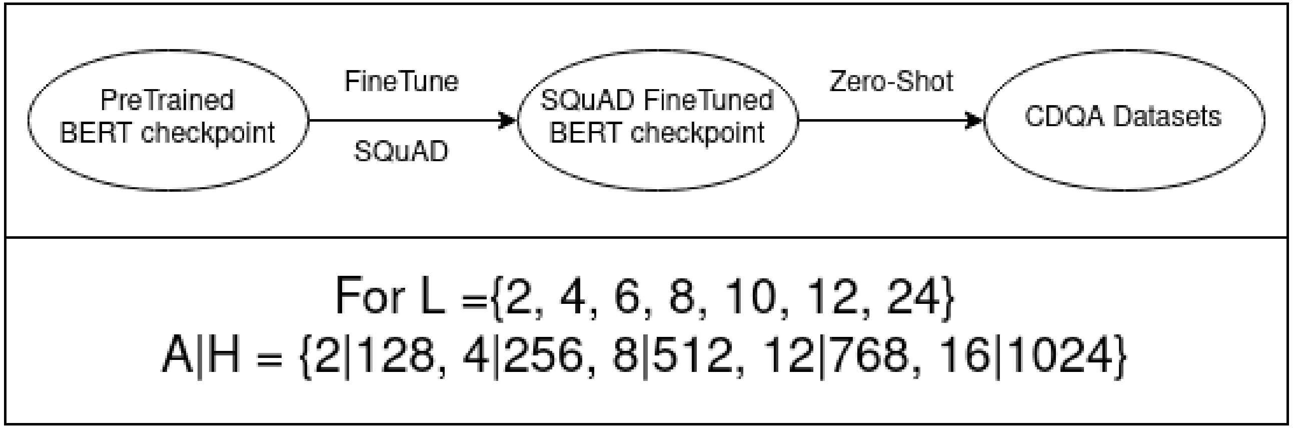
Model Architecture Analysis – Discussion

Experiment Motivation

To see the effects of various architecture configurations on CDQA performance.

Experiment Details - 1

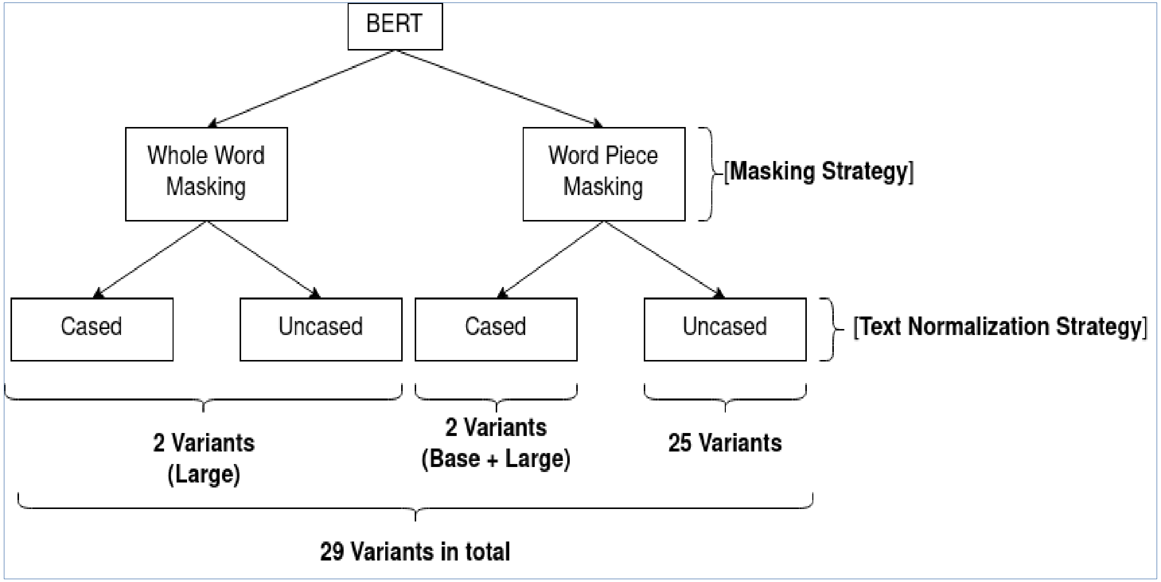
- We looked at **zero-shot CDQA performance changes** in *children (layers = 2, 4, etc.) & parent (layer = 24) configurations* of $\text{BERT}_{\text{BASE}}$ (layers = 12, attention_heads = 12, hidden_dim = 768)
- The overall pipeline is simply,



