction

Detailed mathematical construction can be found in Oswald et al. 2023 [9].

Intuitively:

- Self-attention is a high-capacity function and can approximate many math operations
- The reference model (the one who does SGD) is a simple linear regression model
- Lost of non-linearity removed to facilitated the construction

[9] Oswald, et al. "Transformers Learn In-Context by Gradient Descent." ICML 2023.

Learning in In-Context Learning: Gradient Construction

Detailed mathematical construction can be found in Oswald et al. 2023 [9].

Intuitively:

- Self-attention is a high-capacity function and can approximate many math operations
- The reference model (the one who does SGD) is a simple linear regression model
- Lost of non-linearity removed to facilitated the construction

A very toy-ish set up, but a good thought process and a starting point to understand complicated LLMs

- Similar assumptions are often taken in current deep learning theory research

The gradient decent Transformer T_{GD} is learn in-context by gradient decent by construction

[9] Oswald, et al. "Transformers Learn In-Context by Gradient Descent." ICML 2023.

Learning in In-Context Learning: Trained Transformer

TF_{GD} is constructed but not learned

- A constructed measurement target

One can train the toy Transformer TF_{Train} in the same in-context learning set up

- E.g., to perform linear regression task with in-context examples

[9] Oswald, et al. "Transformers Learn In-Context by Gradient Descent." ICML 2023.

Learning in In-Context Learning: Comparison

TF_{GD} is constructed but not learned

- A constructed measurement target

One can train the toy Transformer TF_{Train} in the same in-context learning set up

- E.g., to perform linear regression task with in-context examples

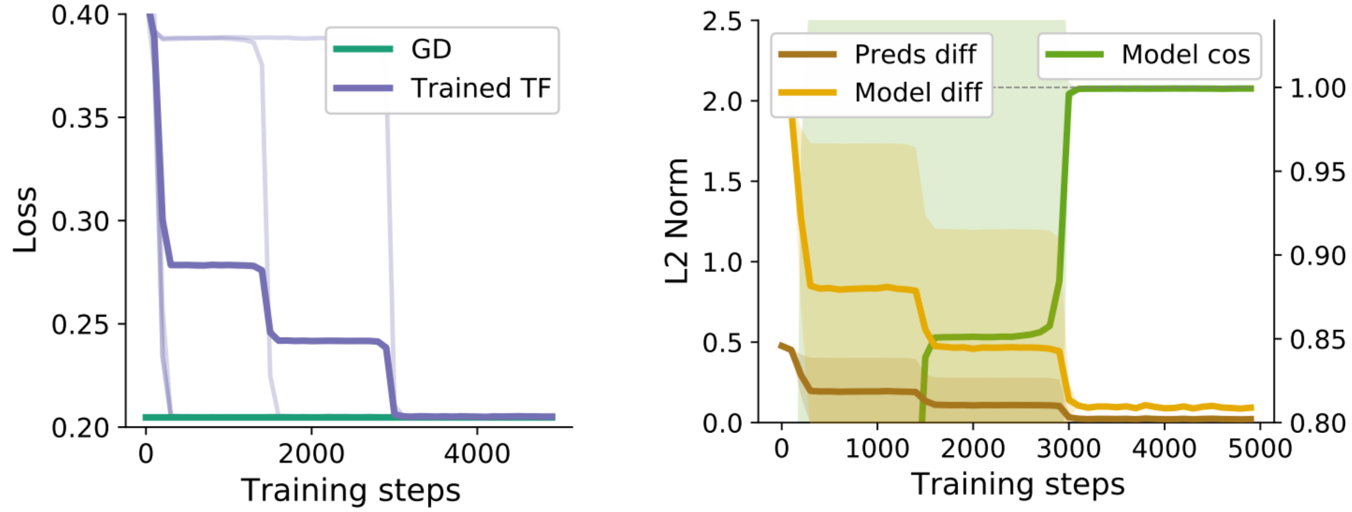

Figure 23: Comparison of constructed TF_{GD} and Trained TF_{Train}. [9]

