1. Supplementary Material

1.1. Empirical manifold capacity and theoretical manifold capacity

1.1.1. Empirical manifold capacity

In this section, we provide detailed description about how to find empirical manifold capacity. Given P object manifolds, N_c, the critical number of feature dimensions, is defined as the necessary number of feature dimensions so that P object manifolds, with randomly assigned $+/-$ labels for each manifold, can be linearly separated half the time on average (see (Stephenson et al., 2019)). The empirical manifold capacity is defined as P/N_c, which is the ratio between number of object manifold and the critical number of feature dimensions. To find N_c, a bisection search is performed until either the linearly separated fraction is within an error tolerance range $\epsilon = 0.05$ or the number of iteration exceeds 100. If the number of feature dimensions N is larger than N_c, then the fraction of linearly separable dichotomies is close to 1, and the data is in the linearly separable regime. Conversely, if the number of feature dimensions N is smaller than N_c, then the fraction of linearly separable dichotomies is close to 0, and the data is in the linearly inseparable regime.

In our experiments, we first randomly sample 20 instances for each manifold to perform the analysis. Then, for each candidate feature dimension in the bisection search, we sample 51 randomly assigned dichotomies to compute the linearly separable fraction. We use features extracted from pre-trained `bert-base-cased` model. Note that we exclude the embedding layers in this analysis due to the overlapping data point between manifolds as reported in Section 1.3.

1.1.2. Theoretical manifold capacity

Theoretical capacity used here is Mean-Field Theoretical Manifold Capacity described in Section 2 of the main text. We use $\kappa = 10^{-8}$ and $n_t = 300$, in which κ is the margin size and n_t is the number of Gaussian vectors to sample per manifold (see (Chung et al., 2018)). We also use the same randomly chosen 20 instances from the simulation capacity analysis for each manifold.

Figure 1 shows a close match between simulation capacity and the MFT manifold capacity observed in various linguistic tasks, measured across the hierarchy of `bert-base-cased` model.

1.2. Model architecture details

1.2.1. Pre-trained Models Details

We present briefly the pre-trained models that we used for the experiments.

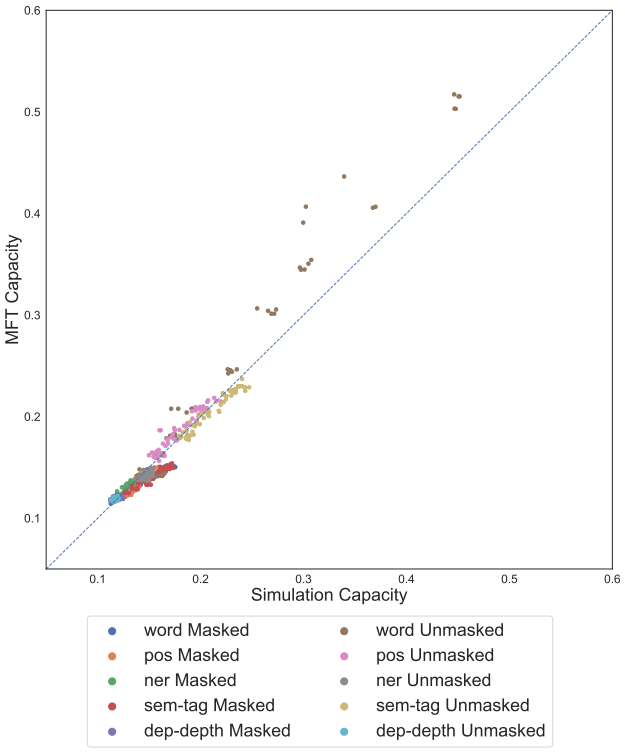

Figure 1. Simulation capacity vs. MFT capacity in `bert-base-cased` model.