For L ={2, 4, 6, 8, 10, 12, 24}
A|H = {2|128, 4|256, 8|512, 12|768, 16|1024}

Experiment Details - 2

- We classify the models on 2 basis, to see the effect of each on performance
  - Masking – Whether it makes any difference to use whole word/word piece masking in CD's.
  - Normalization – Since it is known that CD data might contain many *exotic entities* not encountered in OD data, perhaps Cased models would perform better than uncased models.



Experiment Details - 3

Finally, we measure perplexity of $\text{BERT}_{\text{BASE}}$ on the training data of all five datasets to see whether there exists any correlation between the PPL. & performance.

Questions Asked & Results

[1st **3 slides relate to uncased word piece models** & next include across the board]

Datasets/Domains Used

- SQuAD [Open Domain]
- COVID-QA [Biomedical]
- DuoRC [Movies]
- TechQA [Customer technical support queries]
- CUAD [Legal]

Does scaling layers improve performance?

- Generally **yes** for Open-Domain data

SQuAD scores

L	A|H	EM|F1
2	2|128	29.58 | 41.44
4	2|128	38.59 | 49.19
8	2|128	44.88 | 55.38

- **Not always** for Closed-Domain data

COVID-QA scores

L	A|H	EM|F1
8	12 | 768	**21.3 | 38.02**
10	12 | 768	**19.47 | 35.45**

CUAD scores

L	A|H	EM|F1
10	2 | 128	**1.15 | 2.4**
12	2 | 128	**0.55 | 1.86**

Does scaling attention heads/hidden dimension improve performance?

- Generally **yes** for Open-Domain data

SQuAD scores

L	A|H	EM|F1
10	2 | 128	51.58 | 62.44
10	4 | 256	71.41 | 80.44
10	8 | 512	78.13 | 86.08

- **Not always** for Closed-Domain data

TechQA scores

L	A|H	EM|F1
12	4 | 256	**1.94 | 6.03**
12	8 | 512	**1.61 | 6.86**

CUAD scores

L	A|H	EM|F1
6	4 | 256	**1.23 | 2.74**
6	8 | 512	**0.74 | 1.95**

Does scaling both together improve performance?