Trained Transformer matches the constructed gradient decent Transformer

- Near identical
  - Prediction L2 difference
  - Model sensitivity cosine/L2 difference
  - Model sensitivity L2 difference

Learning in In-Context Learning: Comparison

TF_{GD} is constructed but not learned

- A constructed measurement target

One can train the toy Transformer TF_{Train} in the same in-context learning set up

- E.g., to perform linear regression task with in-context examples

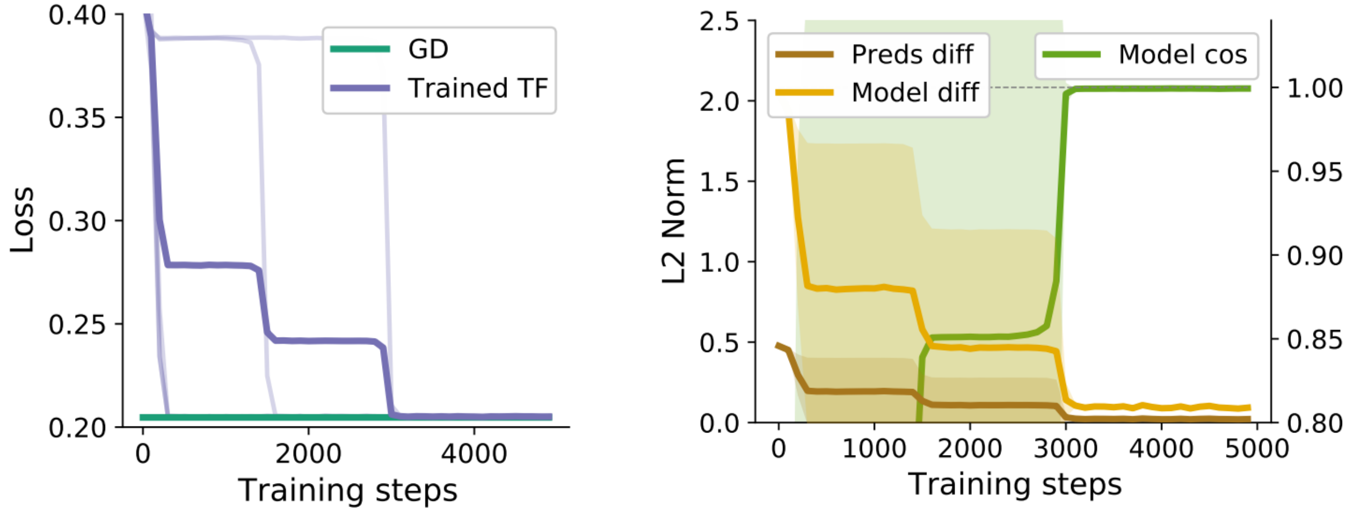

Figure 23: Comparison of constructed TF_{GD} and Trained TF_{Train}. [9]

Trained Transformer matches the constructed gradient decent Transformer

- Near identical
  - Prediction L2 difference
  - Model sensitivity cosine/L2 difference
  - Model sensitivity L2 difference

Transformers (<u>with strong assumptions and simplifications</u>) learn in-context by gradient descent (<u>of a linear regression model</u>)

