The central contribution of T5 is the text-to-text framework, which reformulates all NLP tasks as a problem where both input and output are text strings. Unlike traditional NLP models that might have task-specific architectures (e.g., classification tasks with softmax layers), T5 inputs text and produces text, no matter the task. This simplifies task formulation and enables the model to handle a diverse set of NLP tasks without modifying the architecture.

--> Scaling Up Transfer Learning

T5 investigates how transfer learning performance scales with larger models, datasets, and compute. The study follows the typical transfer learning approach:

Pretrain on a large, diverse corpus with an unsupervised task.
Fine-tune on a specific task with supervised examples.

However, T5 differs in two key ways:

Pretraining Task: Instead of typical masked language modeling (MLM) like BERT, T5 uses "span corruption" (a denoising objective).
Massive Scaling: T5 experiments with models ranging from 60 million to 11 billion parameters, showing that larger models generalize better.

--> Span Corruption as Pretraining Objective

Rather than BERT's token-level MLM, T5’s pretraining strategy randomly masks contiguous spans of text and replaces each with a unique placeholder token, then trains the model to generate the missing spans.

Why Span Corruption?
Captures Long-Range Dependencies: Unlike BERT, which only masks single tokens, T5 forces the model to learn broader sentence-level representations.
More Natural for Seq2Seq Learning: The model inherently learns how to generate text conditioned on context, making it more suited for downstream generative tasks like summarization or translation.
This forces T5 to learn strong bidirectional and autoregressive representations.

--> Model Architecture (Transformer-Based)

T5 adopts a standard Transformer encoder-decoder model, similar to sequence-to-sequence architectures used in neural machine translation. Key architectural details:

Fully Transformer-Based: Unlike BERT (encoder-only) or GPT (decoder-only), T5 is an encoder-decoder model.
Relative Positional Embeddings: T5 drops traditional absolute position embeddings in favor of relative ones.
LayerNorm Placement: Normalization is moved before attention layers instead of after.
Simplifications: T5 removes dropout during pretraining and uses simplified activation functions.
This design ensures that T5 can generate coherent sequences while leveraging the full bidirectional context from the encoder.

--> The C4 Dataset (Colossal Clean Crawled Corpus)

Instead of using commonly used corpora like Wikipedia or BookCorpus, the paper constructs a massive web-crawled dataset called C4 (Colossal Clean Crawled Corpus). C4 is derived from the Common Crawl dataset but aggressively filtered to remove noisy and low-quality content.

Size: 750GB of text, much larger than typical NLP pretraining datasets.
Diversity: Includes various domains, ensuring robust generalization.
This large, diverse dataset helps the model learn richer representations and transfer effectively across tasks.

--> Scaling Laws and Model Size

A key finding is that scaling up both model size and dataset size consistently improves performance.
Larger models benefit more from fine-tuning, even with a relatively small task dataset.
Compute efficiency matters—training bigger models on more data is better than training small models longer.
This aligns with later findings in deep learning that larger models generalize better given sufficient data.

Core Contributions
Unified Text-to-Text Framework → Simplifies multitask learning.
Span Corruption Pretraining → Enhances bidirectional understanding.
Massive Scaling with C4 Dataset → Improves generalization.
Transformer Encoder-Decoder Model → Works well for diverse NLP tasks.
Large-Scale Empirical Study → Provides insights into transfer learning scaling laws.