- **BERT** `bert-base-cased`. 12-layer, 768-hidden, 12-heads, 110M parameters.
- **RoBERTa** `roberta-base`. 12-layer, 768-hidden, 12-heads, 125M parameters.
- **ALBERT** `albert-base-v1`. 12 repeating layers, 128 embedding, 768-hidden, 12-heads, 11M parameters.
- **DistilBERT** `distilbert-uncased`. 6-layer, 768-hidden, 12-heads, 66M parameters. The model distilled from the BERT model `bert-base-uncased` checkpoint.
- **OpenAI-GPT** `openai-gpt`. 12-layer, 768-hidden, 12-heads, 110M parameters.

For each pre-trained model, input text is tokenized using its default tokenizer and features are extracted at token level.

1.2.2. Fine-tuned Model Details

We fine-tuned BERT `bert-base-cased` model on POS downstream task with the following hyper-parameters:

- Epsilon for Adam optimizer: $1e{-}8$.

- Initial learning rate for Adam: $5e{-5}$.
- Max gradient norm: 1.
- Maximum total input sequence length after tokenization: 128. Longer sequences are truncated and shorter sequences are padded.

1.3. Datasets and Manifolds Details

In this section, we provide some information about the labels defining the manifolds for each task with some additional details (e.g., overlapping).

1.3.1. WORD

Labels are the following: `the, of, to, in, and, for, that, is, it, said, on, at, by, as, from, with, million, was, be, are, its, he, but, has, an, will, have, new, or, company, they, this, year, which, would, about, says, market, more, were, his, billion, had, their, up, one, than, some, who, been, stock, also, other, share, not, we, when, last, if, years, shares, all, president, first, two, sales, after, inc., because, could, out, trading, there, only, business, do, such, can, most, into.`

Note that, by definition, there is no overlapping between the manifolds.

1.3.2. PART-OF-SPEECH

Labels are the following: `NN, IN, NNP, DT, JJ, NNS, CD, RB, VBD, VB, CC, TO, VBZ, VBN, PRP, VBG, VBP, MD, POS, PRP$, WDT, JJR, NNPS, RP, WP, WRB, JJS, RBR, EX, RBS, PDT, FW, WP$.`

Labels are described in https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

There is 0.032% of overlapping pairs of words in the embedding layer due to the occurrence of a same word at the same position in multiple sentences with a different POS label. However, as expected, there is no overlapping for higher layers.

For the POS open-word class and closed-word class analysis, we used the following assignment of POS tags:

- Open-word class: `JJ, JJR, JJS, RB, RBR, RBS, NN, NNS, NNP, NNPS, VB, VBD, VBG, VBN, VBP, VBZ, FW`
- Closed-word class: `IN, DT, CD, CC, TO, PRP, MD, POS, PRP$, WDT, RP, WP, WRB, EX, PDT, FW, WP$`

For the ambiguous words analysis, we used the following words with associated POS tags: `back (RP, RB, JJ, NN), cut (VBN, VBD, NN, VB), set (VBD, VB, NN, VBN), close (NN, RB, JJ, VB), lower (RBR, VB, JJR), closed (VBD, VBN, JJ), estimated (JJ, VBD, VBN), call (NN, VB, VBP), come (VB, VBN, VBP), earlier (JJR, RBR, RB), pay (VB, VBP, NN), up (RP, RB, IN), over (IN, RB, RP), proposed (JJ, VBN, VBD), face (VBP, VB, NN), continued (JJ, VBD, VBN), down (IN, RB, RP), show (VB, VBP, NN), off (RP, RB, IN), better (JJR, RBR, RB), longer (RBR, RB, JJR), half (NN, PDT, DT), expected (VBN, JJ, VBD), buy (VB, NN, VBP), look (VB, NN, VBP)`

1.3.3. SEMANTIC TAGS

Labels are the following: `CON, REL, IST, DEF, LOC, PST, ORG, PER, DIS, SUB, EXS, NOW, PRO, HAS, AND, EXG, EXV, QUA, GPE, EXT, ENT, TIM, COO, APP, EPS, YOC, FUT, DOM, NOT, MOR, MOY, ENG, INT, TOP, ALT, ENS, ETV, POS, PRX, BUT, EPT, UOM, DST, QUE, NEC, EPG, IMP, ART, HAP, ETG, ROL, DOW, SCO, REF, COM, DEC, EXC, NAT, RLI, LES, EFS.`

Labels are described by Abzianidze & Bos (2017).