Learning in In-Context Learning: Multi-Layer Transformer

Compare the constructed and learned Transformer in multi-layer setting

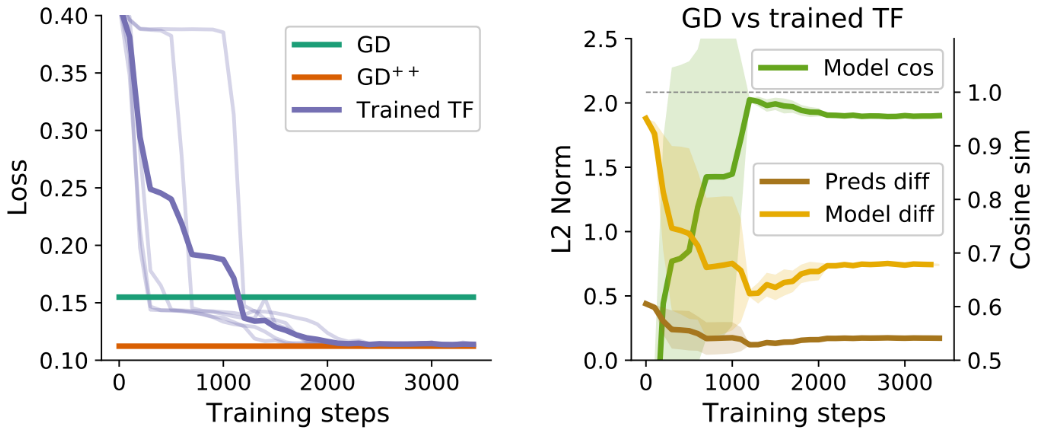

Figure 24: Two-layer TF_{GD} versus TF_{Train}. [9]

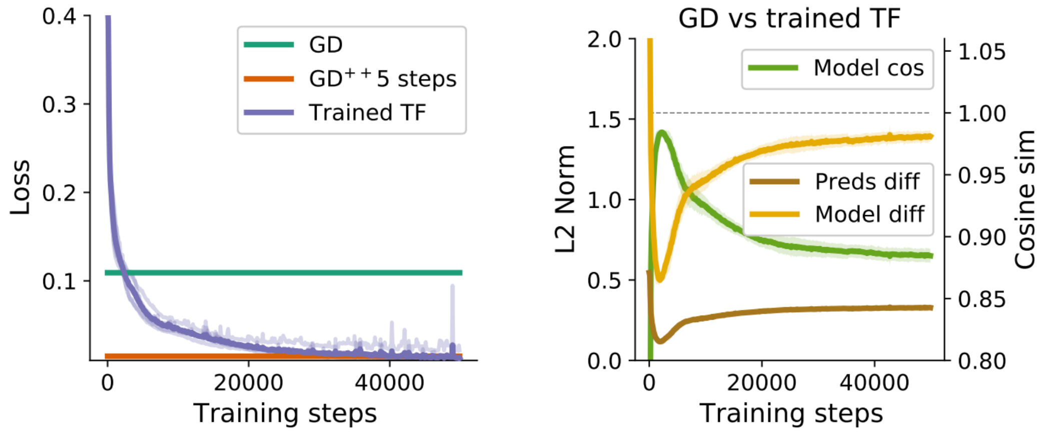

Figure 25: Five-layer TF_{GD} versus TF_{Train}. [9]

Learning in In-Context Learning: Multi-Layer Transformer

Compare the constructed and learned Transformer in multi-layer setting

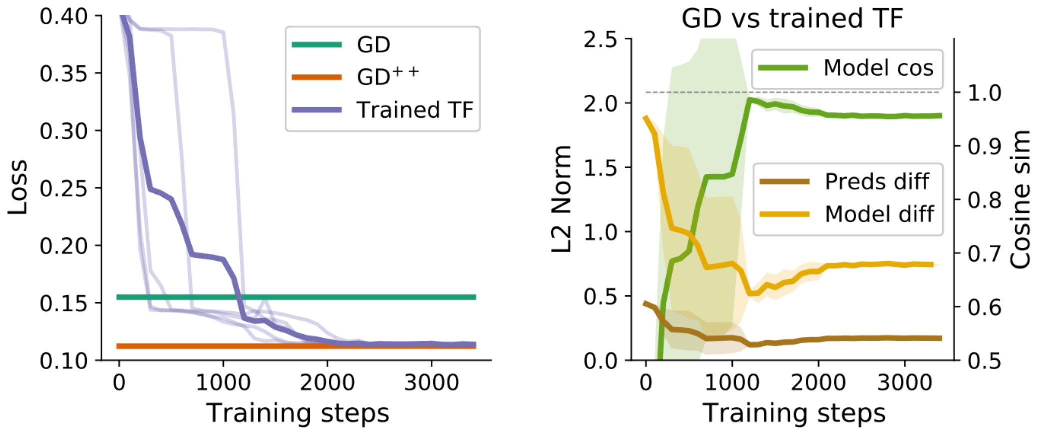

Figure 24: Two-layer TF_{GD} versus TF_{Train}. [9]

