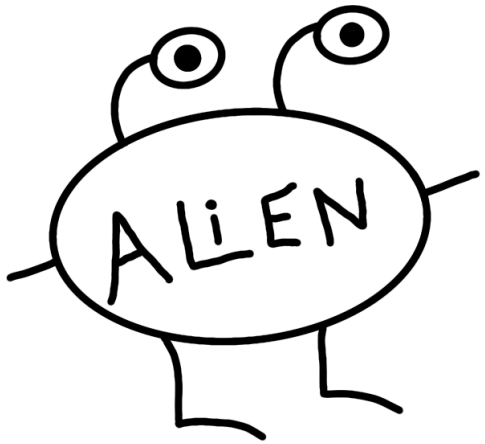
Exploring large language models through the lens of intrinsic dimensionality



Marco Baroni

LangTech @BSC Seminar

11/4/2025

upf. Universitat Pompeu Fabra Barcelona

Computational Linguistics and Linguistic Theory

ICREA

Greetings from El Poblenou!

upf. Universitat Pompeu Fabra Barcelona

Computational Linguistics and Linguistic Theory

Outline

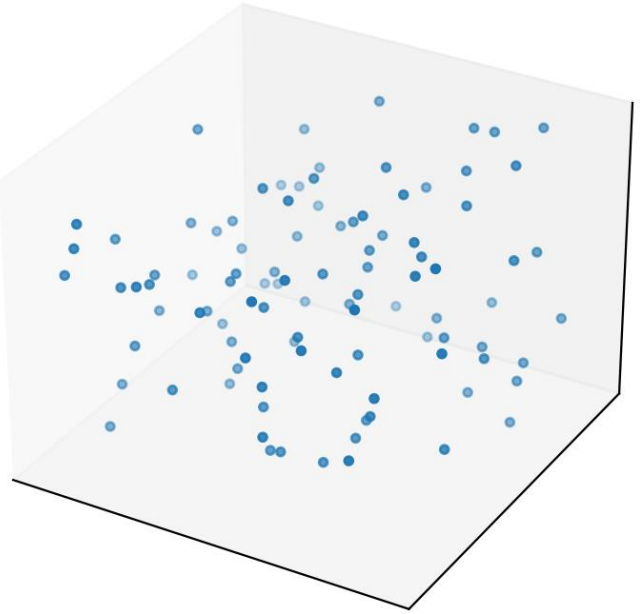
- Motivation and background
- Intrinsic dimensionality as a footprint of linguistic processing
- A more granular look at intrinsic dimensionality and syntactic complexity
- Conclusion

Large language models are wonderful but…

- … we have very little idea of how they work!
- This is an important matter because
  - Safety
  - Inefficient “agnostic” training
  - Understanding LMs might help us understand how information processing systems work in general
- As a consequence, a lot of interpretability research is going on today
  - From very coarse black-box testing (e.g., benchmarking)
  - … to very granular mechanistic analyses
- We seek a middle ground by adopting a geometric view of LM representation spaces

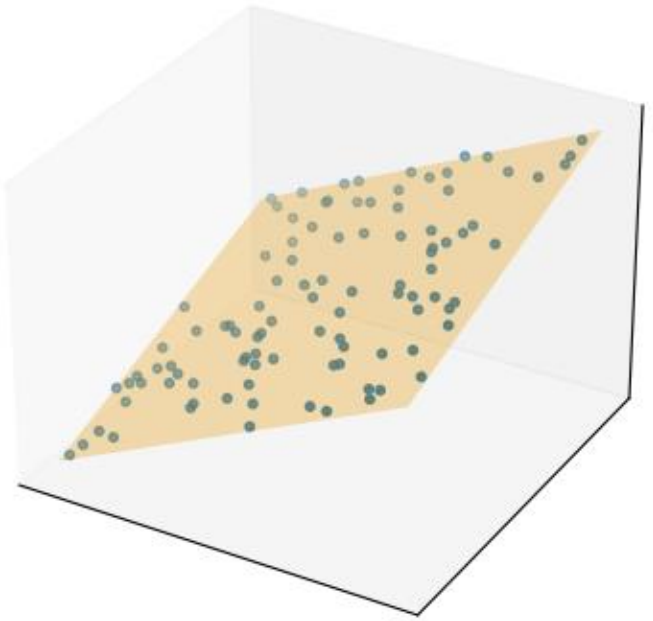
A geometric view of LM representations

- Points are vectors representing a data-set of input texts on a layer of a model
- Dimensionality of space defined by data-set is 3 (or something like 4096 in a more realistic example)



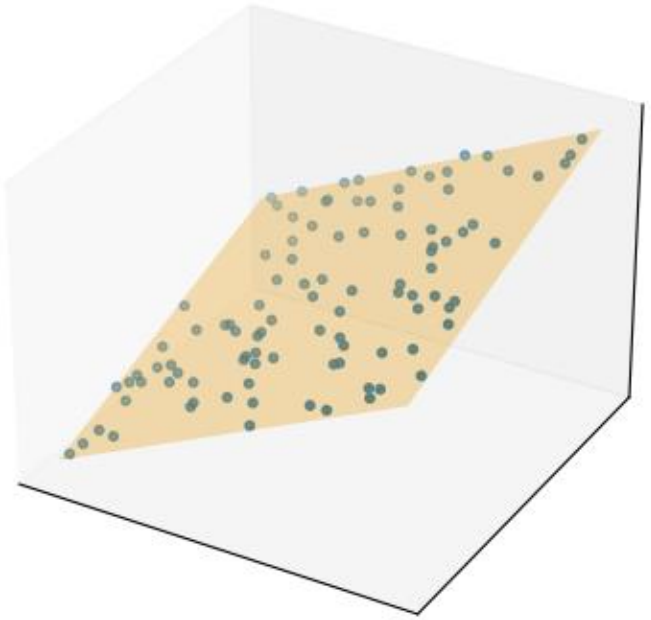
The manifold hypothesis and intrinsic dimensionality

- Points are vectors representing a data-set of input texts on a layer of a model
- ***Extrinsic dimensionality*** of space defined by data-set is 3 (or something like 4096 in a more realistic example)
- ***Intrinsic dimensionality*** is 2

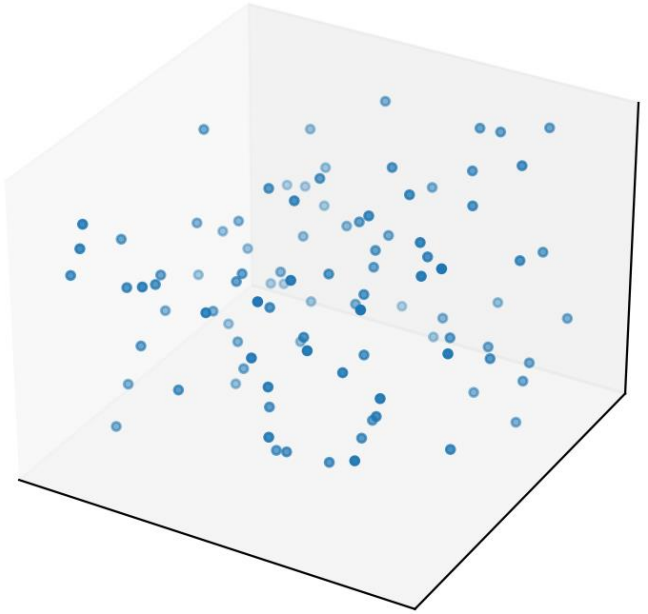


NB: intrinsic dimensionality is a property of a data-set, not a single point

Evolving intrinsic dimensionality across layers of a deep network

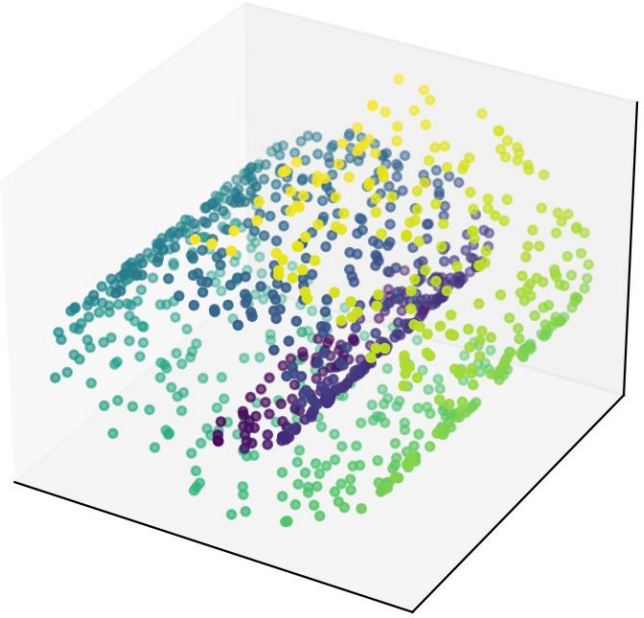


layer N



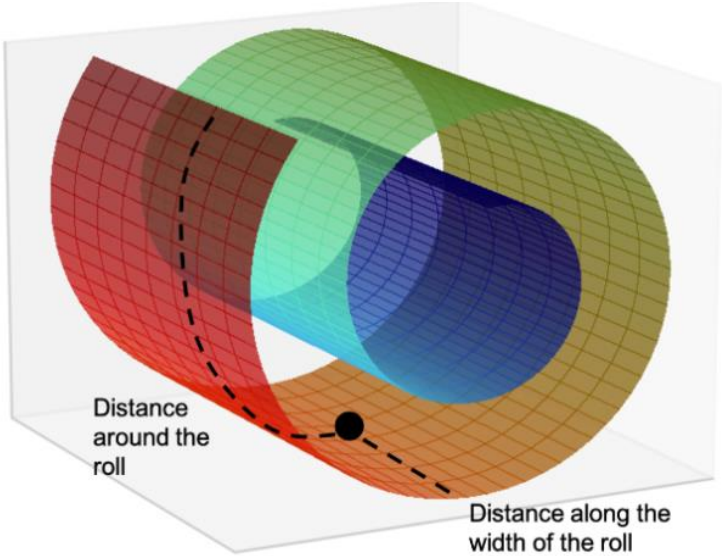
layer N+1

The manifold need not be linear



Estimating the intrinsic dimensionality of a manifold without linearity assumptions

