Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

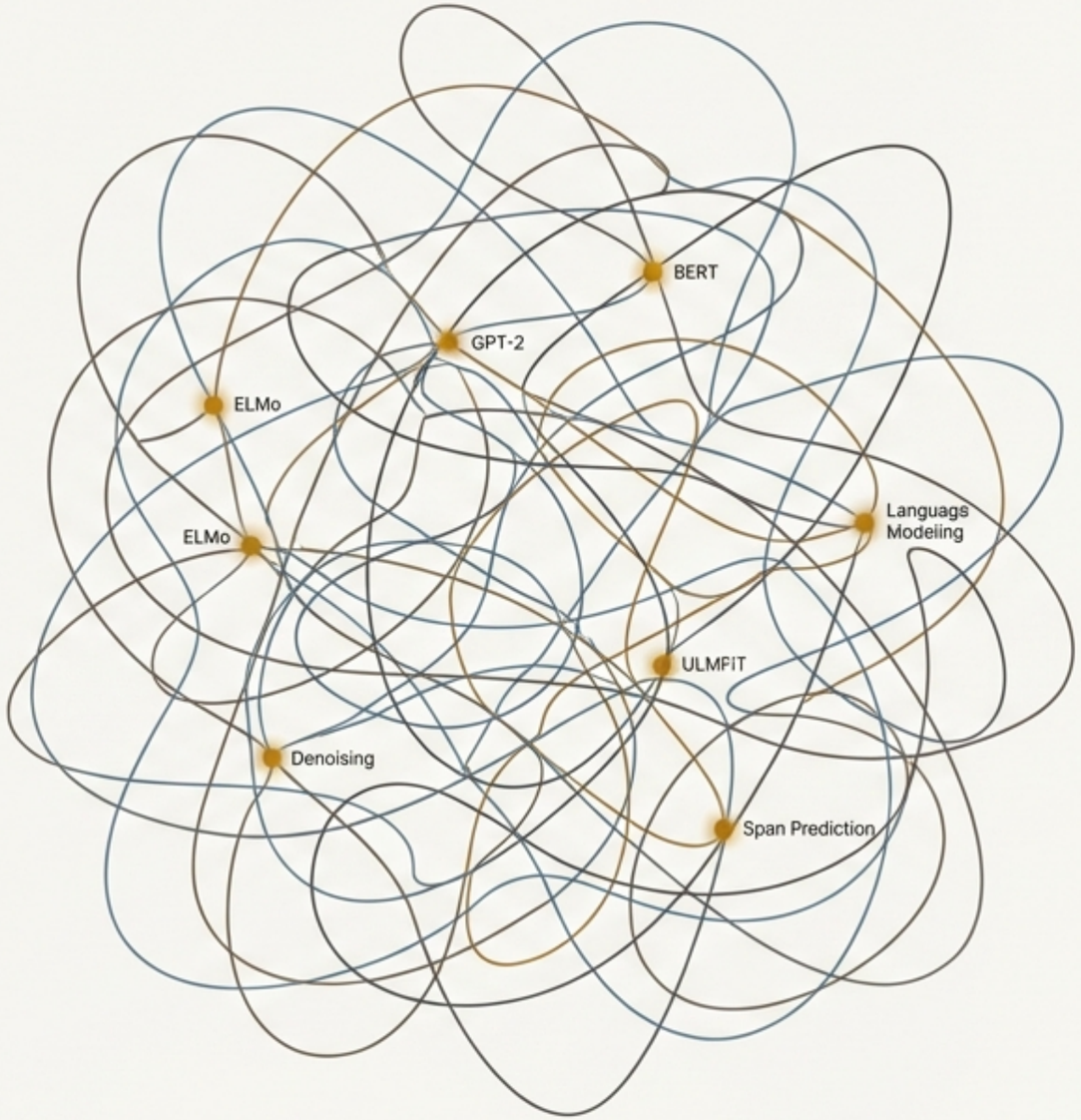
Colin Raffel, Noam Shazeer, Adam Roberts, et al.

Google

Journal of Machine Learning Research (2020)

The NLP transfer learning landscape is powerful, but fragmented and hard to navigate.