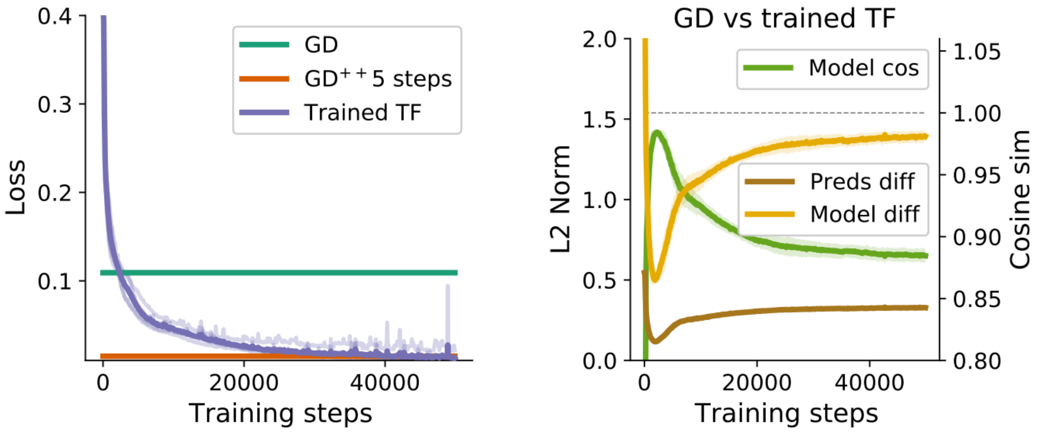

Figure 25: Five-layer TF_{GD} versus TF_{Train}. [9]

- Learned Transformer outperforms the constructed TF_{GD}
- Upgraded gradient decent TF_{GD} with manually tuned data transformation matches better
- Divergence increases with deeper (five only, still) networks
- But still remarkable similarity of in-context learning and gradient decent

Learning in In-Context Learning: Theory versus Empirical

Empirical Observation

- Larger Transformers better learn in-context
- More in-context examples help larger model more
- Smaller Transformers rely more on existing semantic

Theory

- Transformers perform one gradient step per layer
- And per in-context example
- Smaller models have limited gradient steps built in

Assumptions :

- Linear attention + MLP Transformer
- Simple regression reference model
- Shallow networks

In-Context Learning Interpretation: Summary

Various solid empirical evidence that:

- Larger Transformers do learn in-context
- In-context learning ability correlates with model scale

Theorical connections are build between in-context learning and gradient decent observations

- Good intuitions
- One way to make sense of in-context learning

In-Context Learning Interpretation: Discussion

Likely many not-yet-finished learning theory,

- This interpretation is more for our understanding and inspiration
- Strong assumptions are introduced to make the theory

Personal views:

- In-context learning is different from SGD and is more powerful in some scenarios
- Connecting with existing, well-known techniques is a good starting point
- Eventually researchers will develop new theorical frameworks to explain the amazing capabilities of LLM

Outline

What is captured in BERT?

- Attention patterns
- Probing capture capabilities in representations

Why pretrained models generalize?

- Loss landscapes
- Implicit bias of language models

What does in-context learning do?

- Semantic Prior or Input-Label Mapping
- Connection with Gradient Decent

Quiz: Why the order of in-context example matters?