- GRIDE estimator of Denti et al. 2022
- Based on relation between intrinsic dimensionality and expected distribution of distances between points in space
- Intuition: intrinsic dimensionality (ID) estimates the number of independent parameters needed to locally characterize a manifold



https://mbernste.github.io/posts/intrinsic_dimensionality/

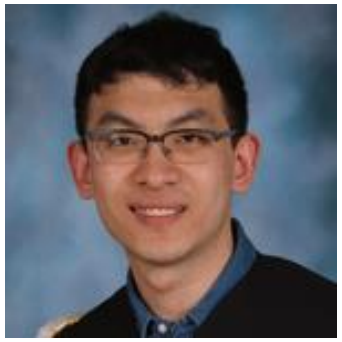
Outline

- Motivation and background
- **Intrinsic dimensionality as a footprint of linguistic processing**
- A more granular look at intrinsic dimensionality and syntactic complexity
- Conclusion

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio and Marco Baroni: *Emergence of a High-Dimensional Abstraction Phase in Language Transformers*, to be presented at ICLR 2025








https://arxiv.org/abs/2405.15471

Methodology

- **LMs:** Pyhia-6.9B, OPT-6.7B, Llama-3-8B, OLMo-7B, Mistral-7B
- **Datasets:** 10k fixed-length token fragments from the Pile, Bookcorpus, WikiText-103
  - also, shuffled-corpus versions to compare them against

Methodology

- **LMs:** Pyhia-6.9B, OPT-6.7B, Llama-3-8B, OLMo-7B, Mistral-7B
- **Datasets:** 10k fixed-length token fragments from the Pile, Bookcorpus, WikiText-103
  - also, shuffled-corpus versions to compare them against

sane input fragment

the "evil agents of the Roman Church" . By contrast , Shakespeare 's King John , a relatively anti

Methodology

- **LMs:** Pyhia-6.9B, OPT-6.7B, Llama-3-8B, OLMo-7B, Mistral-7B
- **Datasets:** 10k fixed-length token fragments from the Pile, Bookcorpus, WikiText-103
  - also, shuffled-corpus versions to compare them against

shuffled input fragment

into on defeated that's ERA, named. sisters "preparation's Ivo", follows Paul team creating

Methodology

- Feed each set of 10k fragments to a model
- Extract hidden vector representing the last token of each fragment at each layer
- Compute ID of space defined by set of 10k last-token vector representations at each layer

Methodology

- Feed each set of 10k fragments to a model
- Extract hidden vector representing the last token of each fragment at each layer
- Compute ID of space defined by set of 10k last-token vector representations at each layer

In autoregressive/causal models, last token representations are the only ones having access to the whole sequence through attention

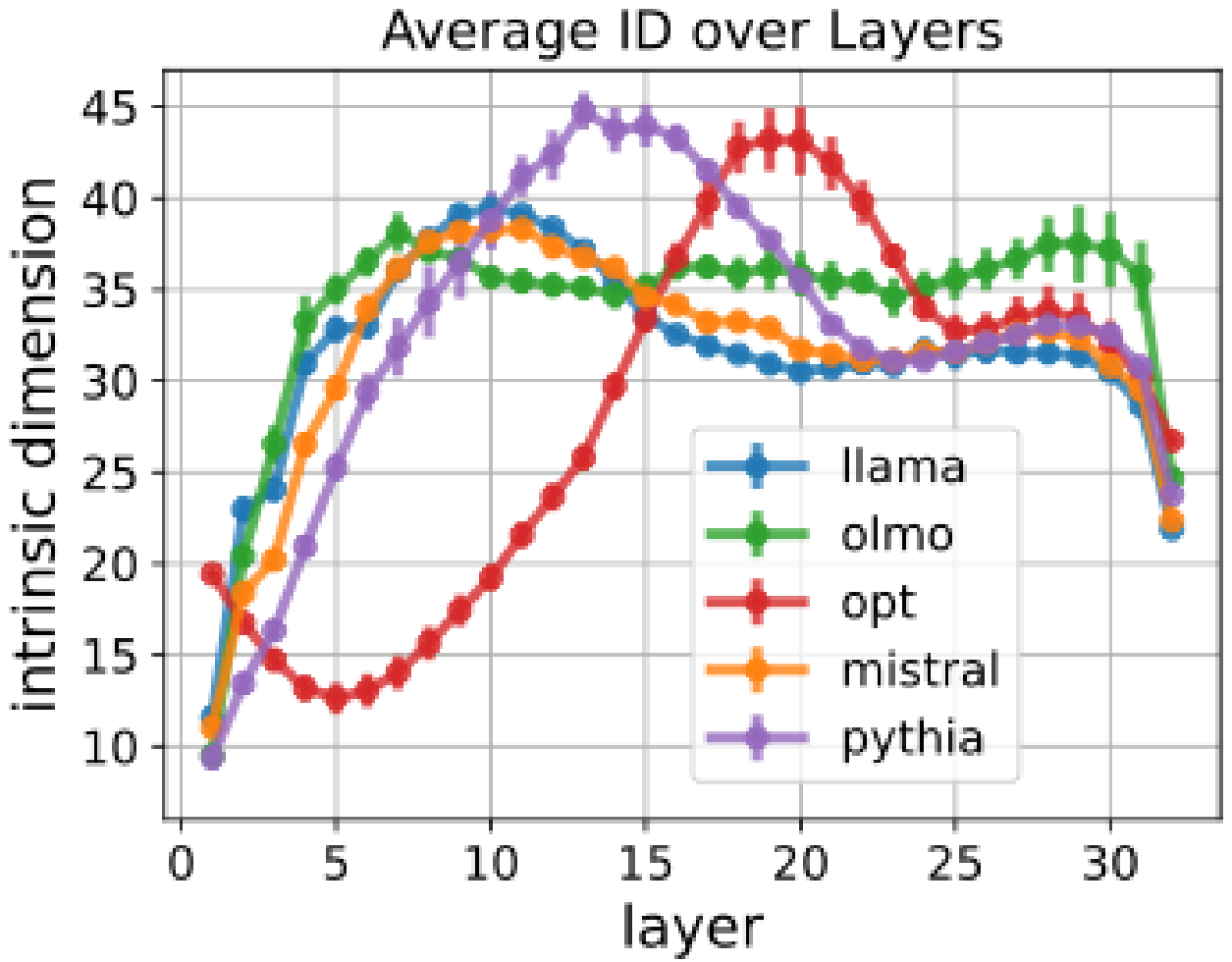
Methodology

- Feed each set of 10k fragments to a model
- Extract hidden vector representing the last token of each fragment at each layer
- Compute ID of space defined by set of 10k last-token vector representations at each layer

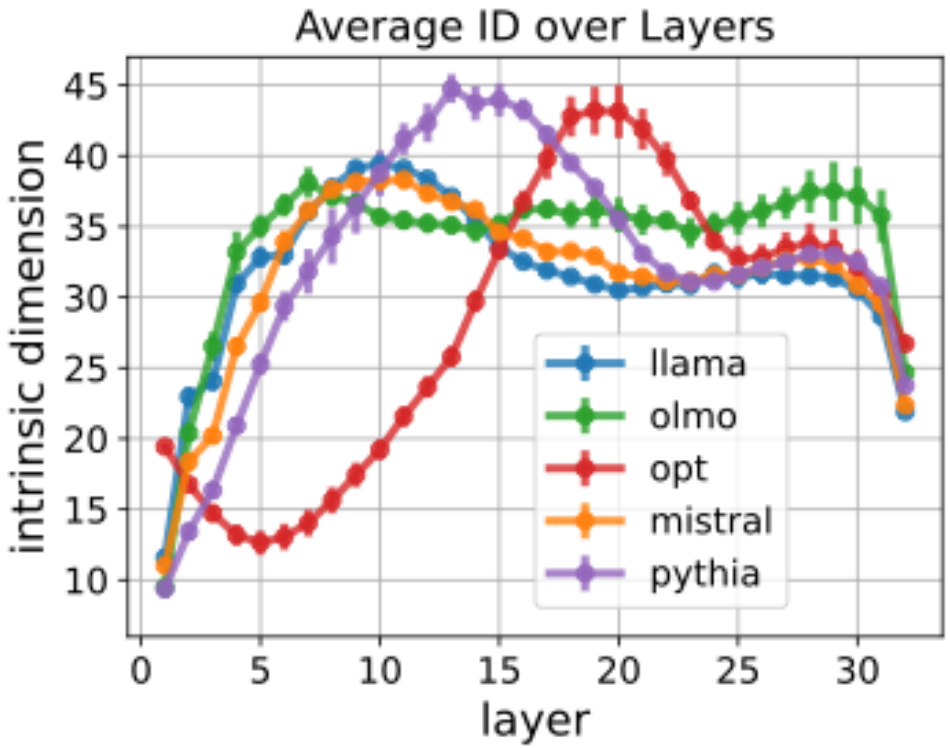
- Process repeated 5 times per corpus with different samples to estimate ID variance
- Here, results averaged across corpora and repetitions, see paper for detailed results

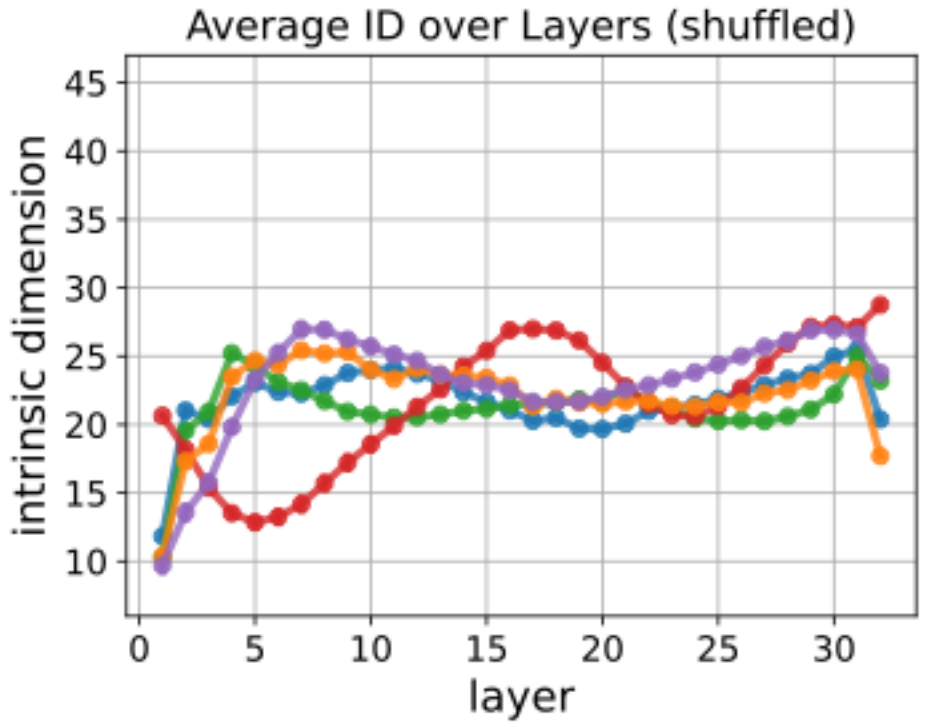
An ID “peak” consistently emerges across LMs and datasets