Note that there is no overlapping between the manifolds.

1.3.4. NAMED-ENTITY RECOGNITION

The NER dataset includes 18 labels described by Weischedel et al. (2011), consisting of 11 types (`GPE, LOCATION, WORK_OF_ART, EVENT, LAW, PRODUCT, LANGUAGE, PERSON, ORG, NORP, FAC`) and 7 values (`DATE, PERCENT, CARDINAL, TIME, QUANTITY, ORDINAL, MONEY`). With BIO tagging scheme, each label can occur either with `B-` (*beginning*) prefix or with `I-` (*inside*) prefix; there is an additional `O` (*outside*) label for words that are not named-entities.

There is 0.014% of overlapping pairs of words in the embedding layer due to the occurrence of a same word at the same position in multiple sentences with a different NER label. However, as expected, there is no overlapping for higher layers.

1.3.5. Dependency Depth

We select dependency depths from 0 to 21. From depth 18 to 21, we have respectively 12, 12, 5, 4 samples occurring in the corpus (instead of 50 for other depths).

Note that there is no overlapping between the manifolds.

1.4. Additional Experiments

1.4.1. Random baseline control for manifold capacity

We compare in Figure 2 the manifold capacity to three different manifold capacity baselines:

- **Lower bound**. The lower bound capacity LB is defined as the classification capacity of unstructured manifolds and only depends on the number of samples in each manifold.

$$LB = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{l_i}{2}} \tag{1}$$

  where LB is lower bound capacity, n is the number of manifolds, l_i is the number of samples in manifold i (see (Stephenson et al., 2019)).

- **Randomly initialized (untrained) model**. All model weights are set to a random number. Note that this random initialization has also an impact on the embedding layer.

- **Shuffled label manifolds**. The manifolds are shuffled without repetition and the number of samples for each manifold are preserved.

For both masked and unmasked data from `bert-base-cased` model, the capacity of shuffled label manifolds matches closely with the lower bound capacity, suggesting that randomly assigned manifold in different layers and linguistic tasks follow closely with the lower bound capacity.

Concerning the untrained model with random weights, in unmasked data, the capacities in the embedding layer are higher than lower bound and lower than the capacities in the pre-trained model. This reflects the fact that word vectors are already somewhat separated in the embedding layer, and the random weights don't improve or decrease the capacity. For the masked data with untrained model, the manifold capacity decreases across layers. The trends observed here are similar to prior work by Jawahar et al. (2019). Note that as observed by Gaier & Ha (2019), structured manifolds could emerge even in untrained models.

1.4.2. Analysis of Raw SVM Fields Distribution of POS manifold

We report in Figure 3 the raw SVM fields distribution of POS manifold with `bert-base-cased` model. The raw SVM fields distribution, despite of having a different distribution shape, shows similar trend across layers with the normalized SVM fields distribution described in the main text for both masked and unmasked dataset. The accuracy for raw SVM field distribution matches exactly the accuracy for normalized SVM fields distribution because normalization doesn't change sign of the fields. For unmasked data, the peak of the field distribution and the right tail moves slightly to the negative direction in all different train/test splits. For masked data, although the peak shifts to the negative direction, the right tail of the distribution extends to the positive direction in all different train/test splits, representing an increase in accuracy across layers.

1.4.3. Geometric Properties Evolution through Sequential Layers across Linguistic Tasks and Models (Additional Figures)

We report geometric properties (manifold capacity, radius, dimension and center correlation) for word, semantic tags, NER and dependency depth manifolds for the different models.