References: BERTology

- Clark, Kevin, et al. "What does bert look at? an analysis of bert's attention." arXiv preprint arXiv:1906.04341 (2019).
- Tenney, Ian, Dipanjan Das, and Ellie Pavlick. "BERT rediscovers the classical NLP pipeline." arXiv preprint arXiv:1905.05950 (2019).
- Htut, Phu Mon, et al. "Do attention heads in BERT track syntactic dependencies?." arXiv preprint arXiv:1911.12246 (2019).
- Liu, Leo Z., et al. "Probing across time: What does RoBERTa know and when?." arXiv preprint arXiv:2104.07885 (2021).
- Tenney, Ian, et al. "What do you learn from context? probing for sentence structure in contextualized word representations." arXiv preprint arXiv:1905.06316 (2019).
- Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. "A primer in BERTology: What we know about how BERT works." Transactions of the Association for Computational Linguistics 8 (2021): 842-866.
- Carlini, Nicholas, et al. "Extracting Training Data from Large Language Models." USENIX Security Symposium. Vol. 6. 2021.
- Carlini, Nicholas, et al. "Quantifying memorization across neural language models." arXiv preprint arXiv:2202.07646 (2022).
- Izacard, Gautier, and Edouard Grave. "Distilling knowledge from reader to retriever for question answering." arXiv preprint arXiv:2012.04584 (2020).

References: Optimization

- Erhan, Dumitru, et al. "The difficulty of training deep architectures and the effect of unsupervised pre-training." Artificial Intelligence and Statistics. PMLR, 2009.
- Li, Hao, et al. "Visualizing the loss landscape of neural nets." Advances in neural information processing systems 31 (2018).
- Hao, Yaru, et al. "Visualizing and understanding the effectiveness of BERT." arXiv preprint arXiv:1908.05620 (2019).
- Liu, Hong, et al. "Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models." arXiv preprint arXiv:2210.14199 (2022).
- Chiang, Ping-yeh, et al. "Loss Landscapes are All You Need: Neural Network Generalization Can Be Explained Without the Implicit Bias of Gradient Descent." The Eleventh International Conference on Learning Representations. 2023.

References: Knowledge

- Petroni, Fabio, et al. "Language models as knowledge bases?." arXiv preprint arXiv:1909.01066 (2019).
- Roberts, Adam, Colin Raffel, and Noam Shazeer. "How much knowledge can you pack into the parameters of a language model?." arXiv preprint arXiv:2002.08910 (2020).
- Jiang, Zhengbao, et al. "How can we know what language models know?." Transactions of the Association for Computational Linguistics 8 (2020): 423-438.
- Zaken, Elad Ben, Shauli Ravfogel, and Yoav Goldberg. "Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models." arXiv preprint arXiv:2106.10199 (2021).
- Min, Sewon, et al. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?." arXiv preprint arXiv:2202.12837 (2022).
- Geva, Mor, et al. "Transformer feed-forward layers are key-value memories." arXiv preprint arXiv:2012.14913 (2020).
- Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