Extrinsic dimension: 4096

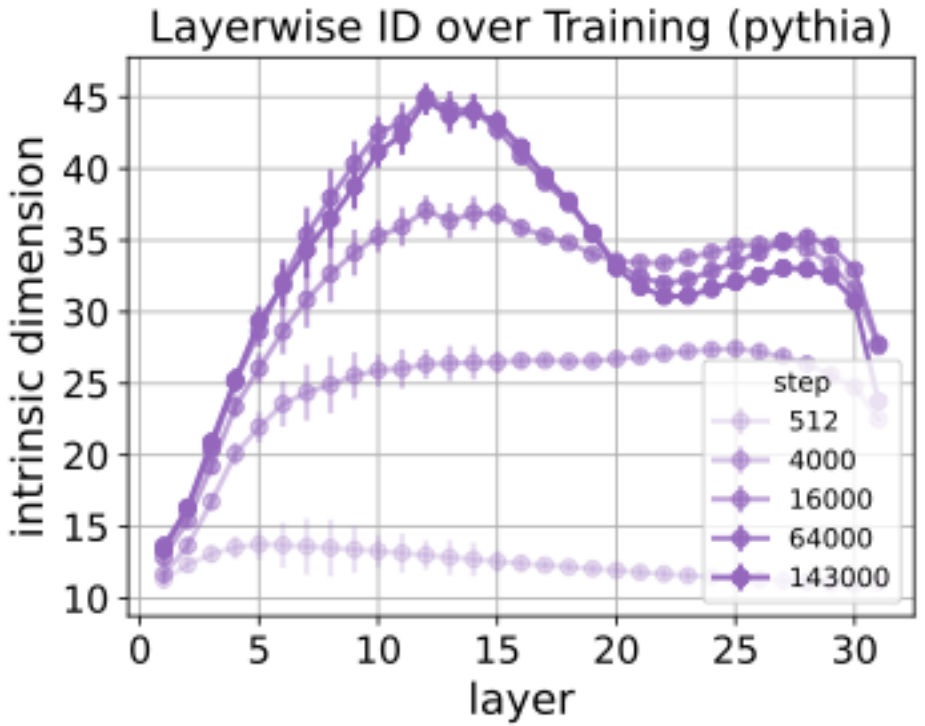


ID peak responds to meaningful text

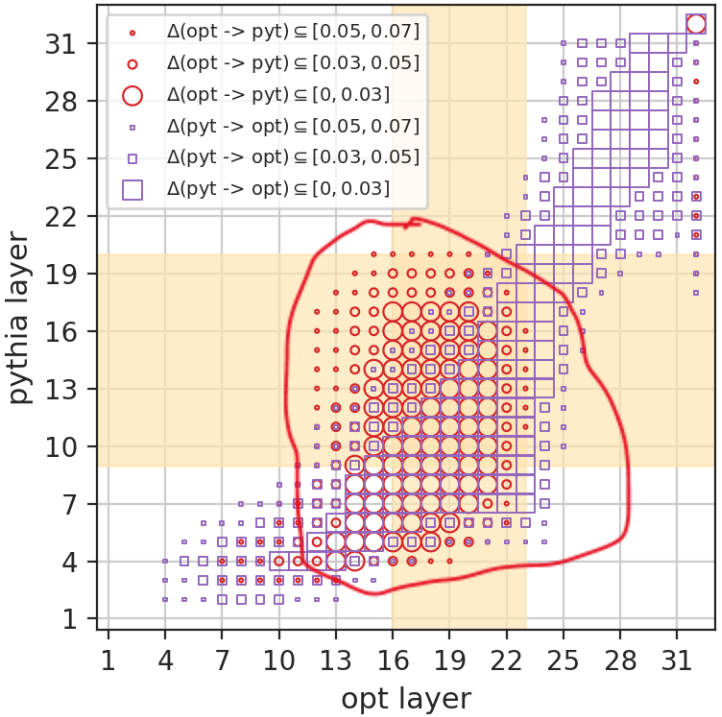


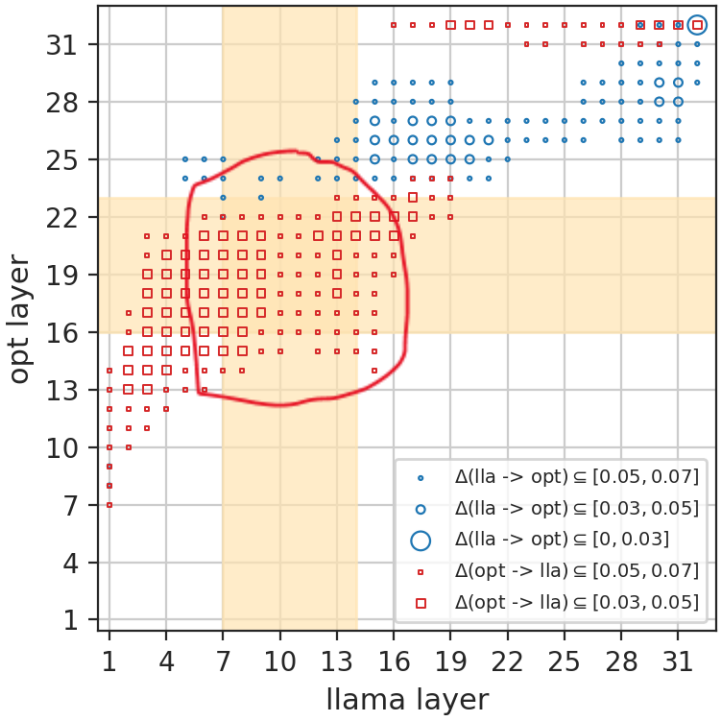


ID peak emerges during training



LM representations resemble each other more under the ID peaks



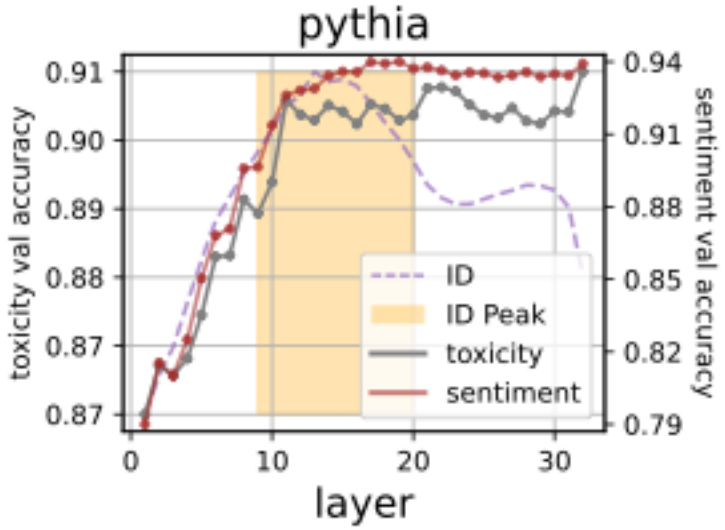


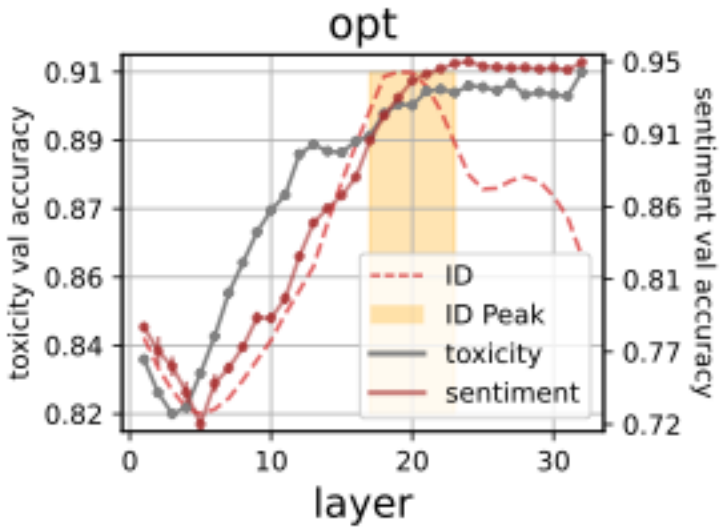
(Asymmetric) similarity measured by Information Imbalance (Glielmo et al. 2022)

What happens under the ID peak?
Downstream tasks

- **Toxicity detection:** predict binary toxicity label from Kaggle's jigsaw toxic comment classification challenge (Adams et al., 2017)
- **Sentiment classification:** Predict positive/negative label in dataset of IMDb movie reviews (Maas et al., 2011)
- Linear classifier trained/tested on frozen last-token representations from each LM layer