Word For word manifolds, as reported in Figure 4, similarly to POS manifold, the capacity increases for unmasked data and decreases for masked data in all the different models. In both masked and unmasked cases, the trend is clear and steep. In the masked case, the inputs are masked and feature vectors values only depend on the positional embedding and are not related to the word strings; since the model is trained to predict the masked word token, the word manifolds emerge across layers. In the unmasked case, the inputs are context-free embedding word vectors and are well separated; since the model tries to contextualize the word using its neighbor words, the word manifolds get entangled and lead to a decrease in word manifold capacity. The radius, dimension and center correlation measures also reflect the observed trend in the capacity. In the unmasked data, the radius, the dimension and center correlation of word manifolds increase across layers, representing manifold entangling. In the masked data, the dimension, radius and center correlation decrease across layers, suggesting manifold untangling.

Semantic Tags For semantic tag manifolds, as reported in Figure 5, similarly to word manifolds and POS manifolds, the capacity decreases in the unmasked dataset and increases in the masked dataset. Similarly to POS tags, semantic tags also have high correlation with context-free word; as reported by Bjerva et al. (2016), the per-word most

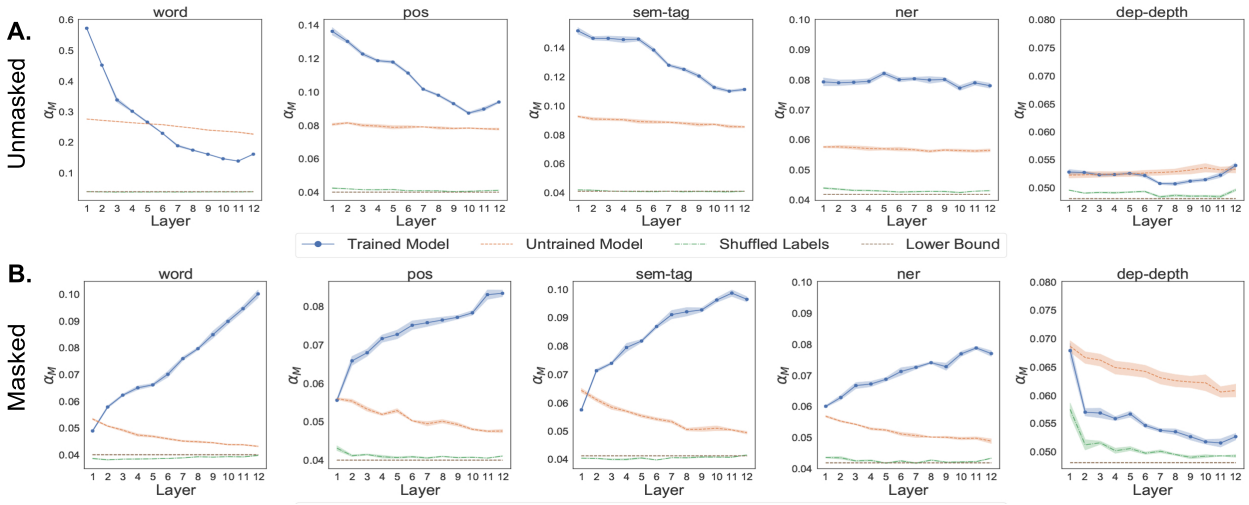

Figure 2. Randomly controls for manifold capacity in `bert-base-cased` model.

frequent class baseline for semantic tags has an accuracy of 77.39%. Therefore, in the masked case, since the model is trained to predict the word tokens which share information with the semantic tags, the manifold capacity increases. In the unmasked case, the inputs are word embedding vectors, which carry information about semantic tags, and the model tries to contextualize the inputs by their neighboring words. Contextualization can both entangle semantic tags manifold by decreasing the correlation between word tokens and their semantic tags and untangle semantic tags manifold by gaining information from neighbor words. These two competing effects lead to an overall decrease in manifold capacity, but this decrease has a much less magnitude than the decrease in word manifold capacity (-0.6 vs. -0.06). Manifold radius, dimension and center correlation also have similar trend as POS and word manifolds.