BERT Attention Patterns: Linguistic Examples

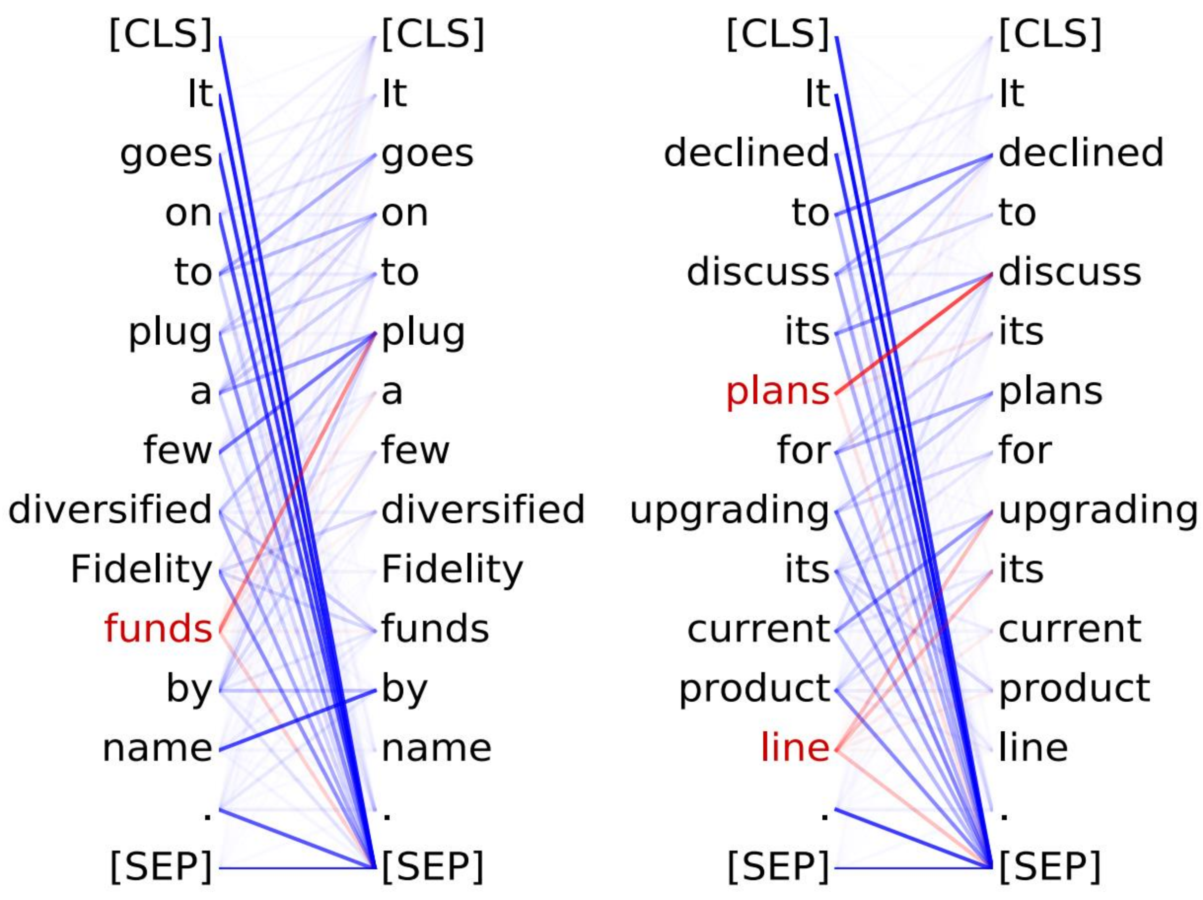

Figure 5: Objects Attend to their Verbs (Left→Right) [1]

[1] Clark Et al. “What Does BERT Look At? An Analysis of BERT’s Attention.” BlackBoxNLP 2019