- Recent progress has created a wide diversity of pre-training objectives, architectures, datasets, and fine-tuning methods.
- This rapid progress “can make it difficult to compare different algorithms, tease apart the effects of new contributions, and understand the space of existing methods.”
- The core challenge: a lack of systematic understanding and a unified framework to measure what truly matters.



T5 proposes a unified framework by treating every NLP task as a text-to-text problem.

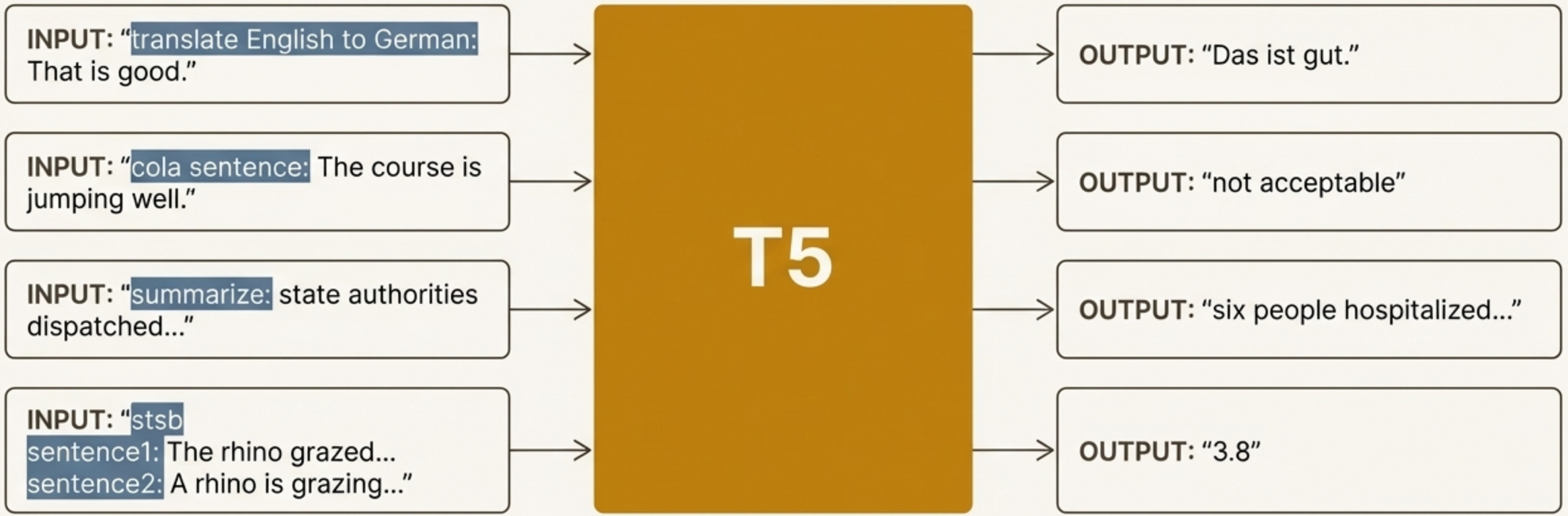
The core idea: convert all text-based language problems into a format of "taking text as input and producing new text as output."

This allows for the direct application of the same model, objective, training procedure, and decoding process to a diverse set of tasks.

This unification provides a standard testbed for a systematic, empirical survey of transfer learning techniques.

Translation, classification, summarization, and regression all become text generation.