Named-entity Recognition For NER manifolds, as reported in Figure 6, the different models express similar trend for both masked and unmasked data. For the unmasked data, the manifold capacity remains mostly unchanged across layers. This trend suggests a balance between losing information from correlation between words and NER label, and gaining information from contextualization by neighbor words. The geometric properties also show a competing effect between decreasing radius and increasing dimension and center correlation. For the masked data, the manifold capacity increases across layers (similar trend as word, POS and sem-tag). This trend is expected because the input tokens are masked and the model objective is to predict the masked word, which can carry some information about NER. Geometric properties show decreasing radius and center correlation, suggesting manifold untangling.

Dependency Depth For dependency depth manifolds, as reported in Figure 7, similar trend is observed for the different models in both masked and unmasked dataset. For unmasked data, the manifold capacity remains mostly unchanged. Manifold radius and dimension do not change significantly, while center correlation peaks at the intermediate layers. Since dependency depths are numerical values, higher center correlation may suggest a structured geometry relationship between different dependency depth clusters. Hewitt & Liang (2019) also reports similar results about syntactic parse tree peaks at the intermediate layers. For masked data, manifold capacity, radius and center correlation decreases across layers, while dimension increases. Generally, the manifold capacity and geometry measures for dependency depth manifolds are quite different from other manifolds. While other manifolds are *categorical* values, dependency depths are *numerical* values. A large capacity implies that category manifolds are well-separated for a classification task; however, since dependency depth manifolds have a numerical and transitive property, its geometry may not be optimized for classification capacity. Instead, dependency depth task may be explained better by a task that reflects such numerical and transitive properties such as a regression task, and the relation between the representation geometry and the regression performance will be explored as future work.

1.4.4. Correlation of Manifold Capacity and Task Performance in POS Fine-Tuned Model

When fine-tuning pre-trained `bert-base-cased` model for POS task, a strong correlation is observed between the POS manifold capacity and F1 score across update steps for

