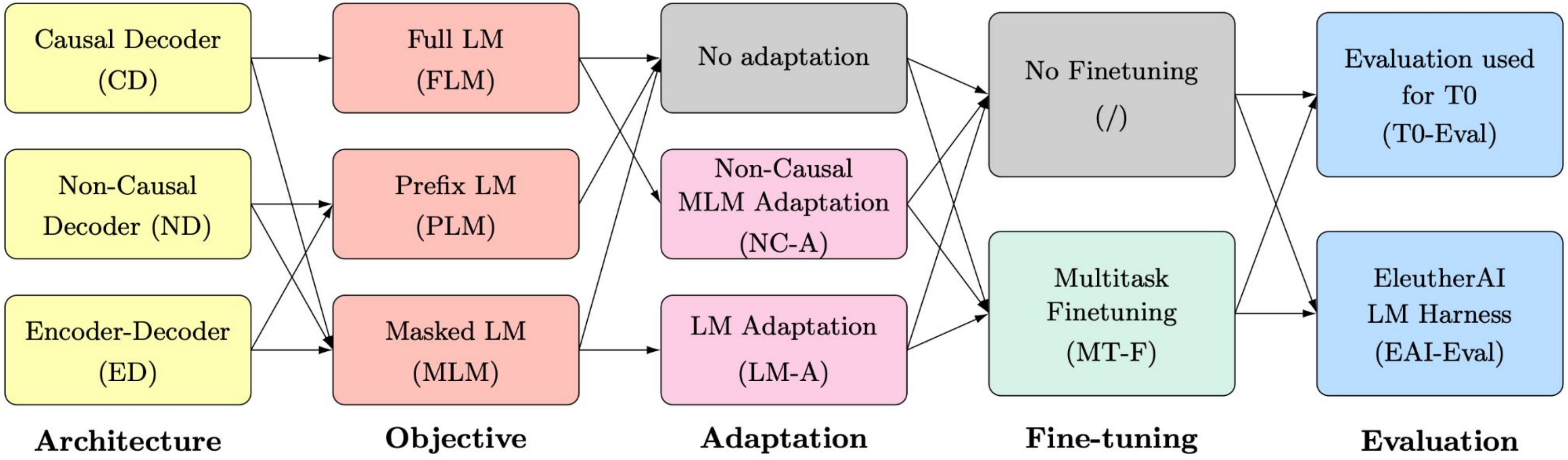
What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?

Thomas Wang[1*] Adam Roberts[2*] Daniel Hesslow[3]
Teven Le Scao[1] Hyung Won Chung[2] Iz Beltagy[4]
Julien Launay[3,5] Colin Raffel[1]

[1]Hugging Face [2]Google [3]Lighton
[4]Allen Institute for AI [5]LPENS, École Normale Supérieure