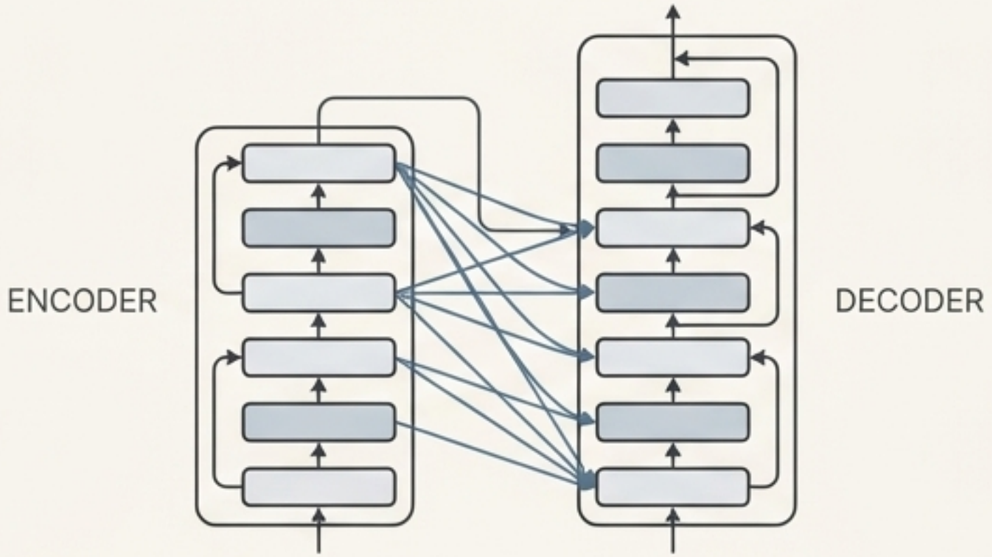
By adding a task-specific prefix to the input, we tell the model what to do.
The model's only job is to generate the correct target text.



The framework is powered by a standard Transformer and a new, massive text corpus.

Model (T5)

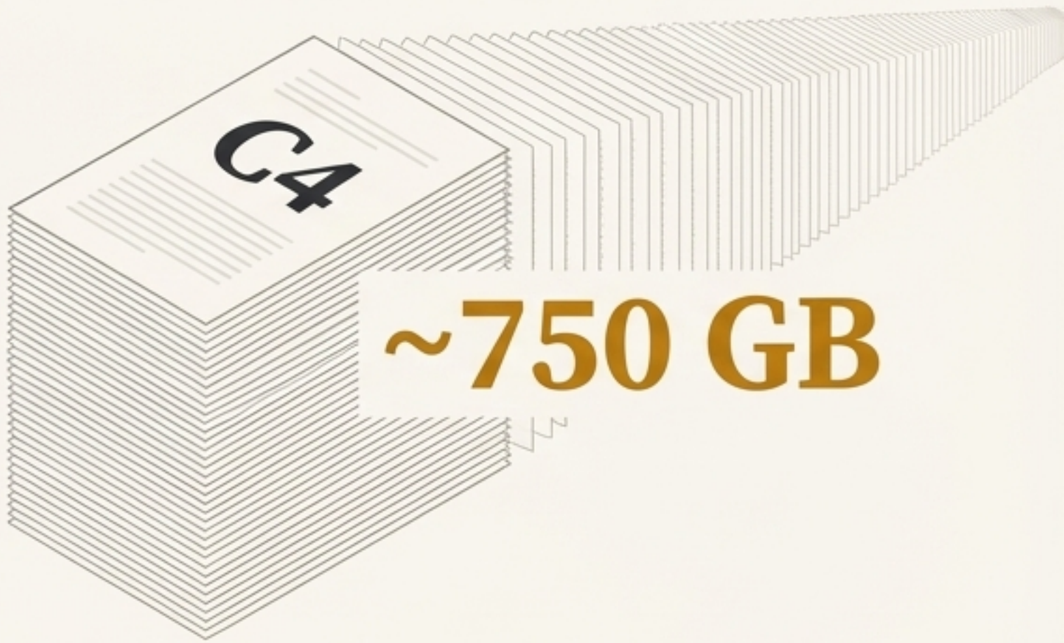
"Text-to-Text Transfer Transformer" based on the original encoder-decoder Transformer architecture. The baseline model has ~220 million parameters, with encoder/decoder stacks similar in size to BERT-BASE.