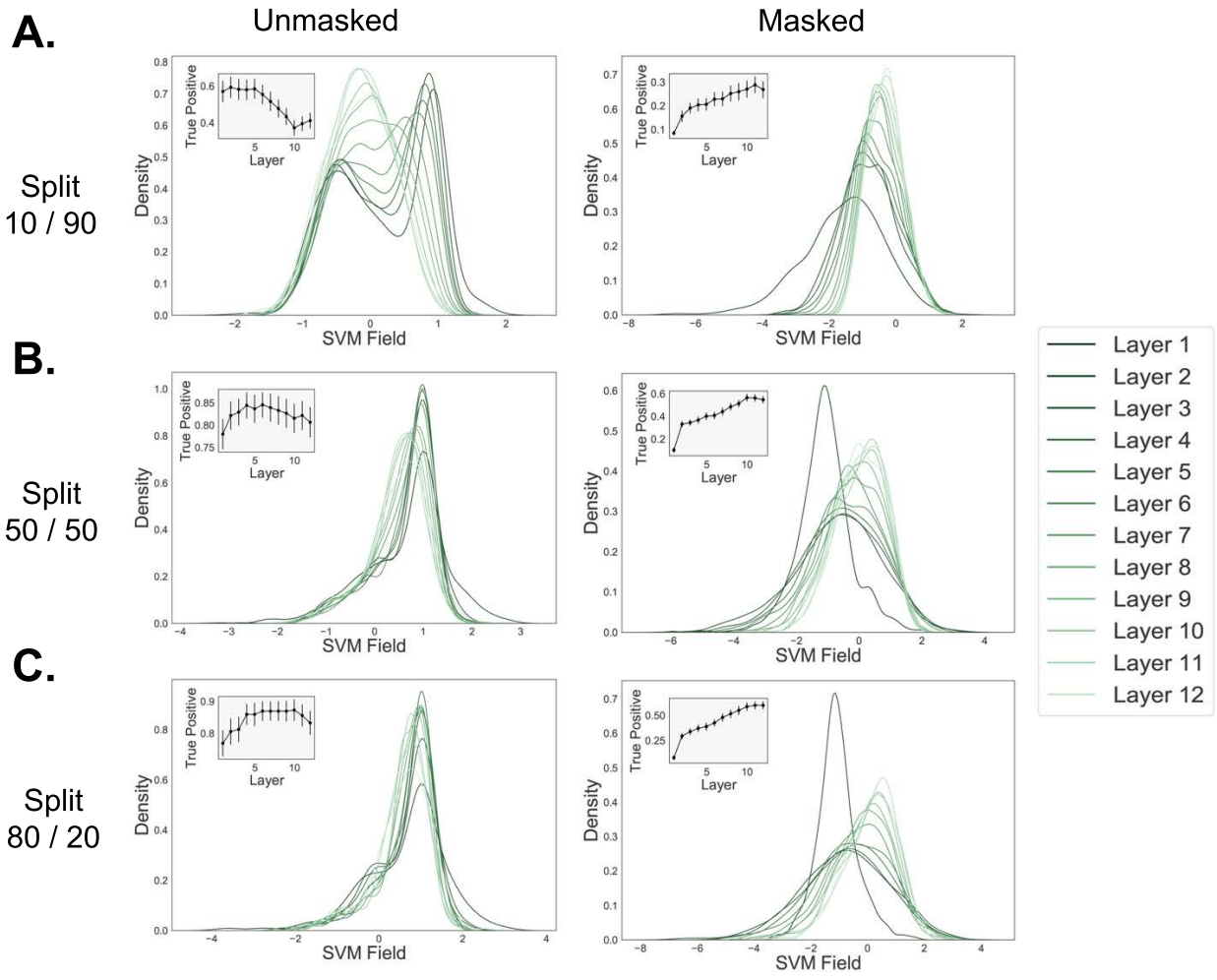

Figure 3. Raw SVM fields of POS manifold with `bert-base-cased` model.

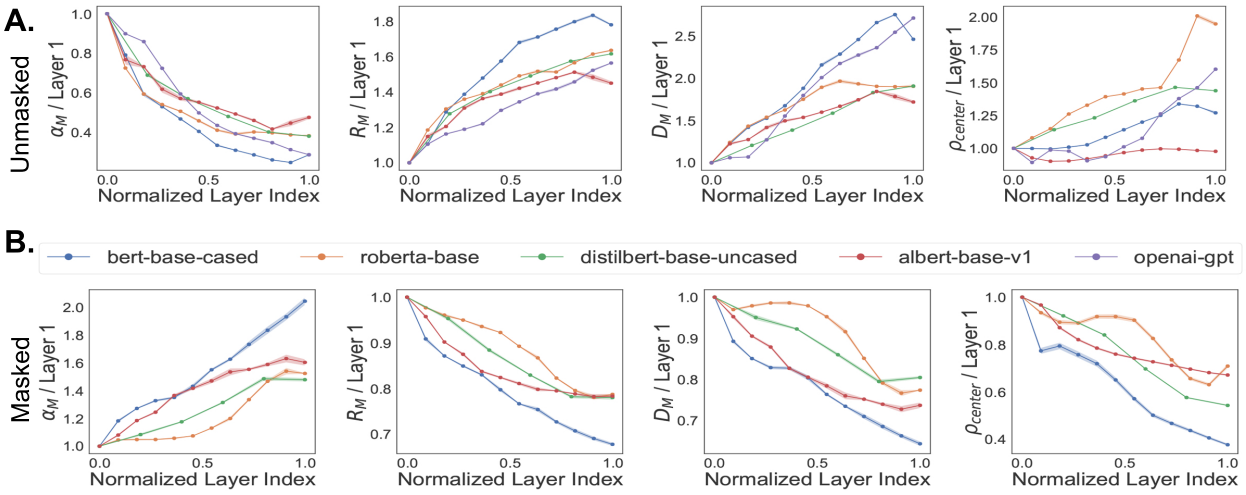

Figure 4. Geometric properties of word manifold in different models.

update step	raw capacity	F1
1	0.0903	0.04
5	0.0915	0.11
10	0.0998	0.34
20	0.1362	0.55
50	0.2361	0.87
Pearson correlation	0.9334	

Table 1. Correlation of raw manifold capacity and F1 in POS fine-tuned model, unmasked data.