BERT Attention Patterns: Linguistic Examples

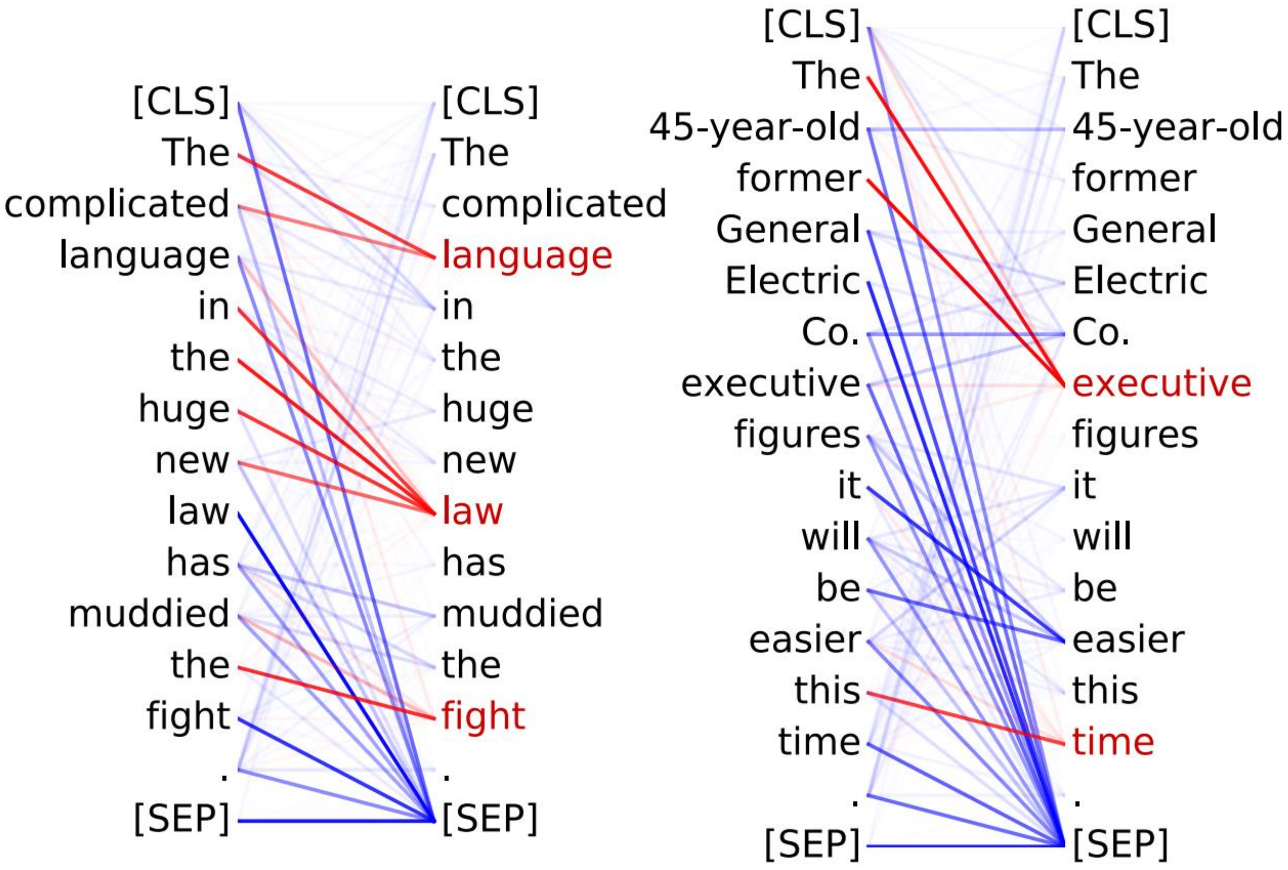

Figure 6: Noun Modifiers Attend to their Noun (Left→Right) [1]

Probing Pretraining Representations: Across Layers

Mixing representations from multiple layers:

$$\boldsymbol{h}_t^{\text{mix}} = \sum_l s^l \boldsymbol{h}_t^l \,; s^l = \text{softmax}(\alpha^l)$$

Definition: Center-of-Gravity

$$E[l] = \sum_l l \cdot s^l$$

- Expected layer to convey the information needed by the probe task
- Larger Center-of-Gravity → information needed captured at higher layers

Definition: Expected Layer

$$\Delta^l = \text{Probing Score}(0\!:l) - \text{Probing Score}(0\!:l-1)$$

$$E[\Delta^l] = \frac{\sum_l l \cdot \Delta^l}{\sum_l \Delta^l}$$

- Δ^l : The benefit of adding layer l in the mix
- $E[\Delta^l]$: The expected layer to resolve the probing task

[3] Tenney, Ian, Dipanjan Das, and Ellie Pavlick. "BERT Rediscovers the Classical NLP Pipeline." ACL. 2019.