Downstream transfer and the ID peak

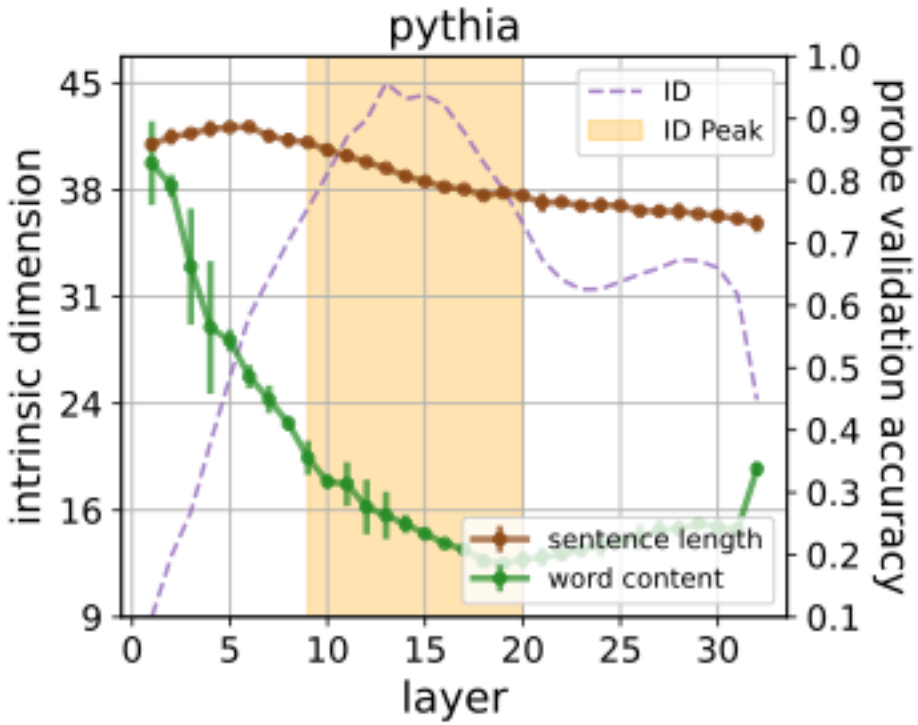


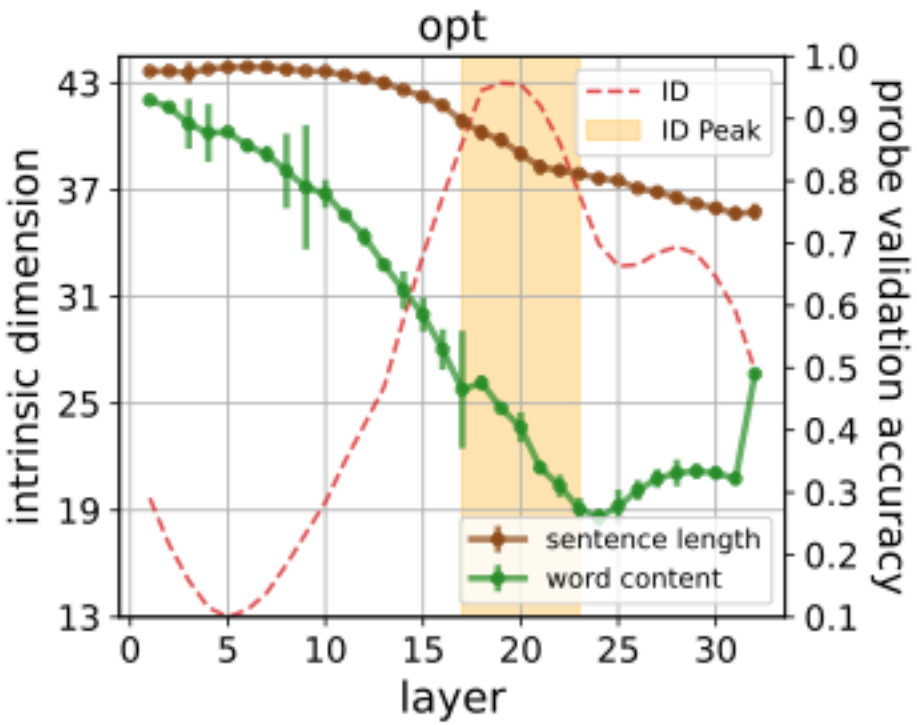


What happens under the ID peak?
Probing tasks from Conneau et al. 2018

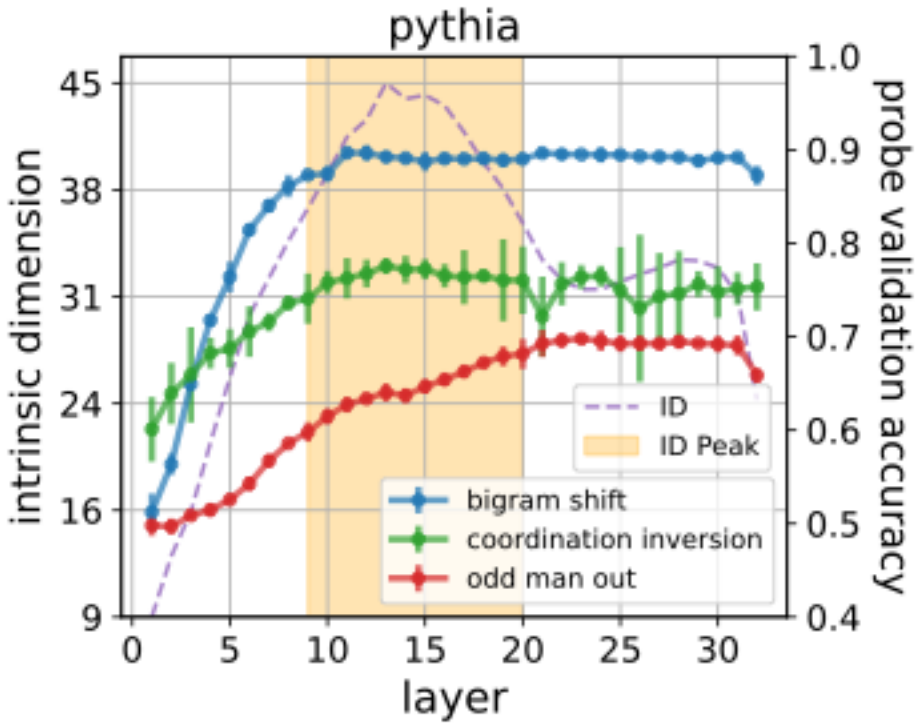
- Surface form tasks
  - **Sentence Length:** predict input sentence length in tokens
  - **Word Content:** tell which of a pre-determined set of words occurs in the input sentence
- Syntactic/semantic tasks
  - **Bigram Shift:** is a sentence well-formed, or corrupted by inverting two adjacent tokens?
  - **Coordination Inversion:** is a sentence well-formed, or corrupted by inverting two coordinated clauses?
  - **Odd Man Out**: is a sentence well-formed, or does it contain a random noun or verb?
- 1-layer MLP-classifier trained/tested on last-token representations from each LM layer

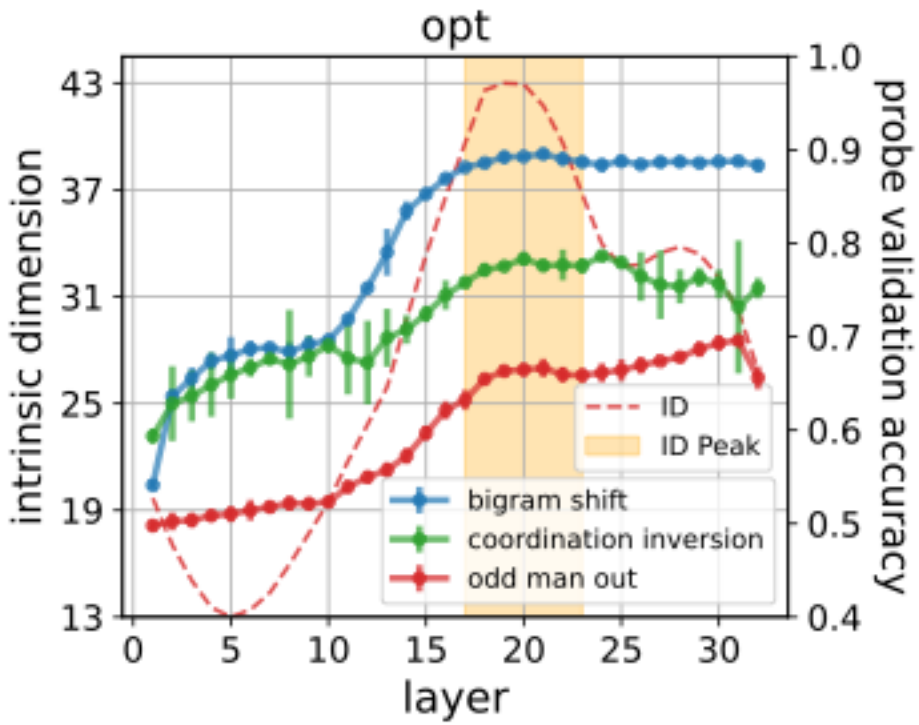
Surface probes and the ID peak



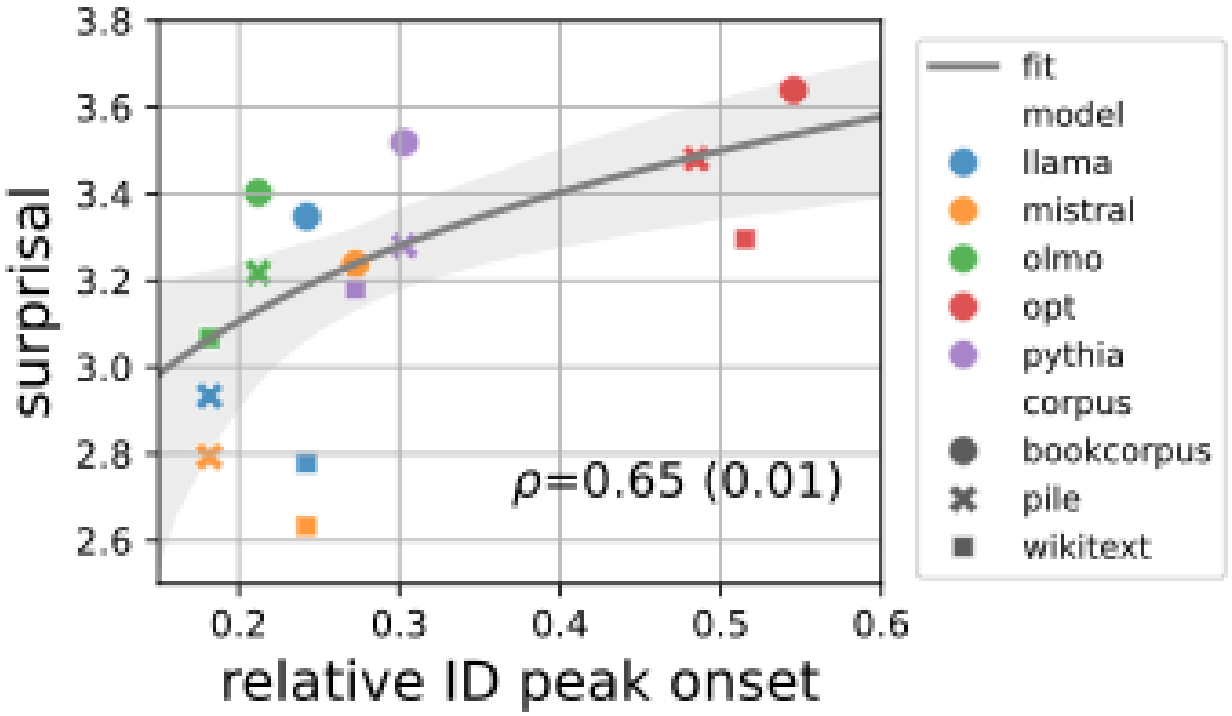


Syntactic/semantic probes and the ID peak





An earlier ID peak correlates with lower surprisal, across LMs and datasets



Ad-interim summary

- Across LMs, *intrinsic dimensionality* orders of magnitude smaller than superficial hidden-state dimensionality
- All models have an early-to-mid-layer ID *peak*
- This expansion in ID might be the locus of “deep” linguistic processing
  - it only emerges during training and it disappears when LM processes random text
  - it marks area of maximum cross-model similarity
  - Downstream/syntactic/semantic tasks reach asymptotic top performance under the peak…
  - … but surface task performance is low/drops
  - The earlier the peak, the lower the surprisal across LMs and datasets

