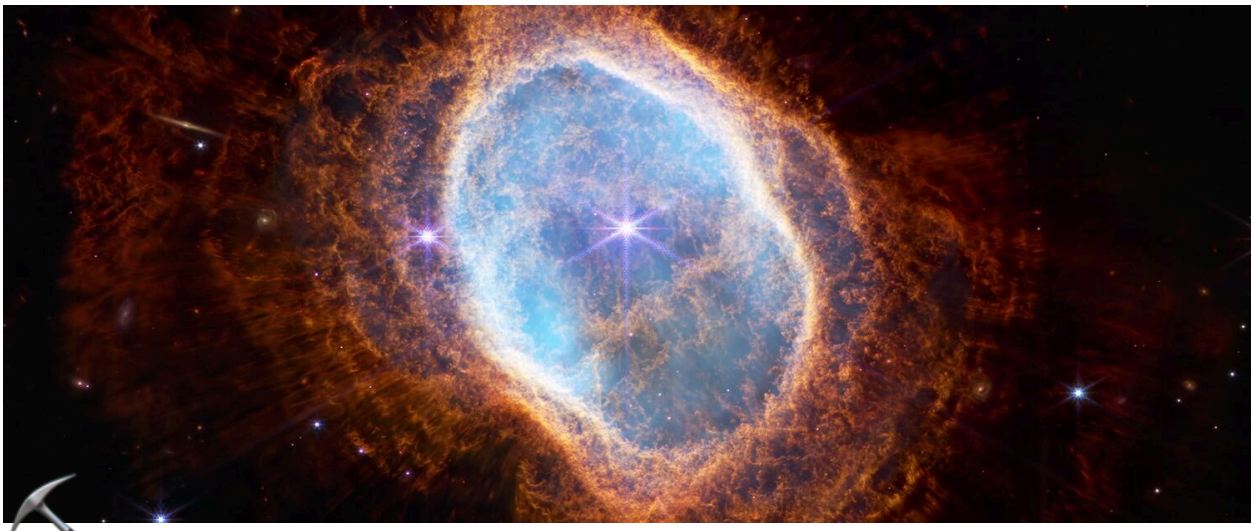

# Datacollator

DataCollatorWithPadding
Collator pads the inputs (input IDs, attention mask, and token type IDs)

data_collator = DataCollatorWithPadding(tokenizer=tokenizer)

DataCollatorForTokenClassification
Collator pads the labels the exact same way as the inputs

data_collator = DataCollatorForTokenClassification(tokenizer=tokenizer)

DataCollatorForSeq2Seq
Data collator will be responsible for preparing the decoder input IDs, which ar e shifted versions of the labels with a special token at the beginning. Since thi s shift is done
slightly differently for different architectures, the DataCollatorForSeq2Seq nee ds to know the model object

collator = DataCollatorForSeq2Seq(tokenizer=tokenizer, model=model)

DataCollatorForLanguageModeling:
Data collator just copies the inputs to create the labels.
Models: Causal Language Model, Mask Language Model

Causal Language Model
For gpt2, Shifting the inputs and labels to align them happens inside the mo
del.

collator = DataCollatorForLanguageModeling(tokenizer=tokenizer, mlm=Fal
se)

Mask Language Model
[MASK] token has been randomly inserted at various locations in our text b
y Data collator. These will be the tokens which our model will have to predict d
uring training.

data_collator = DataCollatorForWholeWordMask(tokenizer=tokenizer, mlm=
True, mlm_probability=0.15)

Note: Collator pads with -100, to make sure the padding tokens are ignored by the loss function