Probing Across Time Tasks

Package	Knowledge	Task	Formulation	Examples
LKT	Linguistic	POS Tagging	Token Labeling	PRON AUX **VERB** ADV ADP DET NOUN PUNCT I 'm **staying** away from the stock .
		Syntactic Chunking		B-NP B-VP B-PP **B-NP** I-NP I-NP O Shearson works at **American** Express Co .
		Name Entity Recognition		O O I-ORG **I-ORG** I-ORG O O O O By stumps Kent **County** Club had reached 108 .
		Syntactic Arc Predication	Token Pair Labeling	Peter and May bought a car . (✗ ✗ ✓)
		Syntactic Arc Classification		Peter and May bought a car . (Conj, Conj, Subj, Obj, Det)
BLiMP	Linguistic	Irregular Forms	Comparing Sentence Scores **Expected:** $\mathbb{S}(\checkmark) > \mathbb{S}(\text{✗})$	✓ Aaron ***broke*** the unicycle. ✗ Aaron ***broken*** the unicycle.
		Determiner-Noun Agree.		✓ Rachelle had bought that ***chair*** ✗ Rachelle had bought that ***chairs***.
		Subject-Verb Agreement		✓ These casseroles ***disgust*** Kayla. ✗ These casseroles ***disgusts*** Kayla.
		Island Effect		✓ Which ***bikes*** is John fixing? ✗ Which is John fixing ***bikes***?
		Filler Gap		✓ Brett knew ***what*** many waiters find. ✗ Brett knew ***that*** many waiters find.
LAMA	Factual	Google RE	Masked LM **Expected:** $\forall w \in V_{\text{RoBERTa}} \setminus \{\checkmark\}$, $\mathbb{P}(\checkmark \mid \mathcal{C}) > \mathbb{P}(w \mid \mathcal{C})$	Albert Einstein was born in ***[MASK]*** ✓: ***[MASK]*** = 1879
		T-REx		Humphrey Cobb was a ***[MASK]*** and novelist ✓: ***[MASK]*** = screenwriter
		SQuAD		A Turing machine handles ***[MASK]*** on a strip of tape. ✓: ***[MASK]*** = symbols
	Commonsense	ConceptNet		You can use ***[MASK]*** to bathe your dog. ✓: ***[MASK]*** = shampoo
CAT	Commonsense	Conjunction Acceptability	Comparing Sentence Scores **Expected:** $\forall$✗, $\mathbb{S}(\checkmark) > \mathbb{S}(\text{✗})$	✓ Jim yelled at Kevin ***because*** Jim was so upset. ✗ Jim yelled at Kevin ***and*** Jim was so upset.
		Winograd		✓ The fish ate the worm. The ***fish*** was hungry. ✗ The fish ate the worm. The ***worm*** was hungry.
		Sense Making		✓ Money can be used for buying ***cars***. ✗ Money can be used for buying ***stars***.
		SWAG		✓ Someone unlocks the door and they go in. ***Someone leads the way in***. ✗ Someone unlocks the door and they go in. ***Someone opens the door and walks out***. ✗ Someone unlocks the door and they go in. ***Someone walks out of the driveway***. ✗ Someone unlocks the door and they go in. ***Someone walks next to someone and sits on a pew***.
		Argument Reasoning		✓ People can choose not to use Google, ***and since all other search engines re-direct to Google***, Google is not a harmful monopoly. ✗ People can choose not to use Google, ***but since other search engines do not re-direct to Google***, Google is not a harmful monopoly.
oLMpics	Reasoning	Taxonomy Conjunction	Multiple Choice Masked LM **Expected:** $\forall$✗, $\mathbb{P}(\checkmark \mid \mathcal{C}) > \mathbb{P}(\text{✗} \mid \mathcal{C})$	A ferry and a floatplane are both a type of ***[MASK]***. ✓ vehicle ✗ airplane ✗ boat
		Antonym Negation		It was ***[MASK]*** hot, it was really cold. ✓ not ✗ really
		Object Comparison		The size of a airplane is usually much ***[MASK]*** than the size of a house. ✗ smaller ✓ larger
		Always Never		A chicken ***[MASK]*** has horns. ✓ never ✗ rarely ✗ sometimes ✗ often ✗ always
		Multi-Hop Composition		When comparing a 23, a 38 and a 31 year old, the ***[MASK]*** is oldest. ✓ second ✗ first ✗ third

In-Context Learning Interpretation: Summary

Various solid empirical evidence that:

- Larger Transformers do learn in-context
- In-context learning ability correlates with model scale

Theorical connections are build between in-context learning and gradient decent observations

- Good intuitions
- One way to make sense of in-context learning
- Very strong assumptions are introduced for the connection, unfortunately