Encoder-Decoder Architecture

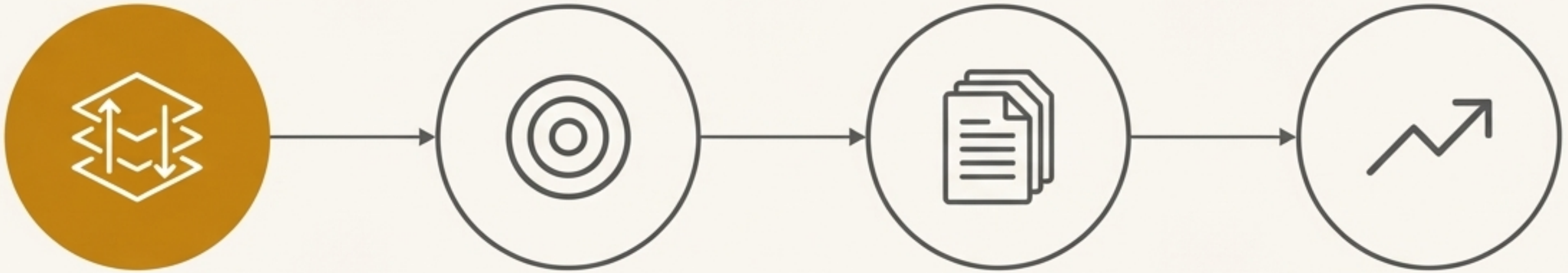
Dataset (C4)

The "Colossal Clean Crawled Corpus," a new dataset created for this work. Sourced from Common Crawl, it consists of ~750 GB of clean English text after extensive heuristic filtering.



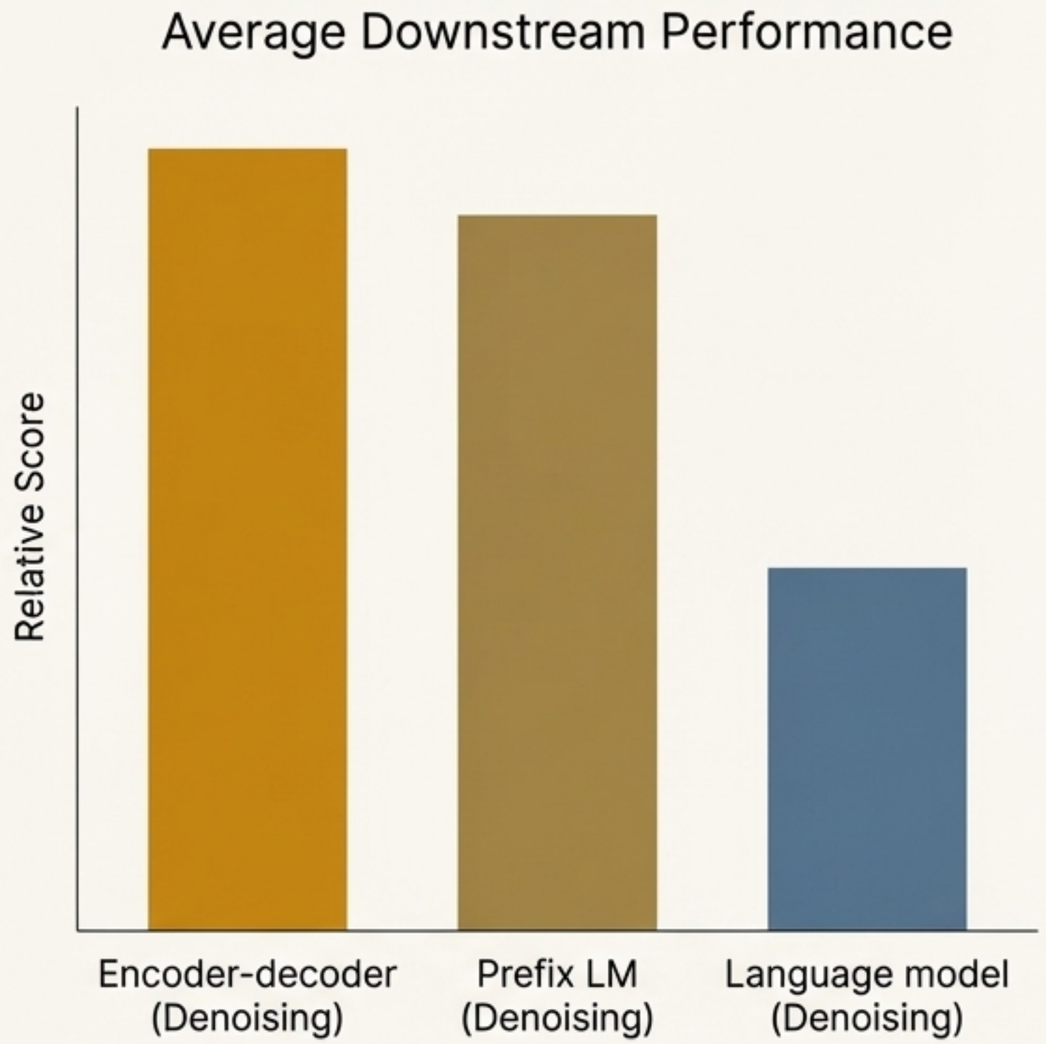
A systematic study was conducted to isolate the impact of key transfer learning variables.