Outline

- Motivation and background
- Intrinsic dimensionality as a footprint of linguistic processing
- **A more granular look at intrinsic dimensionality and syntactic complexity**
- Conclusion

Peeking inside the intrinsic dimensions

- ID might broadly captures some form of “linguistic complexity”
- Can we say something more about *which kind* of linguistic complexity?
- Work in progress with Francesca Franzon, Emily Cheng and Iria De Dios



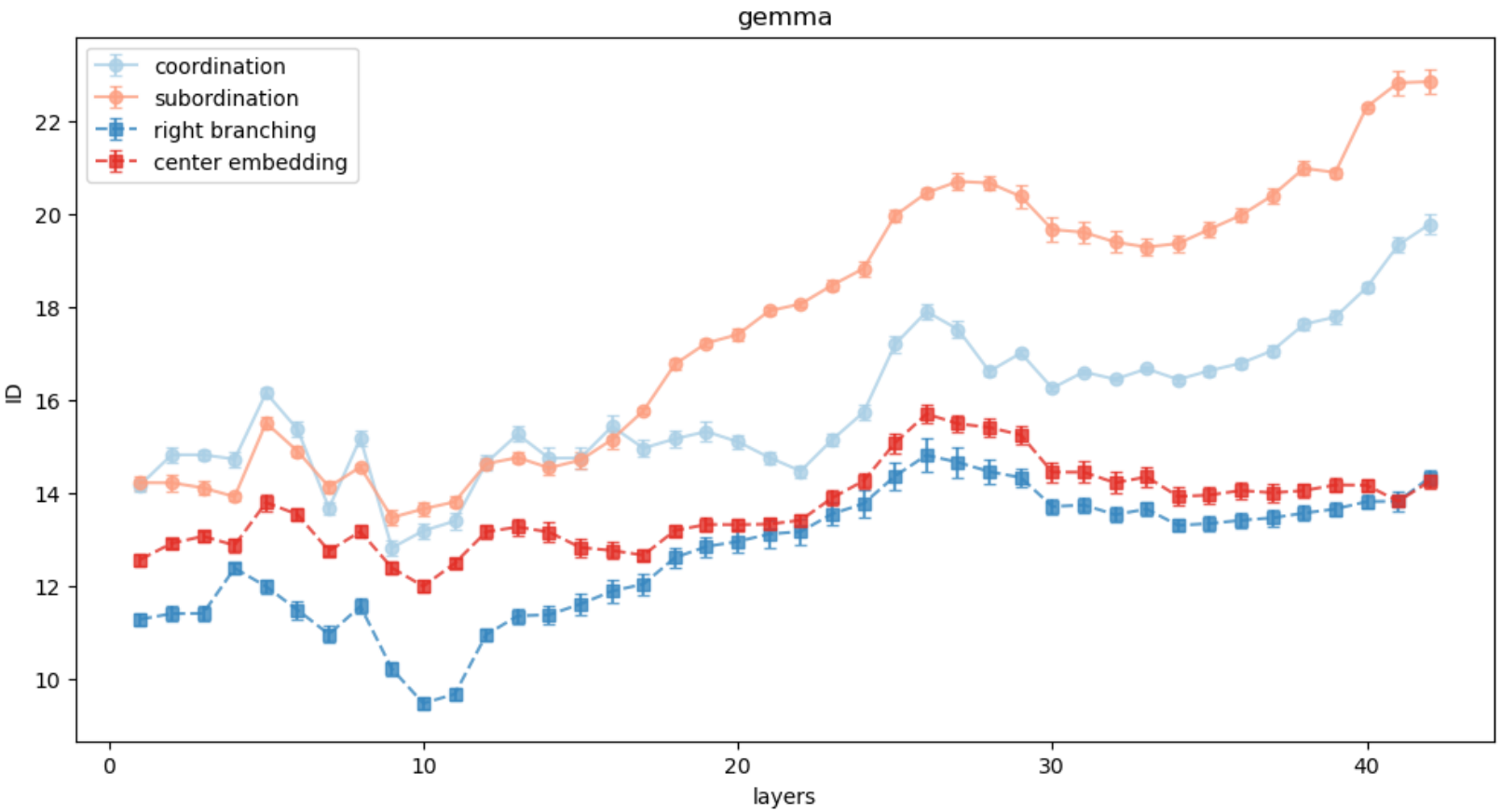
Two kinds of syntactic complexity

- Degrees of nesting
  - Coordination (flat, “easy”):
    - [The blacksmith is babbling] and [Daniel is thinking] and [the librarian is doubting] and [the politician is escaping]
  - Subordination (multiple nesting, “difficult”):
    - [The blacksmith is babbling [that Daniel is thinking [that the librarian is doubting [that the politician is escaping]]]]
- Processing effort
  - Right branching (“easy”):
    - The politician advised the potters [that were waiting]
  - Center embedding (“difficult”):
    - The potters [that the politician advised] were waiting

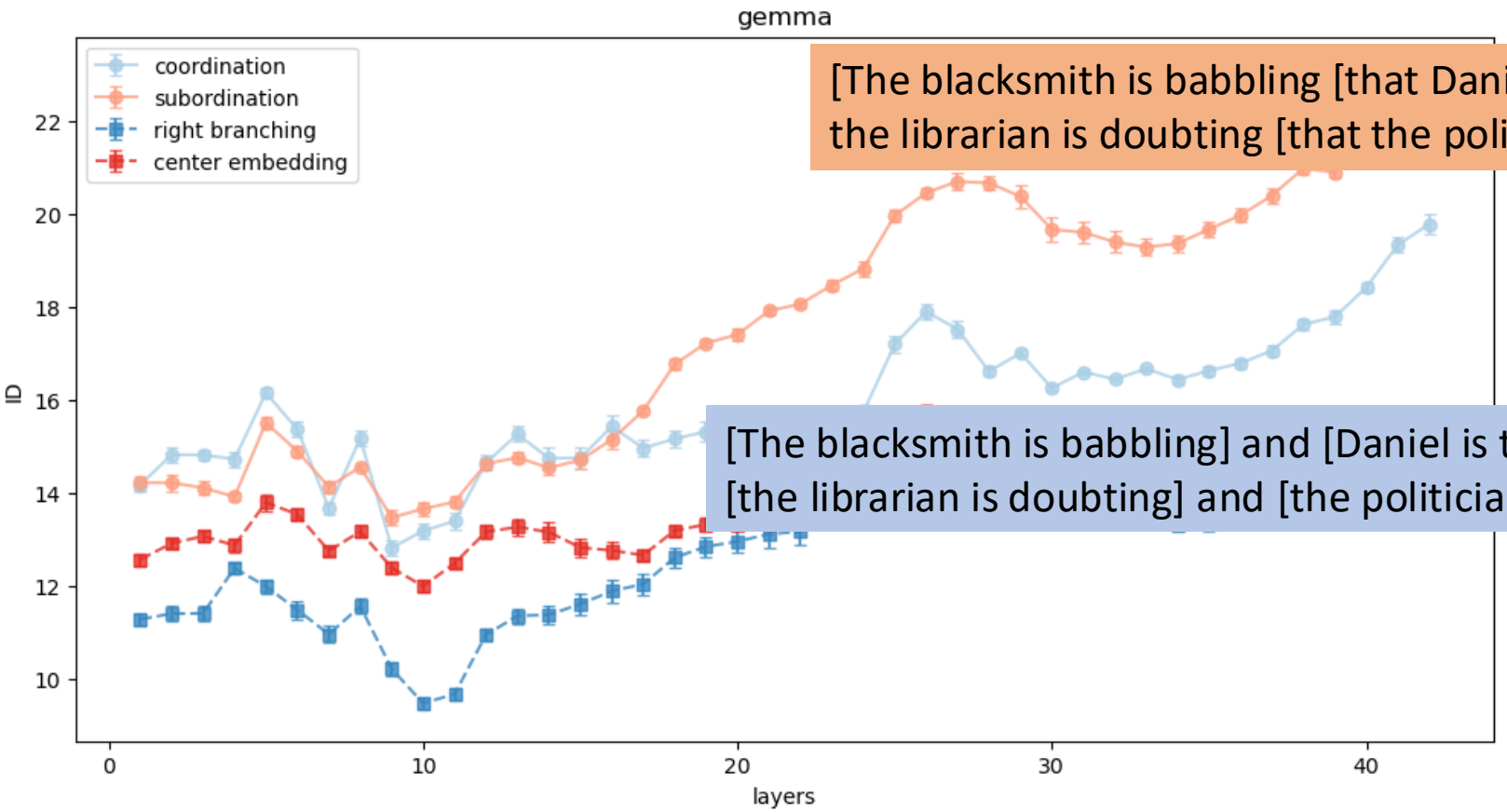
Methodology

- **LMs:** Mistral-7B, Llama-3-8B, pythia-12b, Qwen2.5-14B, gemma-2-9b, OLMo-2-13B
- **Data:** for each condition, 5 sets of 10k automatically generated sentences, lexically matched between coordination/subordination and right/center embedding conditions
- ID computation:
  - Feed each set of 10k sentences to a model
  - Extract last-token hidden-state vectors at each layer
  - Compute ID
- Process repeated 5 times to get variance around ID estimates

