Cornell Bowers C·IS
College of Computing
and Information Science

Investigating the Limitations of Transformers with Simple Arithmetic Tasks

Arnav Joshi, Eric Li, Shawn Chen, Caleb Woo, Allen Wu

Motivation

What can transformer models do?

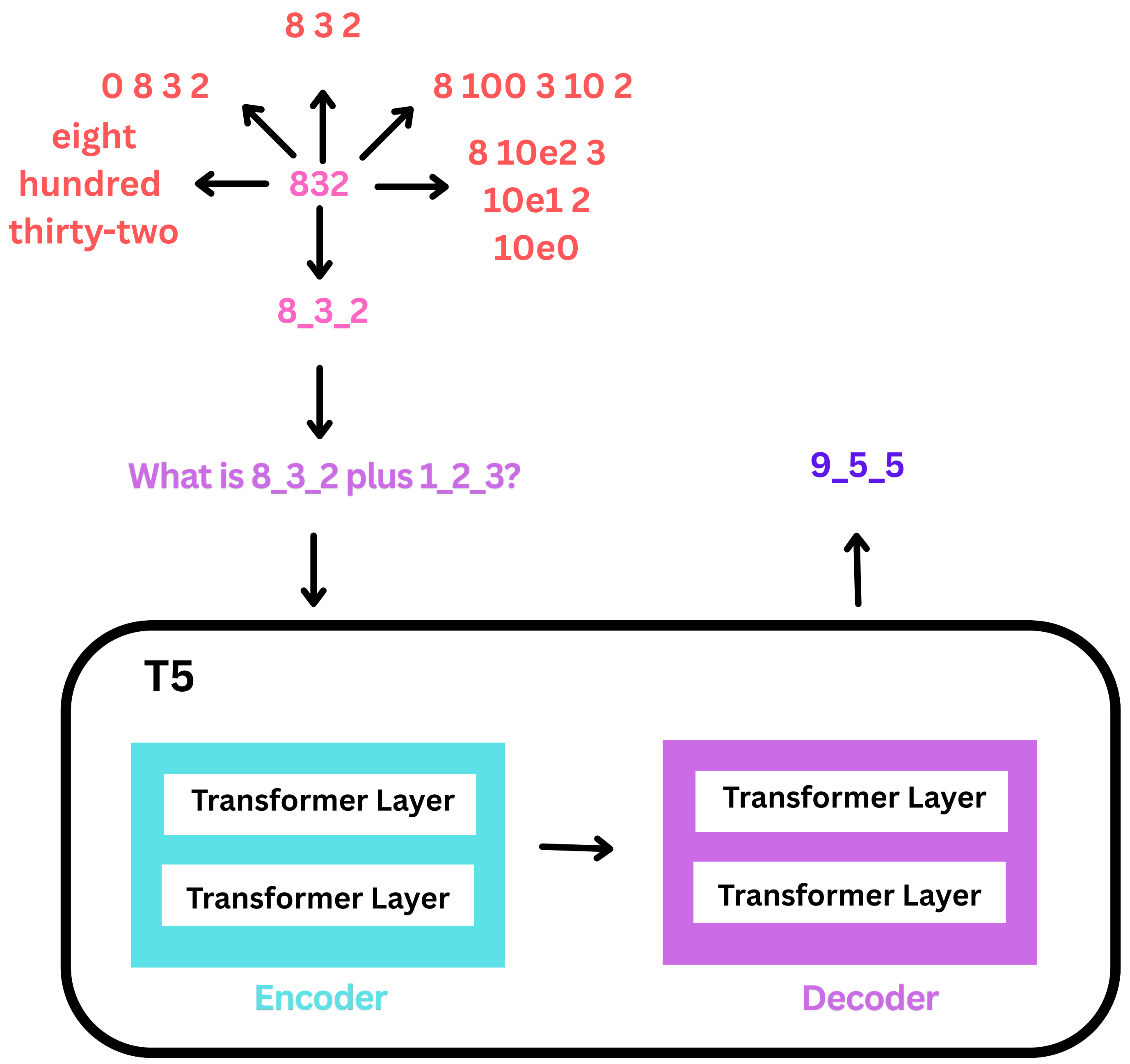
With the rapid advancement of transformer models in recent years, pretrained transformer models have proven to be able to perform complex tasks that borderlines reasoning, through extraction of key representations from textual data.