update step	norm. capacity	F1
1	0.6111	0.04
5	0.6209	0.11
10	0.6839	0.34
20	0.9623	0.55
50	1.6274	0.87
Pearson correlation	0.9417	

Table 2. Correlation of manifold capacity (normalized by embedding layer) and F1 in POS fine-tuned model, unmasked data.

unmasked data, as reported in Table 1 and Table 2. Specifically, Pearson correlation for raw capacity and F1 score and for normalized capacity and F1 score are 0.9334 and 0.9417 respectively. This result suggests that manifold capacity can capture task performance (F1 score) in POS task. Note that asked data is not shown because masked token is never seen during fine-tuning.

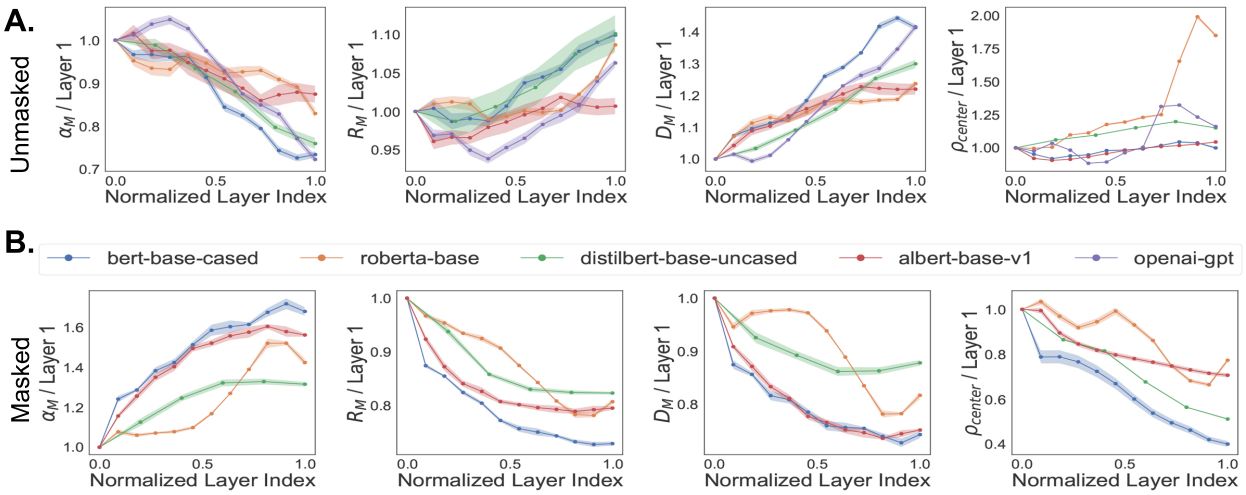

Figure 5. Geometric properties of semantic tags manifold in different models.