Gemma


gemma
coordination
subordination
right branching
center embedding
ID
layers
22
20
18
16
14
12
10
0
10
20
30
40

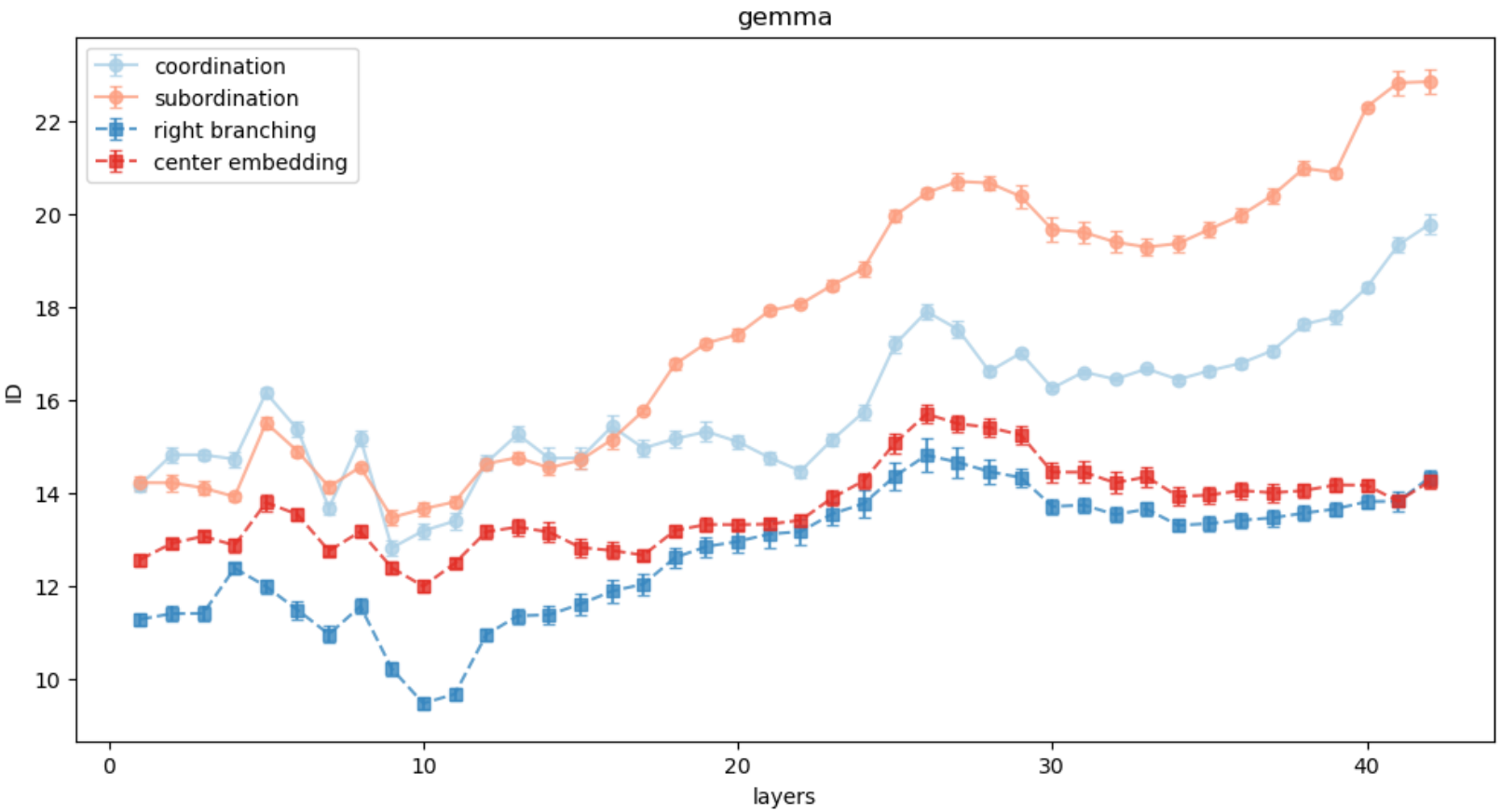
Gemma



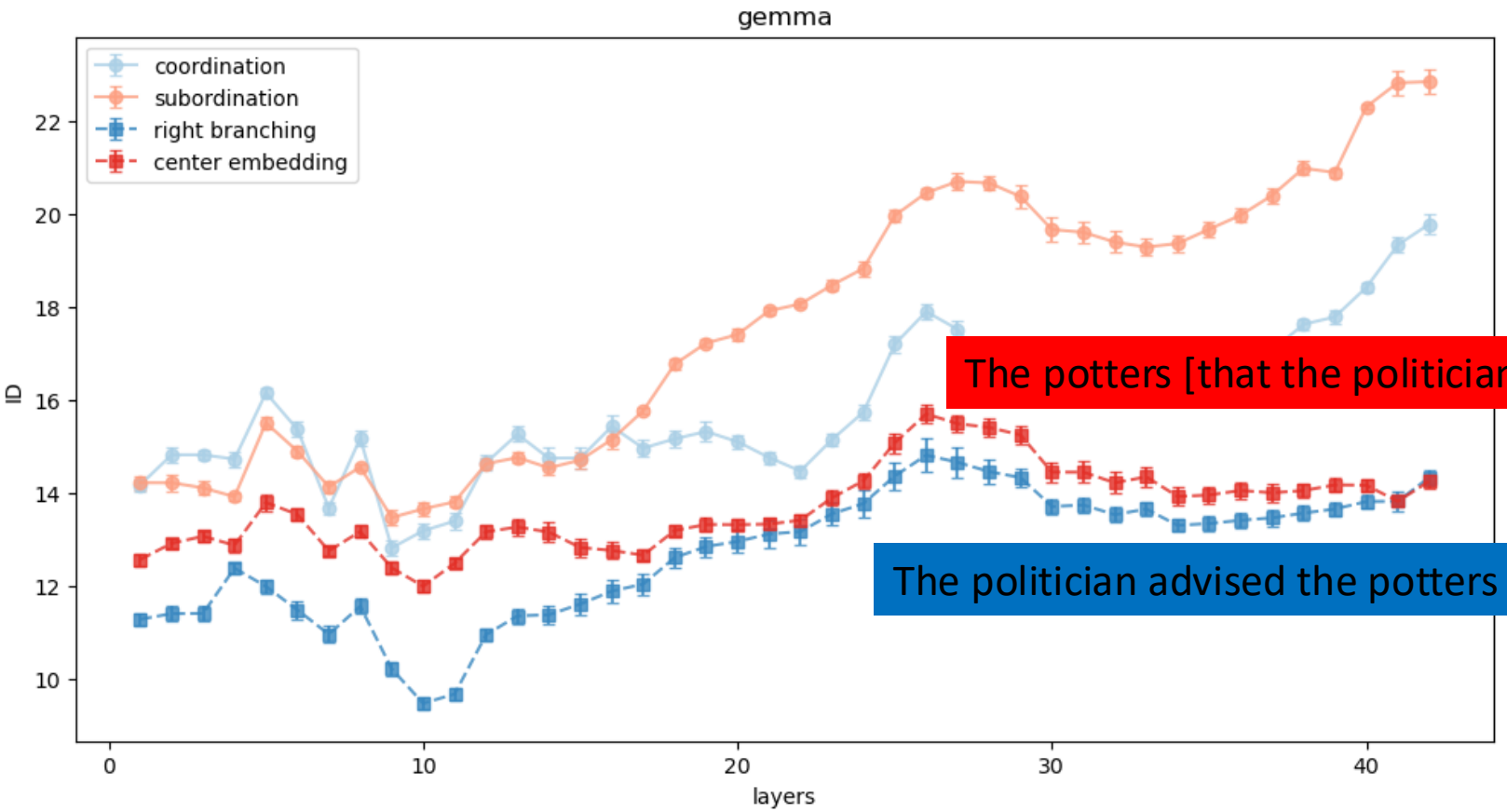
[The blacksmith is babbling [that Daniel is thinking [that the librarian is doubting [that the politician is escaping]]]]

[The blacksmith is babbling] and [Daniel is thinking] and [the librarian is doubting] and [the politician is escaping]

Gemma


gemma
coordination
subordination
right branching
center embedding
ID
layers
22
20
18
16
14
12
10
0
10
20
30
40