The T5 framework allows for a “coordinate ascent” approach: start with a strong baseline and alter one aspect at a time. This allows us to empirically compare and understand the contribution of different factors.



The standard encoder-decoder architecture consistently outperforms decoder-only models.

- Different architectures were compared with equivalent computational cost or parameter counts.
- In all tasks, the encoder-decoder architecture with a denoising objective performed best.
- Sharing parameters across the encoder and decoder performed nearly as well, suggesting an effective way to reduce parameter count.



Denoising objectives are superior, with span corruption offering the best efficiency.

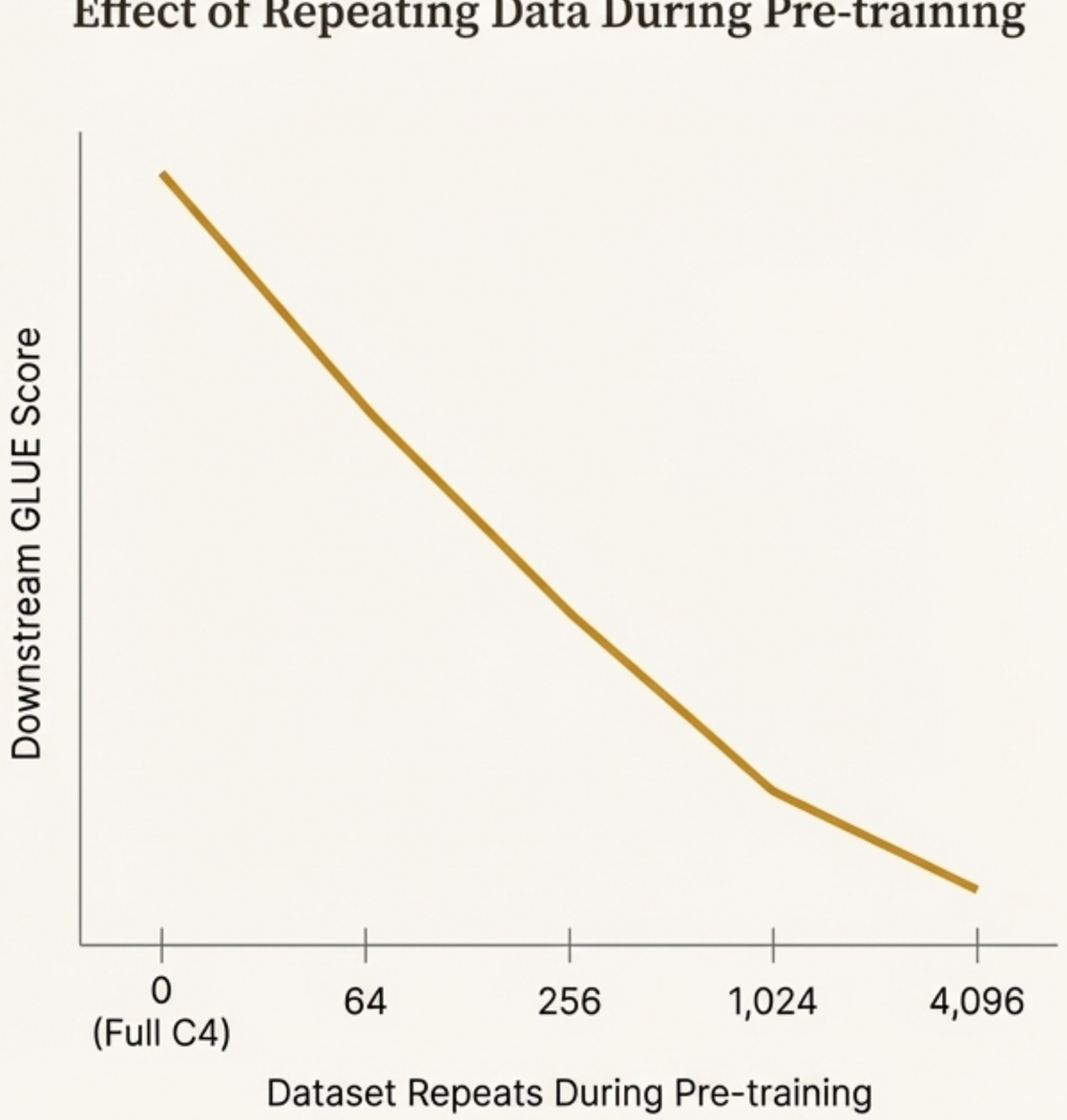
Finding 1: Denoising objectives (like BERT's) significantly outperform prefix language modeling and deshuffling.