Systematic study of variants

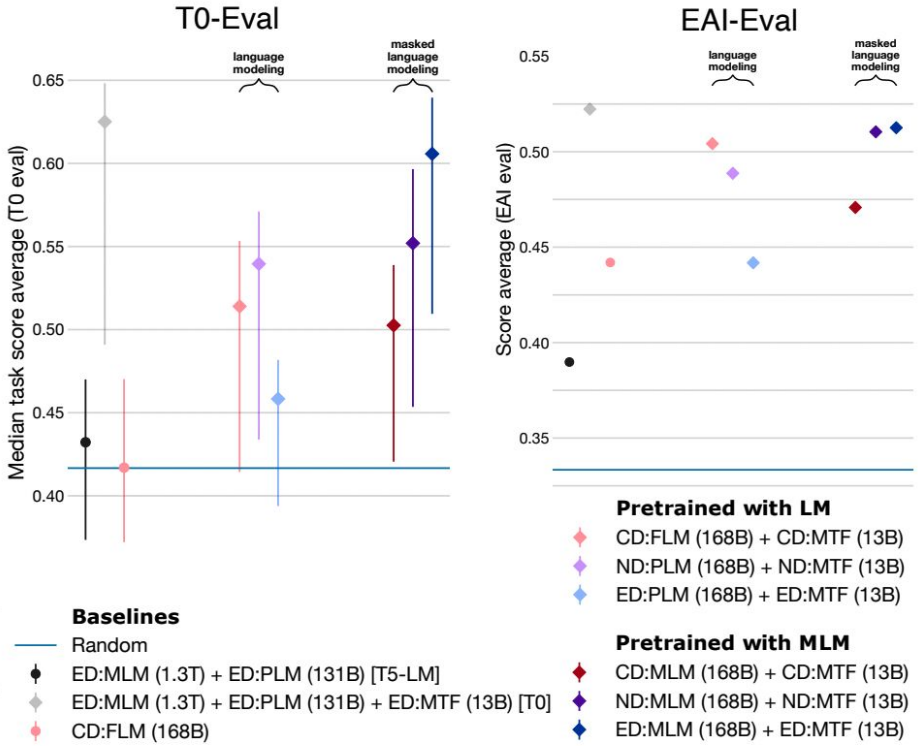


Experiments and results without MTF

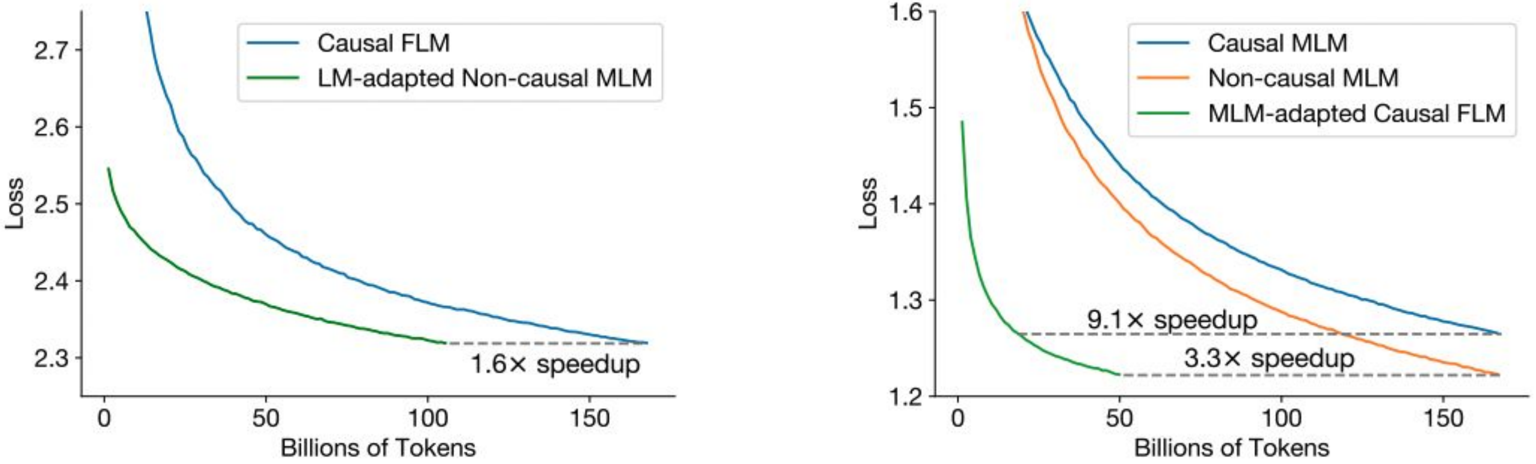
	EAI-EVAL	T0-EVAL
Causal decoder	**44.2**	**42.4**
Non-causal decoder	43.5	41.8
Encoder-decoder	39.9	41.7
Random baseline	32.9	41.7

After full or prefix language modeling pretraining, the causal decoder (FLM) exhibits the best zero-shot generalization abilities

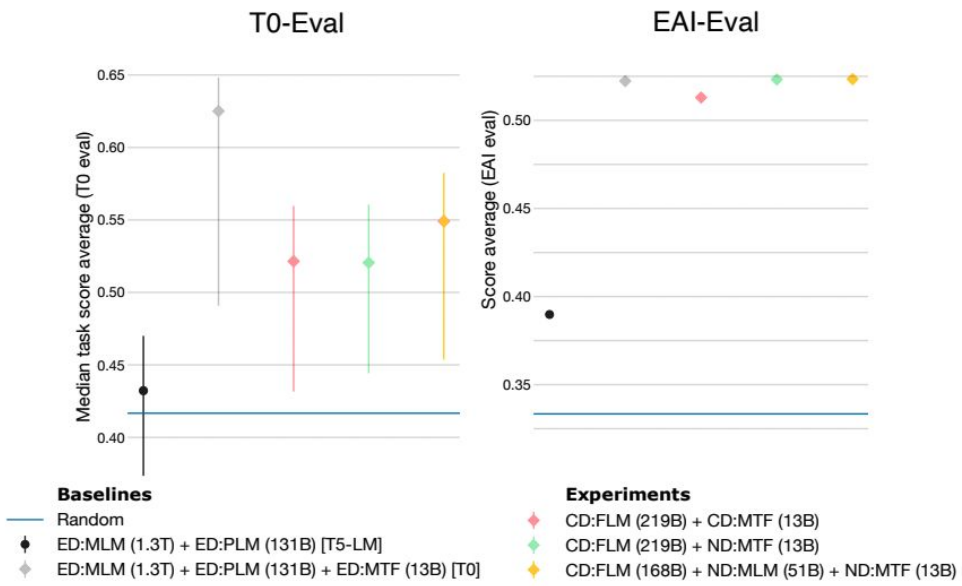
Experiments and results with MTF



Introducing adaptation



Introducing adaptation



Conclusion

- Without multitask finetuning, causal decoder pretrained with full language modeling performs best
- With multitask finetuning, encoder decoder pretrained with masked language modeling performs best
- We can convert a causal decoder model pretrained on full language modeling to a performant non causal decoder model by having a intermediary masked language modeling adaptation.

Acknowledgements

Google Research

This work was granted access to the HPC resources of *Institut du développement et des ressources en informatique scientifique (IDRIS) du Centre national de la recherche scientifique (CNRS)* under the allocation 2021-A0101012475 made by *Grand équipement national de calcul intensif (GENCI)*.

We thank the TPU Research Cloud team for providing us with generous access to TPUv4. We thank the TPUv4 Alpha team for providing technical support for this work.