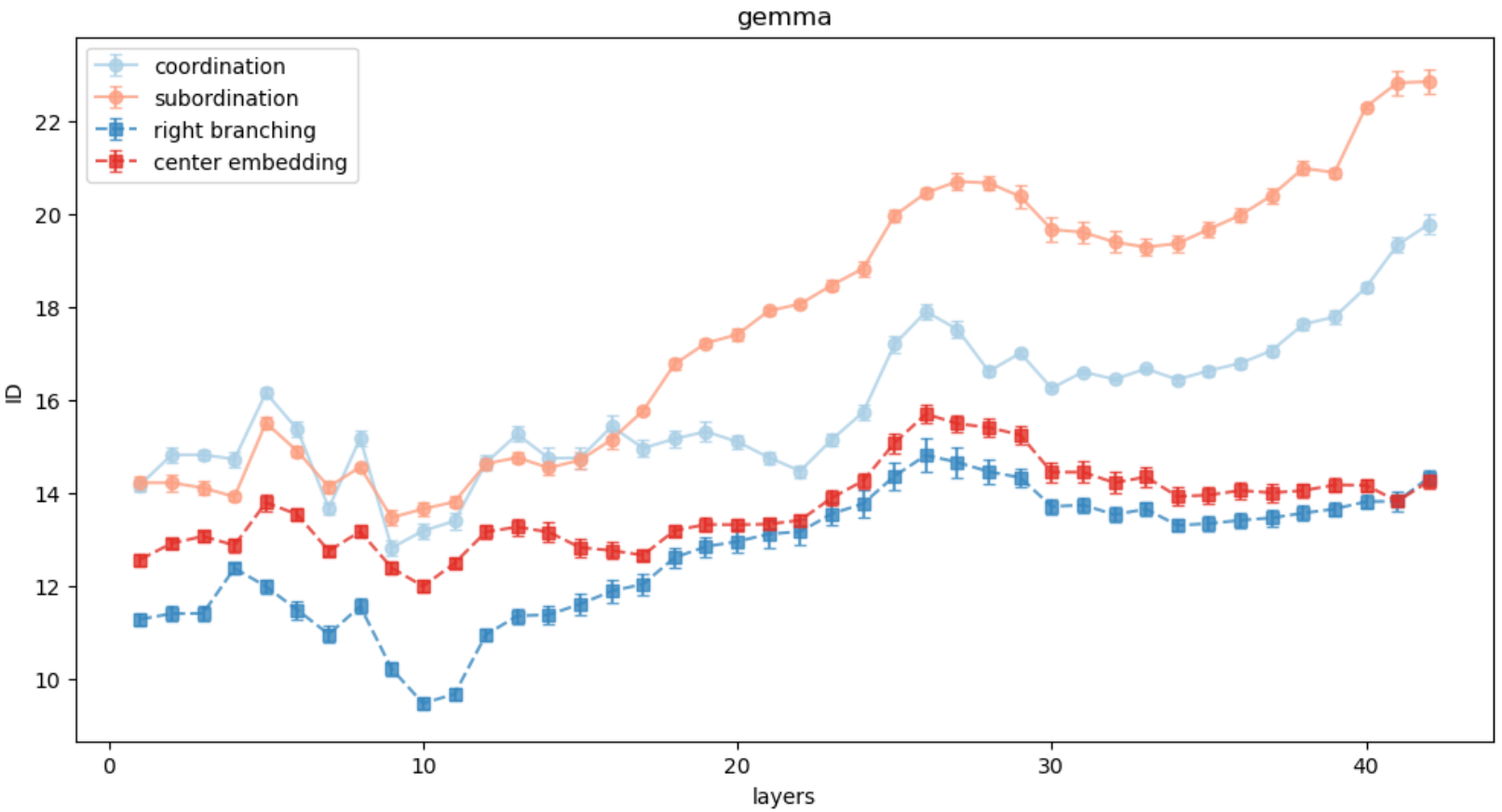
Gemma



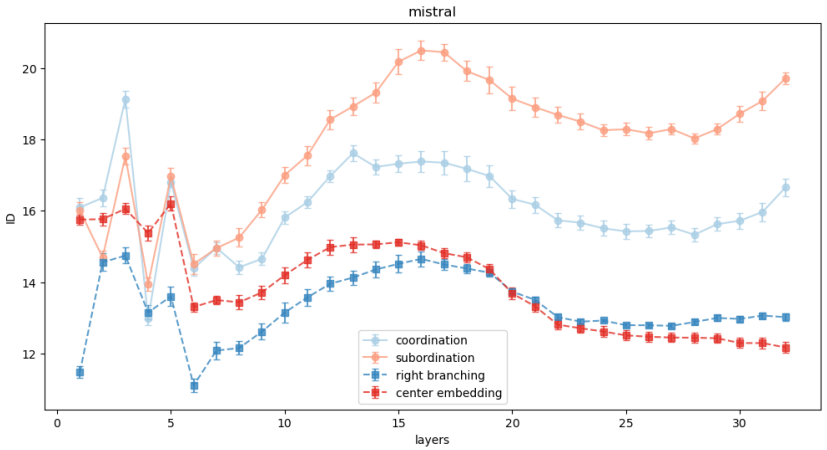
The potters [that the politician advised] were waiting

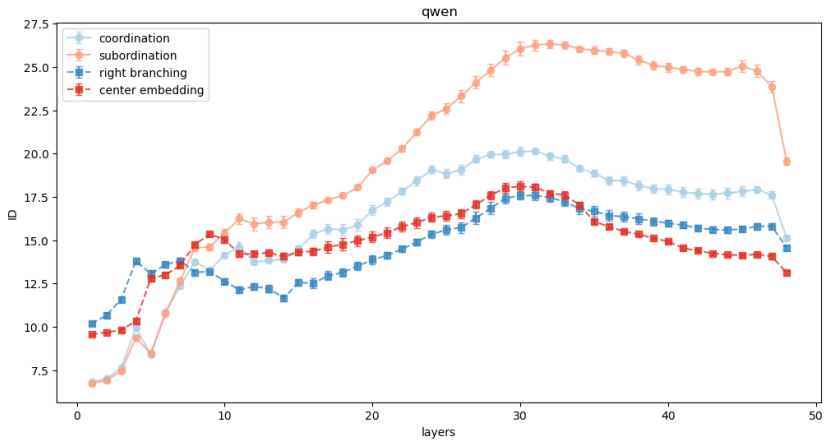
The politician advised the potters [that were waiting]

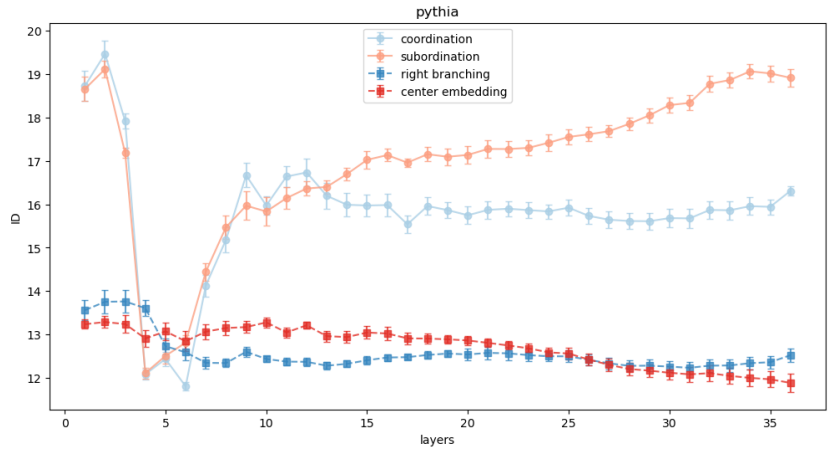
Gemma

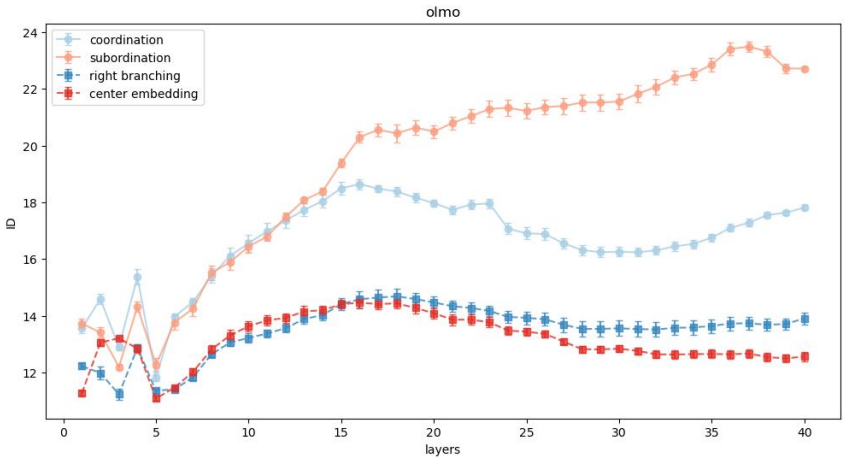

gemma
coordination
subordination
right branching
center embedding
ID
layers
22
20
18
16
14
12
10
0
10
20
30
40