What about numbers?

Whether transformers can handle simple arithmetic is unclear, as most models are trained on textual patterns rather than numbers. If the model can extract representative information about their numerical values, bigger models with better pretraining objectives should lead to models capable of mathematical reasoning.

So, can Transformers do Math?

We aim to examine the performance of transformer models on simple arithmetic tasks (addition and subtraction). To solve the issue of long tokens and limited familiarity with numbers, special formatting is applied to each number to make them more model friendly.



Methodology

Goal:

- Test accuracy of LLMs on addition when using different numerical representations

Numerical Representations:

Orthography	Example	Notes
DECIMAL	832	default representation
CHARACTER	8 3 2	ensures consistent tokenization
FIXED-CHARACTER	0 8 3 2	ensures consistent positions (e.g., max. 4 digits)
UNDERSCORE	8_3_2	underscores provide hints on digit significance
WORDS	eight hundred thirty-two	leverages pretraining
10-BASED	8 100 3 10 2	easy to determine digit significance
10E-BASED	8 10e2 3 10e1 2 10e0	more compact encoding of above

Table 1. Different numerical representations

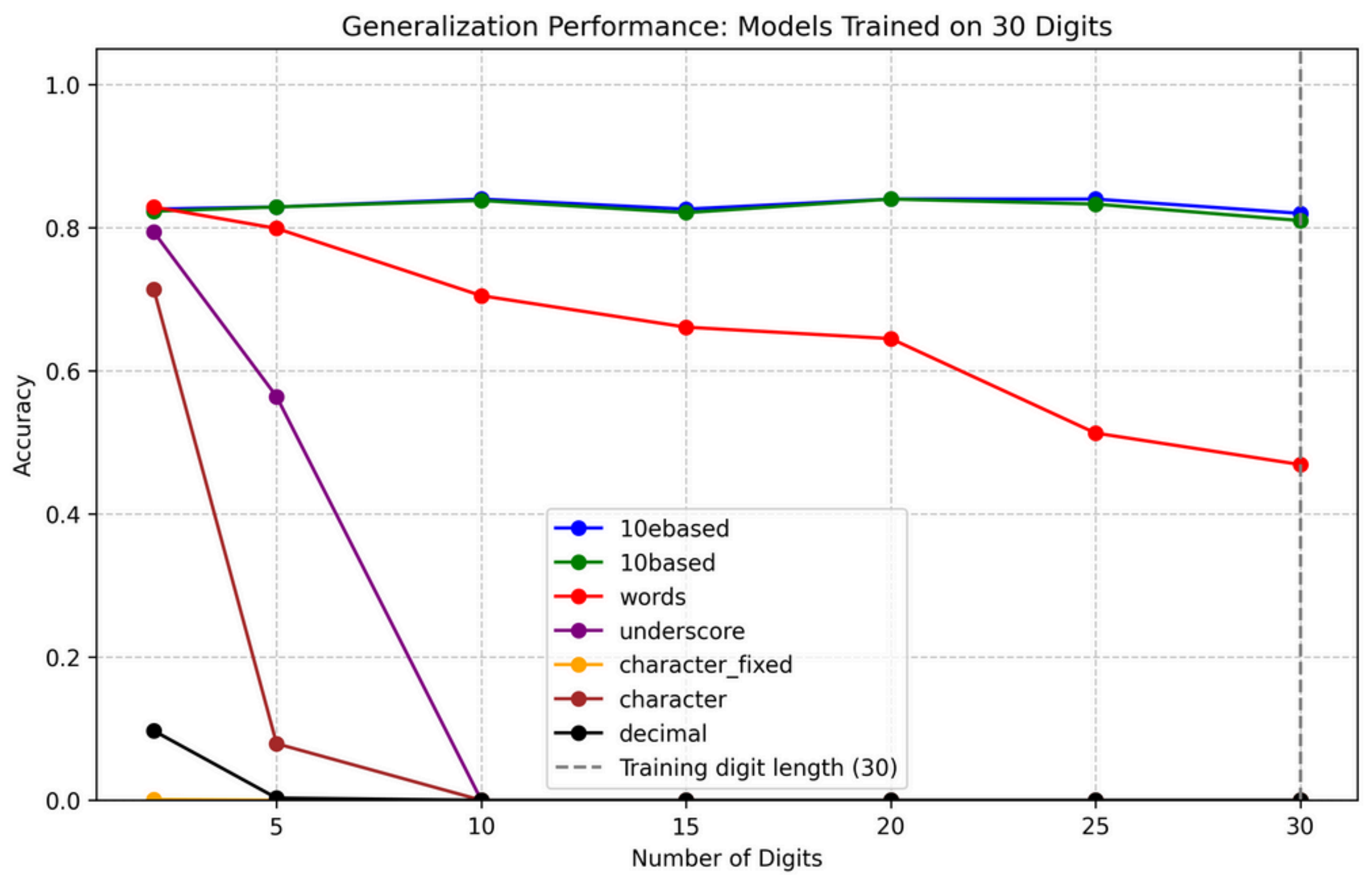
Balanced Sampling:

- Programmatically generate dataset of addition of 2-30 digit numbers
  1. sample d from interval [2, 30]
  2. sample [number1] and [number2] from interval $[10^{d-1}, 10^{d} - 1]$
  3. convert to numerical representation
  4. generate prompt “What is [number1] plus [number2]?” and the answer as label
- This sampling ensures an even mix of numbers with 2 to 30 digits

Approach:

- Fine-tune pretrained T5 models with 220M (base) parameters
- Train on 1000 examples for each representation over 100 epochs, using an additional 1000 as validation data. Select the highest performing on validation

Results



Experiment Results

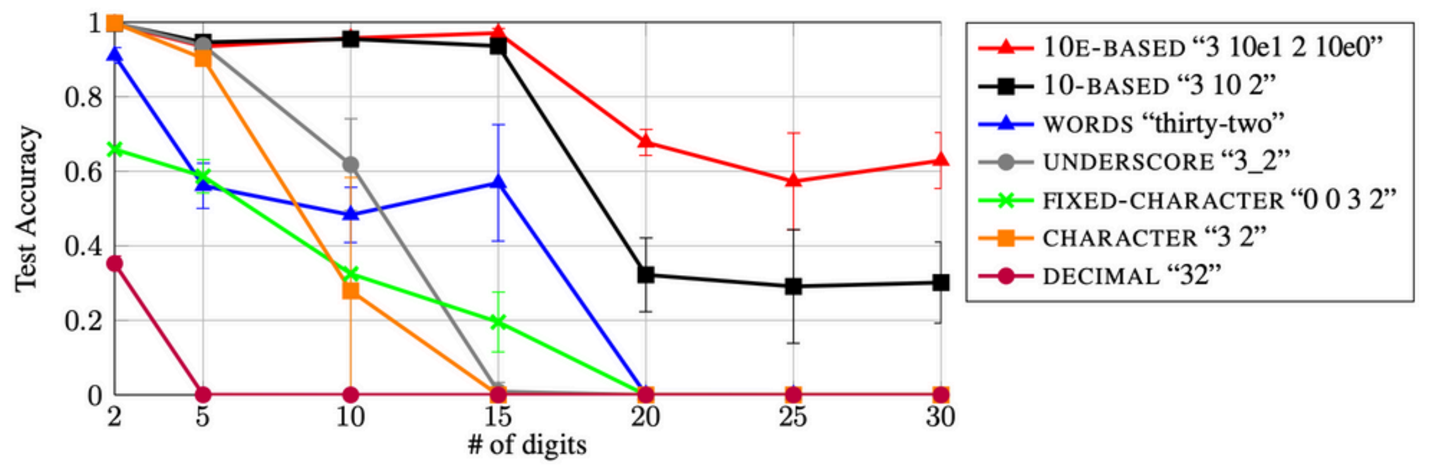


Figure 1: Accuracy of different number representations on the addition task.