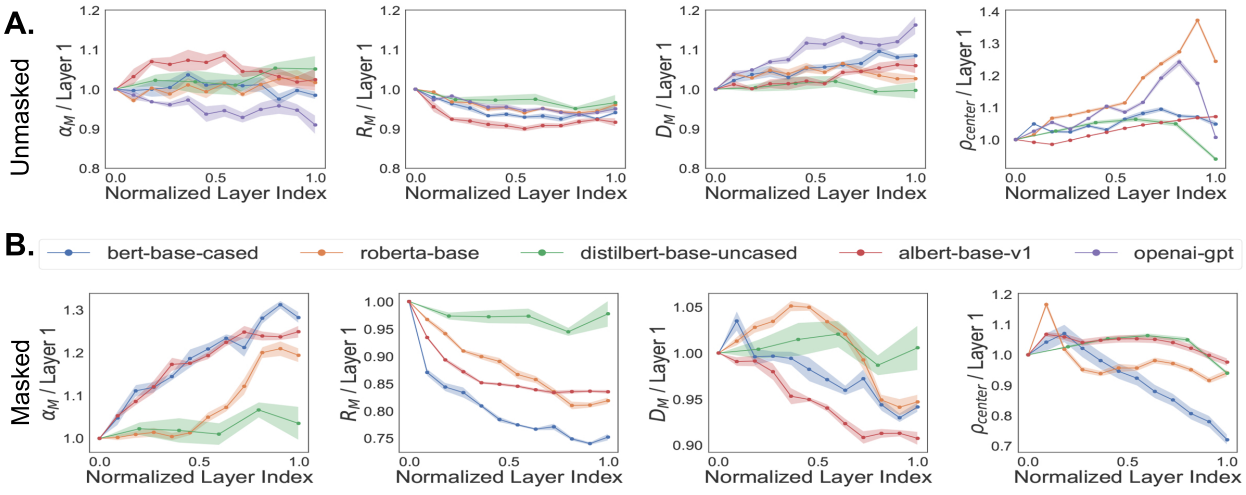

Figure 6. Geometric properties of NER manifold in different models.

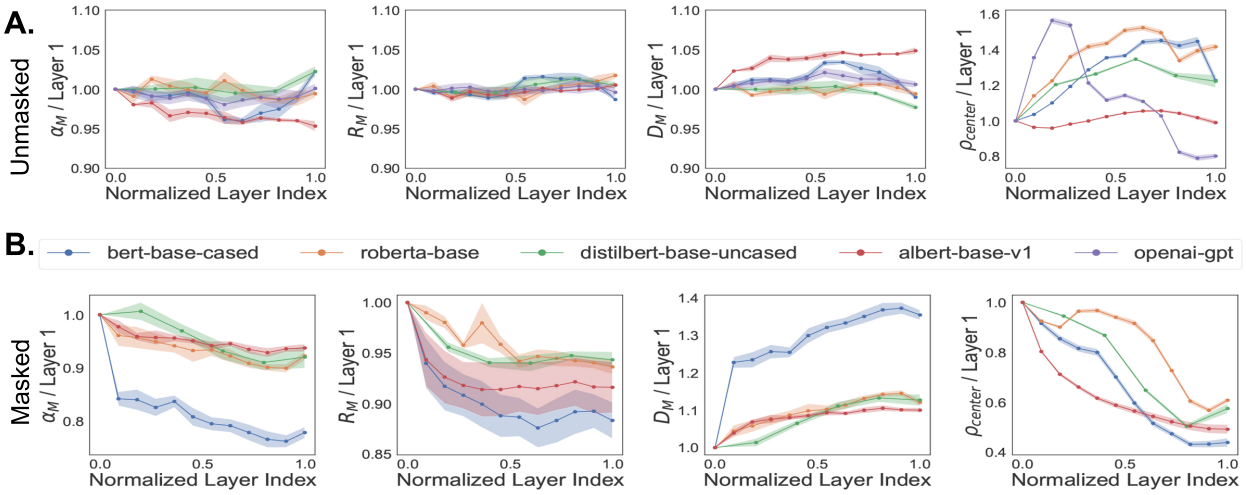

Figure 7. Geometric properties of dependency depth manifold in different models.

References

Abzianidze, L. and Bos, J. Towards universal semantic tagging. *arXiv preprint arXiv:1709.10381*, 2017.

Bjerva, J., Plank, B., and Bos, J. Semantic tagging with deep residual networks. *The 24th International Conference on Computational Linguistics*, 2016.

Chung, S., Lee, D. D., and Sompolinsky, H. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

Gaier, A. and Ha, D. Weight agnostic neural networks. *Advances in Neural InformationProcessing Systems*, 2019.

Hewitt, J. and Liang, P. Designing and Interpreting Probes with Control Tasks. In *Proceedings of EMNLP*, 2019.

Jawahar, G., Sagot, B., and Djame, S. What does bert learn about the structure of language? *Association for Computational Linguistics*, 2019.

Stephenson, C., Feather, J., Padhy, S., Elibol, O., Tang, H., McDermott, J., and Chung, S. Untangling in invariant speech recognition. In *Advances in Neural Information Processing Systems*, pp. 14368–14378, 2019.

Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradhan, S., Ramshaw, L., and Xue, N. Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation. Springer*, pp. 59, 2011.