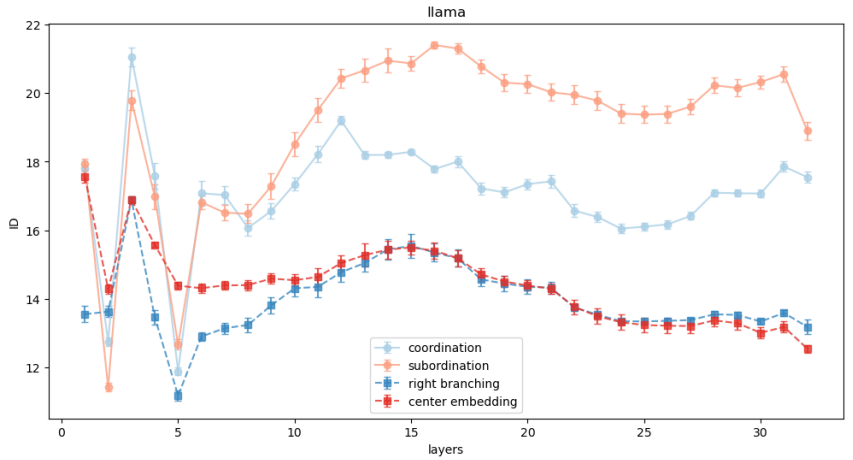
Other models



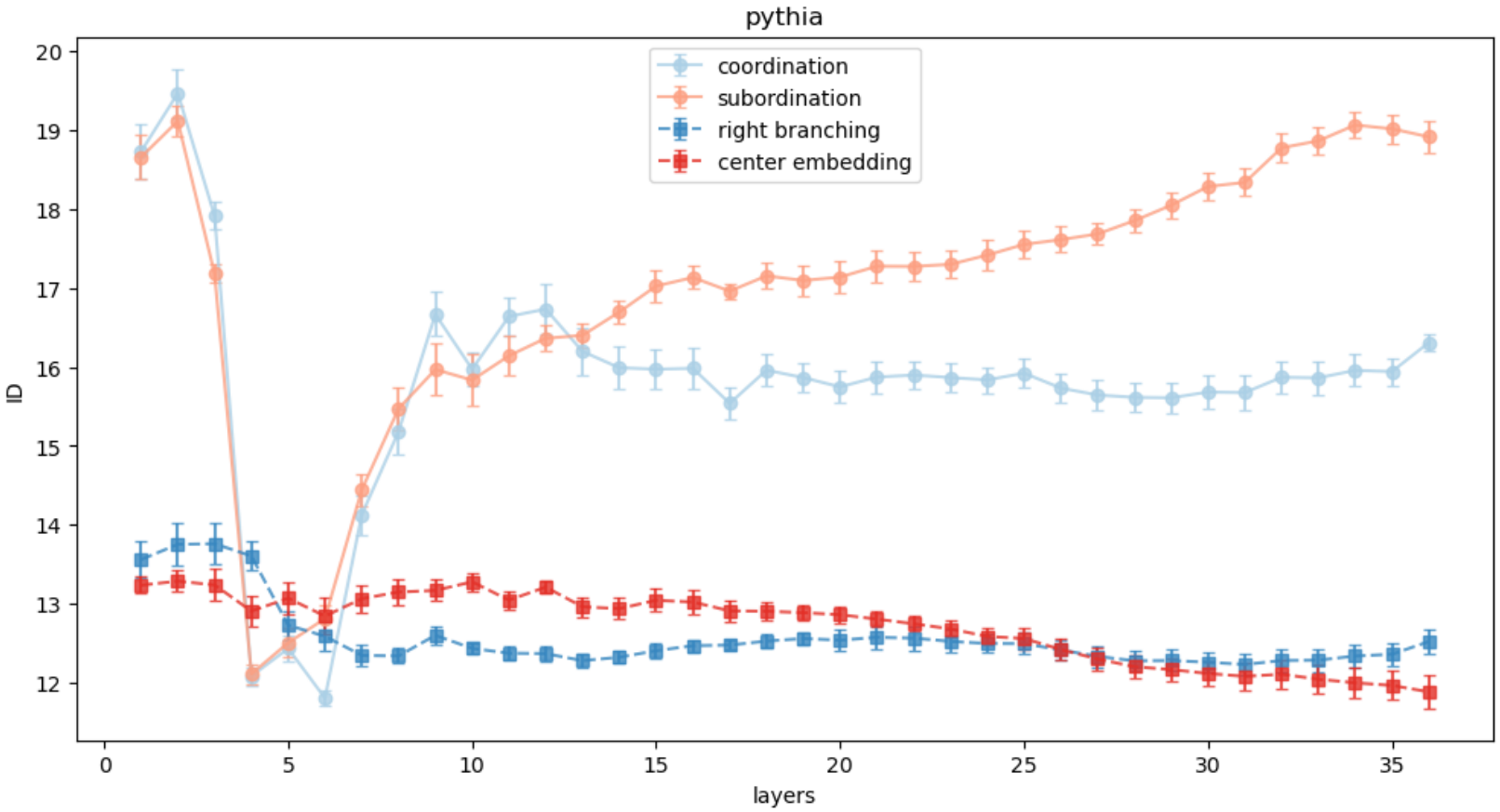




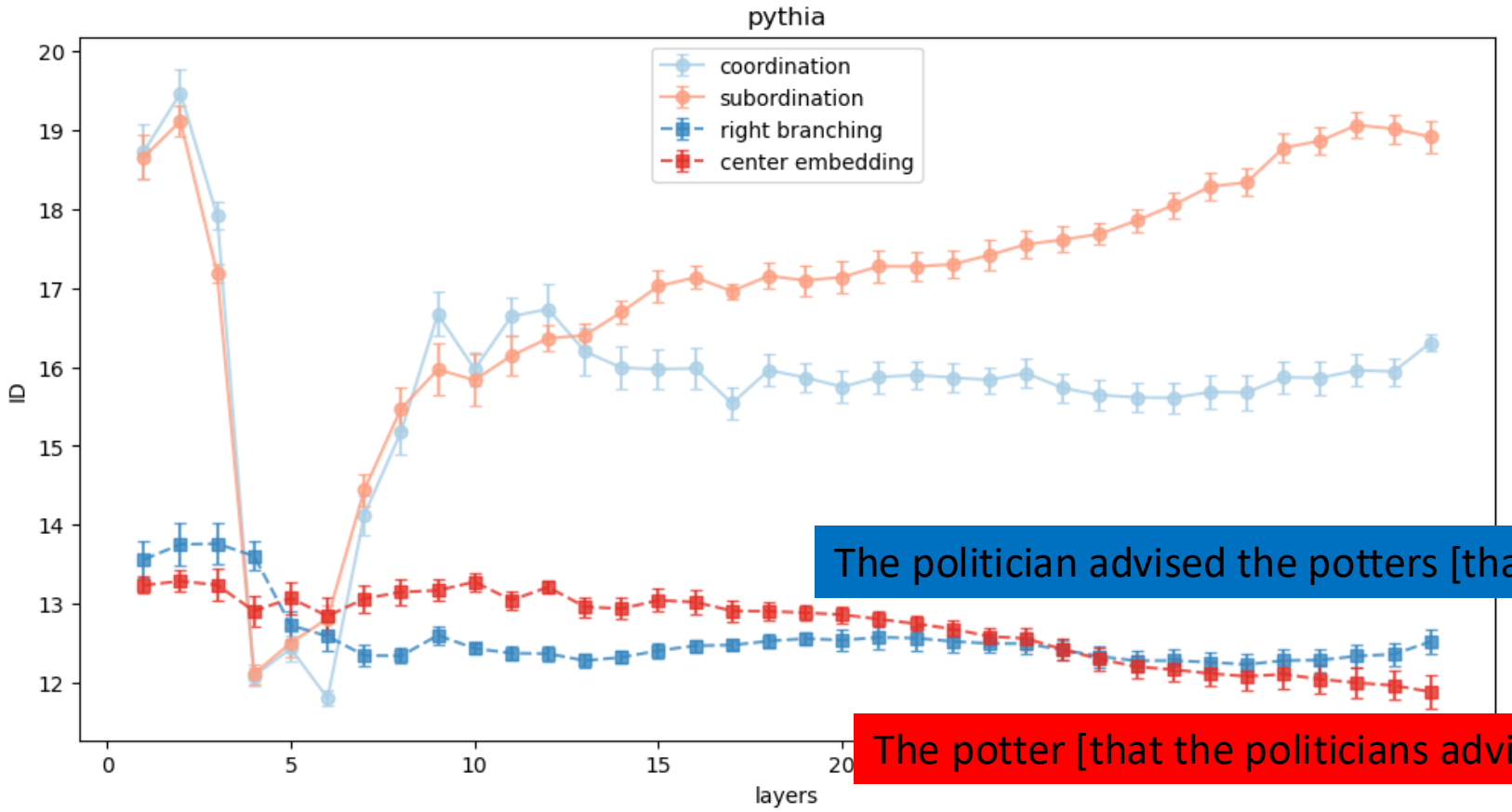




Pythia


pythia
coordination
subordination
right branching
center embedding
ID
layers
20
19
18
17
16
15
14
13
12
0
5
10
15
20
25
30
35

Pythia



The politician advised the potters [that were waiting]

The potter [that the politicians advised] were waiting

Two kinds of complexity detected by different ID profiles

- Deeper nesting leads to higher ID, compared to a similarly long but flatter structure, in the **later** layers of the model
- Processing complexity leads to higher ID, compared to a similar easier-to-parse structure, in the **early-mid** layers of the model
- Effects strikingly consistent across models

Outline

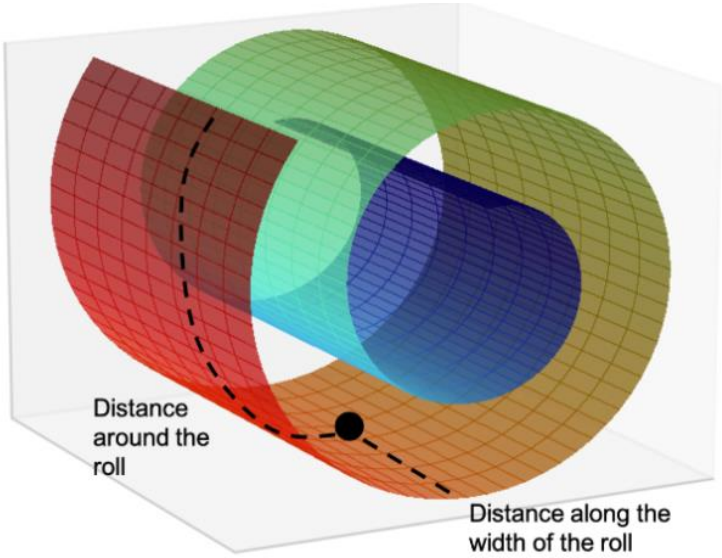
- Motivation and background
- Intrinsic dimensionality as a footprint of linguistic processing
- A more granular look at intrinsic dimensionality and syntactic complexity
- **Conclusion**

Main takeaways

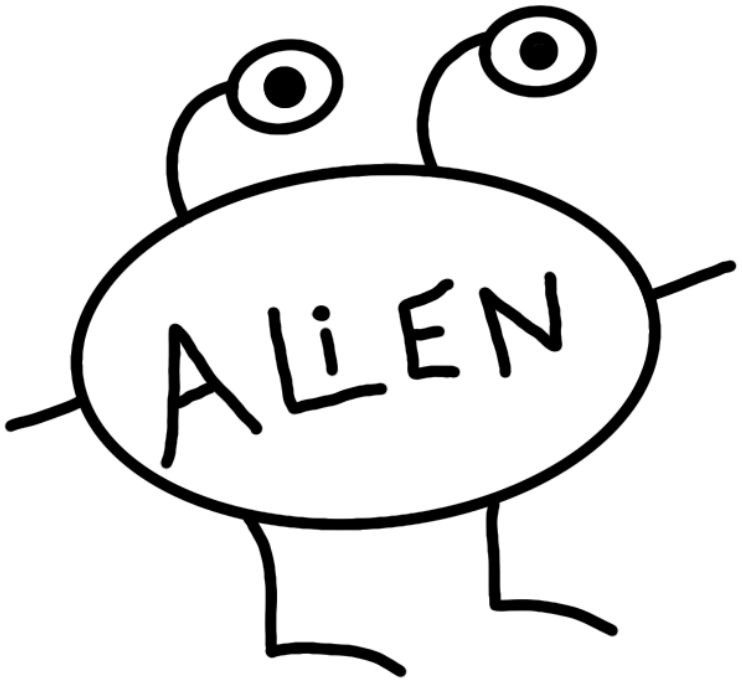
- ID estimation can be used to indirectly characterize what's going on on each layer of a LM
- Our first batch of experiments suggests different phases of processing, with deeper syntactic/semantic processing taking place during a fixed phase of ID expansion
- Ongoing work applying ID estimation to more specific subsets of linguistic data
- Preliminary (but robust) evidence for different footprints of nesting and processing difficulty (still to be related to generic ID peaks)

Moving forward:

Can we say something about what those intrinsic dimensions are?



THANK YOU!!!



https://marcobaroni.org/alien/