



JOHNS HOPKINS

# Seq vs Seq: An Open Suite of Paired Encoders and Decoders



Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, Benjamin Van Durme

## Q: Which is better for a given task – an encoder or a decoder?

**Previous comparisons** were not fair (different architectures, training data, training recipes).

**We pre-trained paired encoder and decoder models from scratch to enable fair comparisons:**

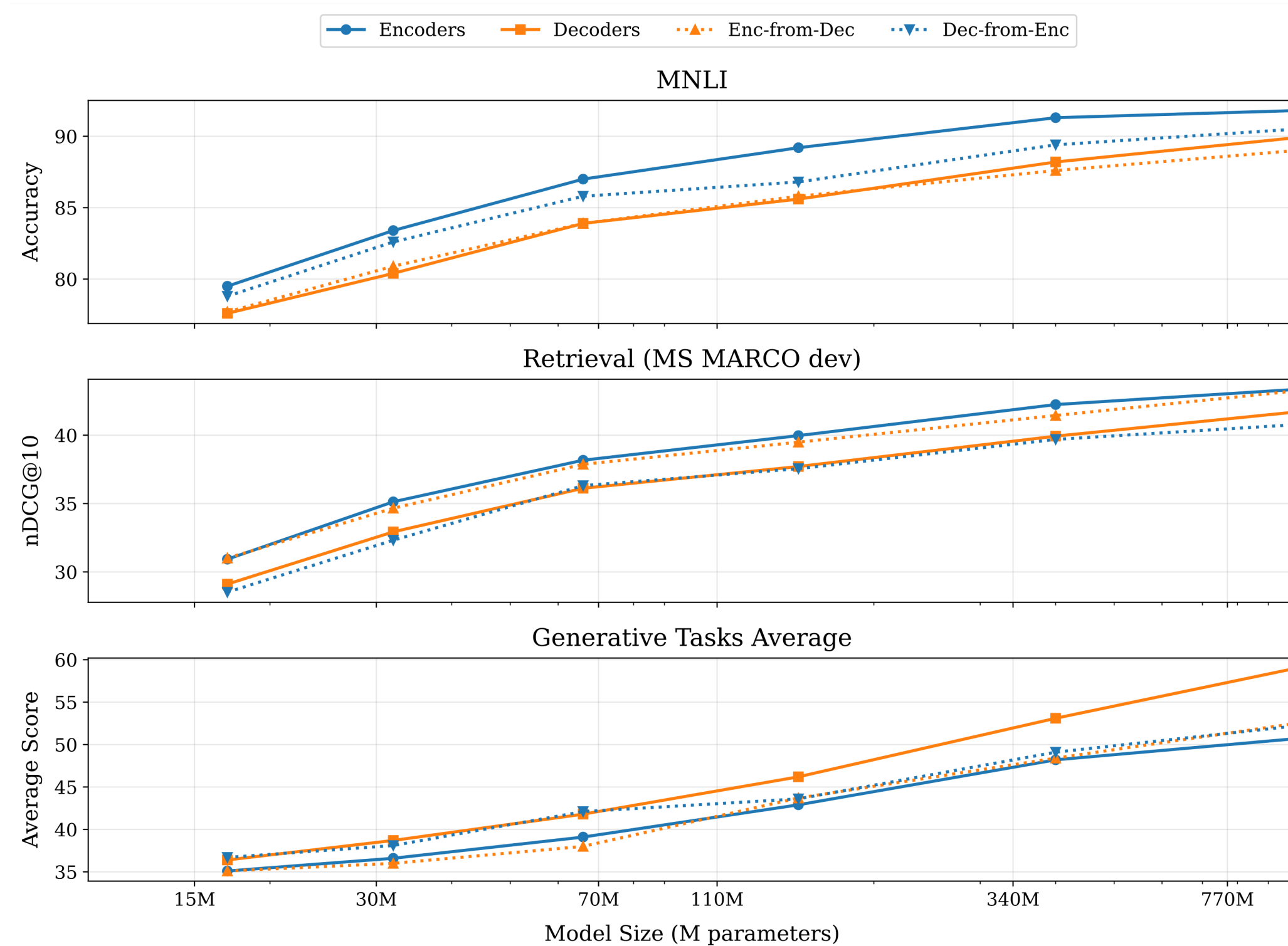
- 6 sizes: 17M, 32M, 68M, 150M, 350M, 1B, pre-trained on 2T tokens in three phases
- For their size, Encoders are SoTA overall, Decoders were SoTA for open-data models

## Q: How does continued pre-training with the reverse objective affect the tasks on which encoders and decoders perform best?

**Under our controlled settings**, we do continue pre-training:

- (a) Train decoders with MNTP to get **encoder-from-decoder** models (similar to LLM2Vec)
- (b) Train encoders on NTP to get **decoder-from-encoder** models

# Are encoders or decoders better?



- **Encoders** are better on **retrieval and classification** tasks.
- **Decoders** are better for **generation**.
- **Encoders** are comparable to an order-of-magnitude bigger decoder on classification: a 400M encoder beats a 1B decoder
- Encoders are **similar on retrieval**, to same-sized LLM2Vec-style **CPT encoders-from-decoders**.
- You can **compare the architectures on any task!** We show that encoders are more gender neutral on gender bias (*right ->*)



## Models & Data & Checkpoints on HF -> Compare the architectures on your task!

