

## G Compositionality scores without anisotropy correction

The raw compositionality scores can be found in [Table 10](#).

## H AUC of approximative probes

Model and representation	probe	AUC
BERT <sub>CLS</sub>	ADD	91.80
	W1	83.82
	W2	91.20
	LIN	92.57
	<b>AFF</b>	<b>93.20</b>
	MLP	92.74
RoBERTa <sub>CLS</sub>	ADD	99.93
	W1	99.84
	W2	99.93
	LIN	99.94
	AFF	99.93
	<b>MLP</b>	<b>99.94</b>
DeBERTa <sub>CLS</sub>	ADD	99.90
	W1	99.75
	W2	99.90
	LIN	99.92
	AFF	99.94
	<b>MLP</b>	<b>99.95</b>
GPT-2 <sub>last</sub>	<b>MLP</b>	<b>99.95</b>
	ADD	96.16
	W1	95.94
	W2	95.97
	LIN	96.21
	<b>AFF</b>	<b>99.18</b>
BERT <sub>AVG</sub>	MLP	98.32
	ADD	82.04
	W1	53.83
	W2	88.10
	LIN	88.68
	<b>AFF</b>	<b>90.63</b>
RoBERTa <sub>AVG</sub>	MLP	88.96
	ADD	97.51
	W1	92.73
	W2	98.49
	LIN	98.56
	<b>AFF</b>	<b>99.00</b>
DeBERTa <sub>AVG</sub>	MLP	98.88
	ADD	92.74
	W1	73.67
	W2	94.38
	LIN	94.89
	<b>AFF</b>	<b>96.21</b>
GPT-2 <sub>AVG</sub>	MLP	95.75
	ADD	99.60
	W1	97.90
	W2	99.64
	LIN	99.69
	<b>AFF</b>	<b>99.81</b>
	MLP	99.76

Table 11: AUC scores for probes trained on various percentages of the training set.

## I Mean deviation of phrase types by tree type

The mean deviation of the most common tree types can be found in [Figure 11](#).

## J Further named entity results

Named entity results can be found in [Figure 12](#) and [Figure 13](#).

Model and representation	Probe	Mean reconstruction score	Standard dev.
BERT <sub>CLS</sub>	ADD	0.9178	0.001159
	W1	0.8382	0.003599
	W2	0.9117	0.0007133
	LIN	0.9258	0.0002285
	<b>AFF</b>	<b>0.9322</b>	0.0002033
	MLP	0.9276	0.0002108
RoBERTa <sub>CLS</sub>	ADD	0.99935	$3.895 \times 10^{-6}$
	W1	0.99850	$2.612 \times 10^{-5}$
	W2	0.99937	$6.866 \times 10^{-6}$
	LIN	0.99946	$4.735 \times 10^{-6}$
	<b>AFF</b>	<b>0.99950</b>	$6.093 \times 10^{-6}$
	MLP	0.99947	$4.719 \times 10^{-6}$
DeBERTa <sub>CLS</sub>	ADD	0.99908	$4.070 \times 10^{-5}$
	W1	0.99762	$2.900 \times 10^{-5}$
	W2	0.99911	$1.399 \times 10^{-4}$
	LIN	0.99928	$8.963 \times 10^{-5}$
	<b>AFF</b>	<b>0.99972</b>	$1.542 \times 10^{-5}$
	MLP	0.99965	$2.323 \times 10^{-5}$
BERT <sub>AVG</sub>	ADD	0.8205	0.0003836
	W1	0.5383	0.007471
	W2	0.8893	0.03071
	LIN	0.8873	0.003071
	<b>AFF</b>	<b>0.9069</b>	0.002566
	MLP	0.8904	0.002988
RoBERTa <sub>AVG</sub>	ADD	0.9752	0.0001306
	W1	0.9274	0.001695
	W2	0.9850	0.0005092
	LIN	0.9858	0.0004573
	<b>AFF</b>	<b>0.9902</b>	0.0003076
	MLP	0.9890	0.0003981
DeBERTa <sub>AVG</sub>	ADD	0.9275	0.002634
	W1	0.7368	0.001575
	W2	0.9438	0.003321
	LIN	0.9493	0.003036
	<b>AFF</b>	<b>0.9625</b>	0.001814
	MLP	0.9590	0.002145
GPT-2 <sub>AVG</sub>	ADD	0.9960	0.0002833
	W1	0.9791	0.0001214
	W2	0.9965	0.0003359
	LIN	0.9970	0.0003036
	<b>AFF</b>	<b>0.9984</b>	0.0002617
	MLP	0.9979	0.0001634

Table 10: Mean reconstruction score (cosine similarity) and standard deviation of each approximative probe across 10 folds. Not corrected for anisotropy in each representation/model type.