Finding 2: Among denoising variants, performance is similar. The key differentiator is computational cost.

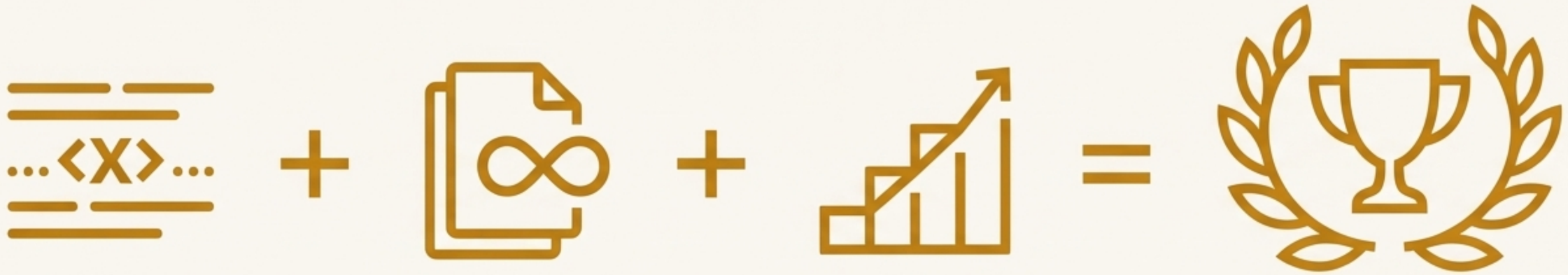
Conclusion: Corrupting contiguous spans of tokens is most effective. It produces marginally better performance and is more efficient due to shorter target sequences.

Large, diverse, and clean datasets are critical; pre-training on in-domain data can provide a targeted boost.