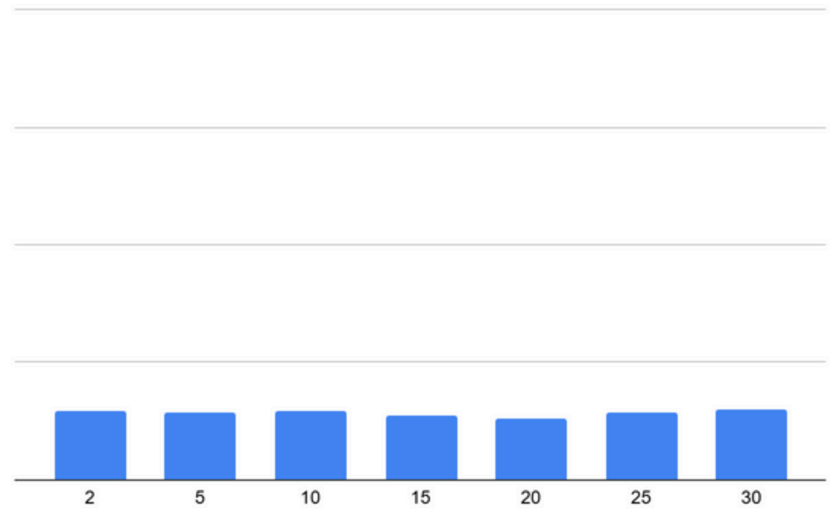
Original Paper Results

Conclusion

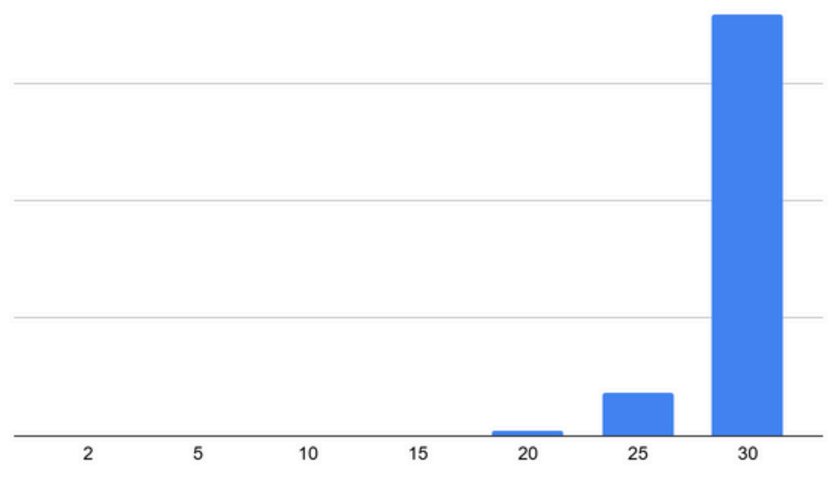
Transformers learn most accurately when introducing position tokens (e.g. “3 10e1 2”) with 10e based words having the highest test accuracy. However, model fails to extrapolate, as they fail to perform the arithmetic tasks when evaluated on inputs whose either length or format distribution differs from the one seen during training, as shown when alternating balanced and random sampling.

		Test	
		Balanced	Random
Train	Balanced	1.000	1.000
	Random	0.014	1.000

Table 2. Accuracy on 60-digit addition, with random and balanced sampling; table from original paper



Balanced Sampling



Random Sampling

Future Work

After validating the paper’s findings, we want to extend to modern transformers, i.e. Llama3, and to chart progress (or persisting gaps) in LLM reasoning. In addition, expanding the range of values to decimal numbers, (e.g “10.49 + 7.77”), quantity of inputs, different operations (multiplication, modular arithmetic, and nested parentheses), on both fine-tuned T5-base and newer models could reveal limits of reasoning.

References

Rodrigo F. Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. Investigating the Limitations of Transformers with Simple Arithmetic Tasks. In *Proceedings of the 1st Mathematical Reasoning in General Artificial Intelligence Workshop (MathAI@ICLR 2021)*. arXiv:2102.13019.