- Generally **yes** for Open-Domain data

SQuAD scores

L	A|H	EM|F1
12	12 | 768	80.9 | 88.2
24	16 | 1024	83.49 | 90.6

- **Not always** for Closed-Domain data

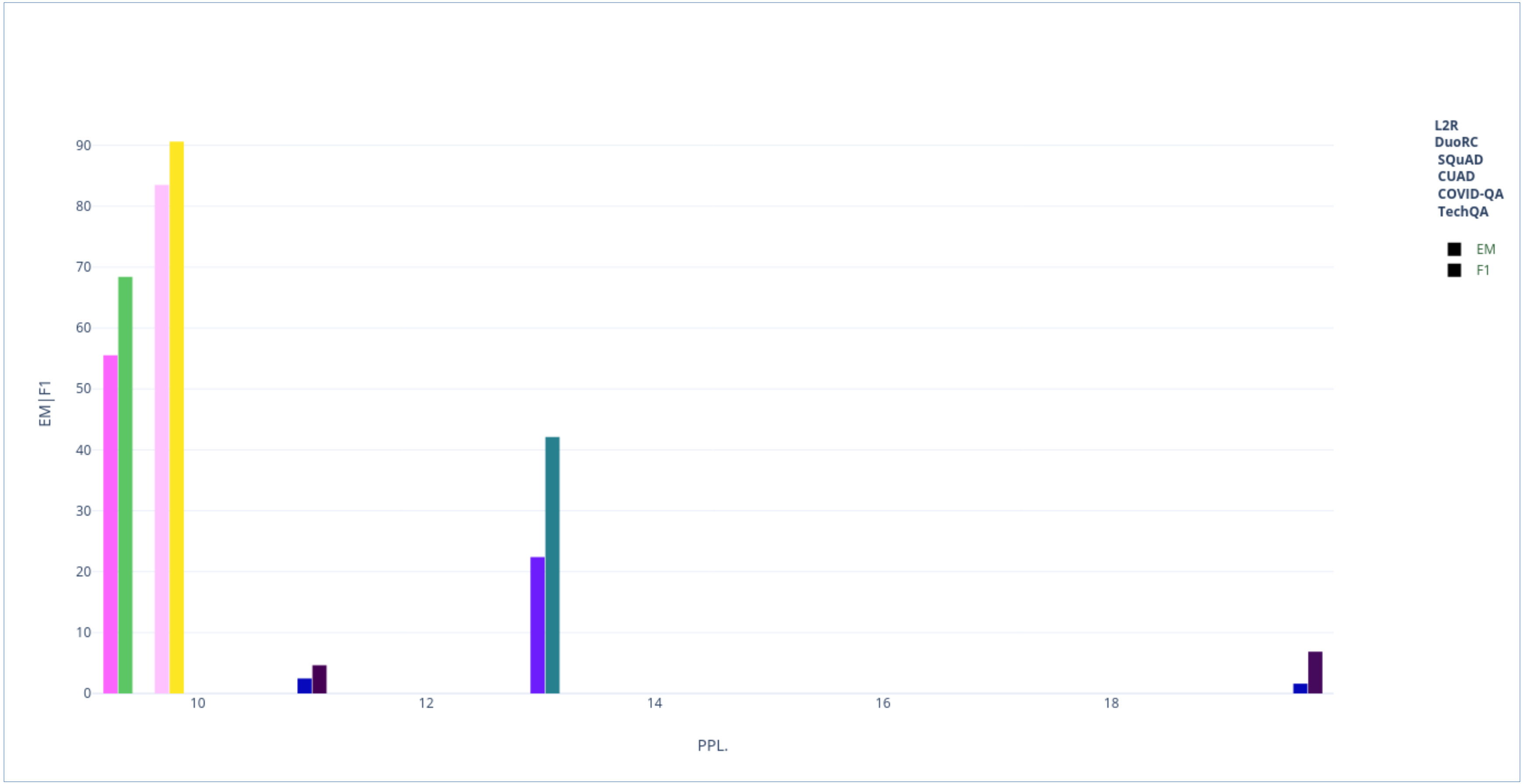
COVID-QA scores

L	A|H	EM|F1
12	12 | 768	**22.39 | 42.11**
24	16 | 1024	**22.14 | 38.52**

CUAD scores

L	A|H	EM|F1
12	12 | 768	**2.46 | 4.63**
24	16 | 1024	**0.78 | 3.56**

How Does PPL. Correlate with Performance



Next Steps

- Training **domain-specific embeddings** and seeing whether they improve performance.
- Motivation – Consider the entity, “HIV-1 infection”
  - BERT tokenizes it as [HIV] + [-] + [1] + [infection]. This leads to a sort of *muddied* representation of the entity.
  - If we could have an embedding for the entity as [HIV-1 infection] *it should, in theory, improve performance.*
- However, acc. to this paper, **all of BERT’s contextualized embeddings occupy a narrow cone in its embedding space**. Thus, adding new domain-specific embeddings shouldn’t help that much? - This is a point I’m trying to reconcile.