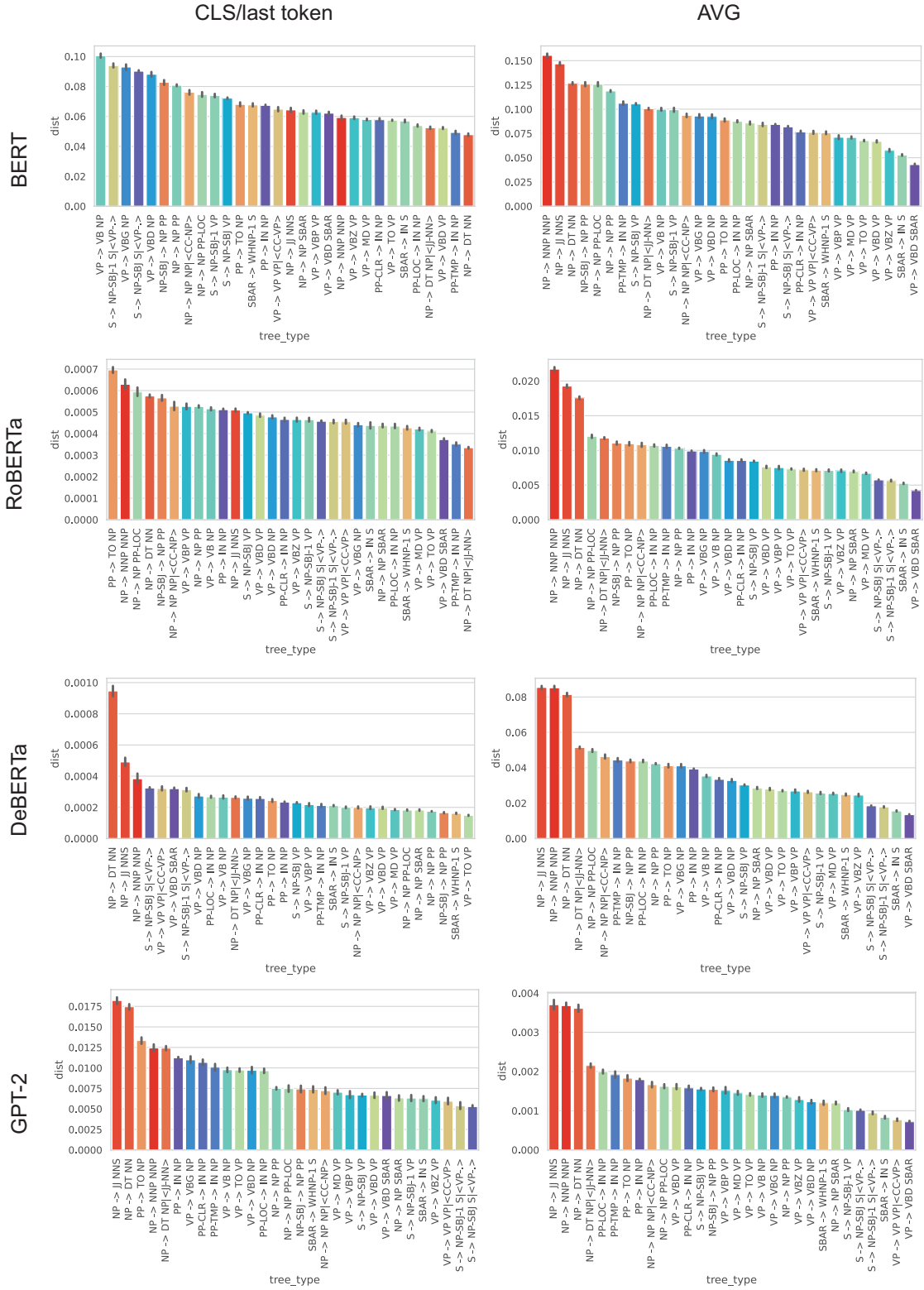


Figure 11: Mean deviation from predicted representation across full tree types.