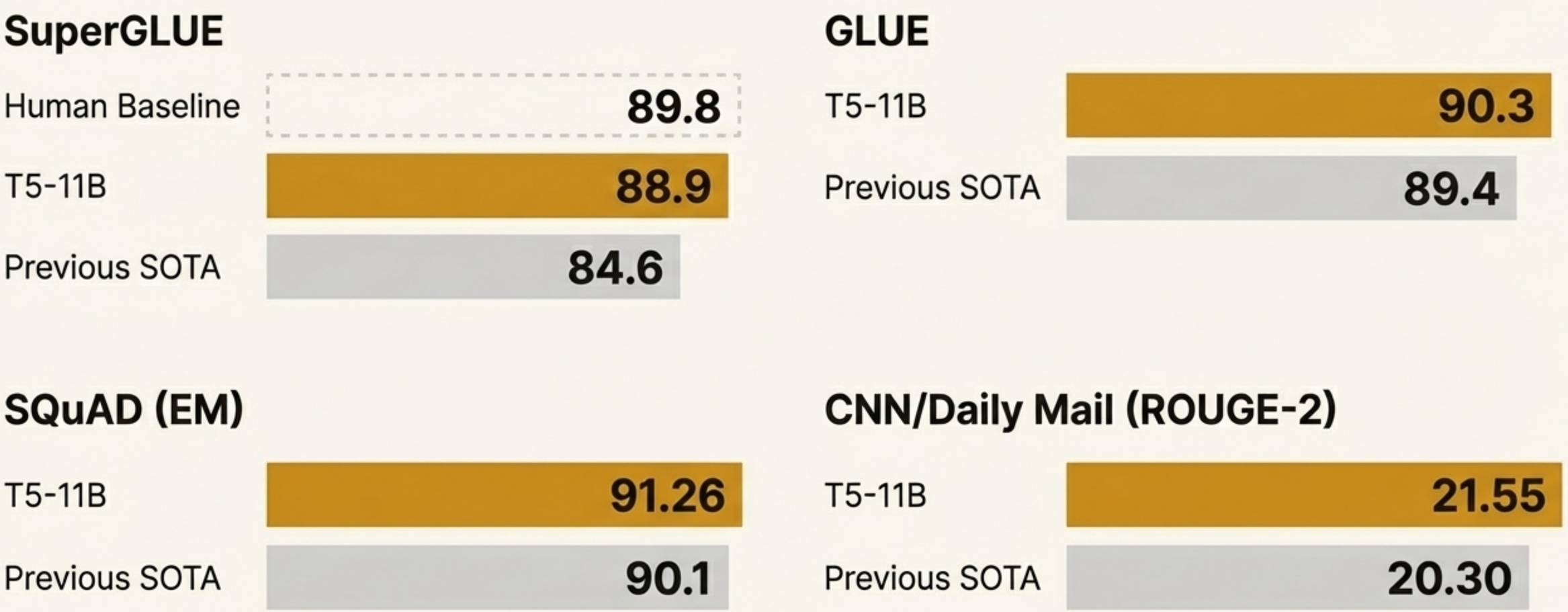
- **Cleanliness matters**: Pre-training on the heuristically-filtered C4 (750GB) uniformly outperforms the unfiltered Common Crawl data (6.1TB).
- **Domain matters**: Pre-training on domain-specific data can improve performance on specific downstream tasks.
- **Size matters**: Performance degrades significantly when pre-training on smaller datasets that must be repeated many times, suggesting the model begins to memorize the data.



Combining these insights with unprecedented scale pushes the state of the art.

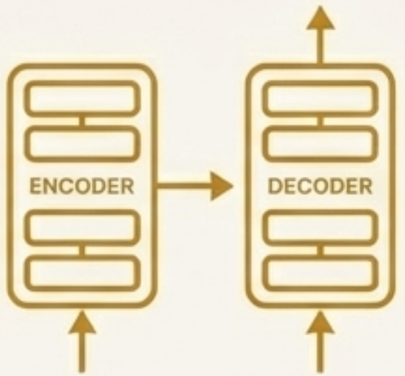


T5-11B achieves state-of-the-art on 18 benchmarks, nearly matching human performance on SuperGLUE.



Key lessons from a comprehensive exploration of transfer learning.

The Text-to-Text framework is a simple yet powerful approach that unifies diverse NLP tasks.



The original Encoder-Decoder Transformer remains a highly effective architecture for this framework.

Scale is a primary driver of performance, with larger models trained on more data consistently achieving better results.

Pre-training on large, diverse, and clean data (like C4) is essential for building general-purpose models.

Important limitations remain, pointing toward future research directions.

Limitations

- **The 'Inconvenience of Large Models'**: Powerful, but computationally expensive for fine-tuning and inference.
- **Translation Performance**: English-only pre-training was insufficient to match SOTA on WMT tasks.
- **Knowledge Extraction Efficiency**: Denoising objectives require massive data. More efficient methods are needed.

Future Directions

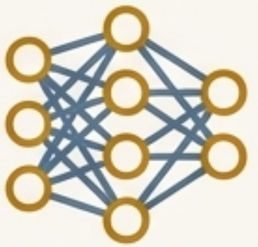
- Research into distillation and parameter sharing remains critical.
- Explore methods like back-translation and cross-lingual pre-training.
- Develop more efficient methods for teaching 'knowledge' to models.

T5 provides a new toolkit to facilitate future work on transfer learning for NLP.

To enable replication, extension, and application of these results, we are releasing our complete framework and assets.

Codebase

github.com/google-research/text-to-text-transfer-transformer



Pre-trained Models

C4 Dataset