# Minimal Clues, Maximal Understanding: Solving Linguistic Puzzles with RNNs, Transformers, and LLMs

**Ishan Khare***, **Anavi Baddepudi***, and **Emma Wang***

**{iskhare, anavib, emmwang}@stanford.edu*

CS 224N | Stanford University

Stanford Computer Science

## Problem and Introduction

- Critical gap exists in ability of deep-learning models to mimic human-like reasoning and understanding
- We uncover limits by solving *Rosetta Stone* linguistic puzzles with minimal bidirectional translations
- Experiments: i) **RNNs**, ii) **fine-tuned transformer based models**, and iii) **GPT-4 in-context learning**
- Our methods surpass baselines (with GPT-4 being the best) but lack human-like reasoning

## Background and Datasets

- Most NLP research predominantly concentrates on merely 0.2% of known languages (20 out of 7,000)
- Puzzles from **PuzzLing Machines**[1] dataset are solvable without knowledge of chosen diverse languages: Swahili, Georgian, Norwegian, Malay
- Explored two datasets: "large" 10,000-sentence **TedTalk corpus**[2] and "small" linguistic puzzles
- Dataset did not include puzzle solutions so we performed data curation for training and test sets

| Chickasaw | English |
|---|---|
| Ofi'at kowi'ã lhiyohli. | The dog chases the cat. |
| Kowi'at ofi'ã lhiyohli. | The cat chases the dog. |
| Ofi'at shoha. | The dog stinks. |
| Ihooat hattakã hollo. | The woman loves the man. |
| Lhiyohlili. | I chase her/him. |
| Salhiyohli. | She/he chases me. |
| Hilha. | She/he dances. |

**Translate the following into Chickasaw:**

| | |
|---|---|
| ? | *The man loves the woman.* |
| ? | *The cat stinks.* |
| ? | *I love her/him.* |

**Translate the following into English:**

| | |
|---|---|
| *Ihooat sahollo.* | ? |
| *Ofi'at hilha.* | ? |
| *Kowi'ã lhiyohlili.* | ? |

*Rosetta Stone* linguistic puzzle

## Methods

- **Baselines**: Implemented *Random Words* and *FastAlign* methods
- **Recurrent Neural Network (RNN)***: Sequence-to-sequence network with LSTM encoder and decoder plus attention
- **Transformer-Based Models***: Pre-trained NMT models with self-attention (6 layers each with 8 attention heads)
- **LLM In-Context Learning:** Provided solved puzzles and prompted GPT-4 for translation without language-specific knowledge

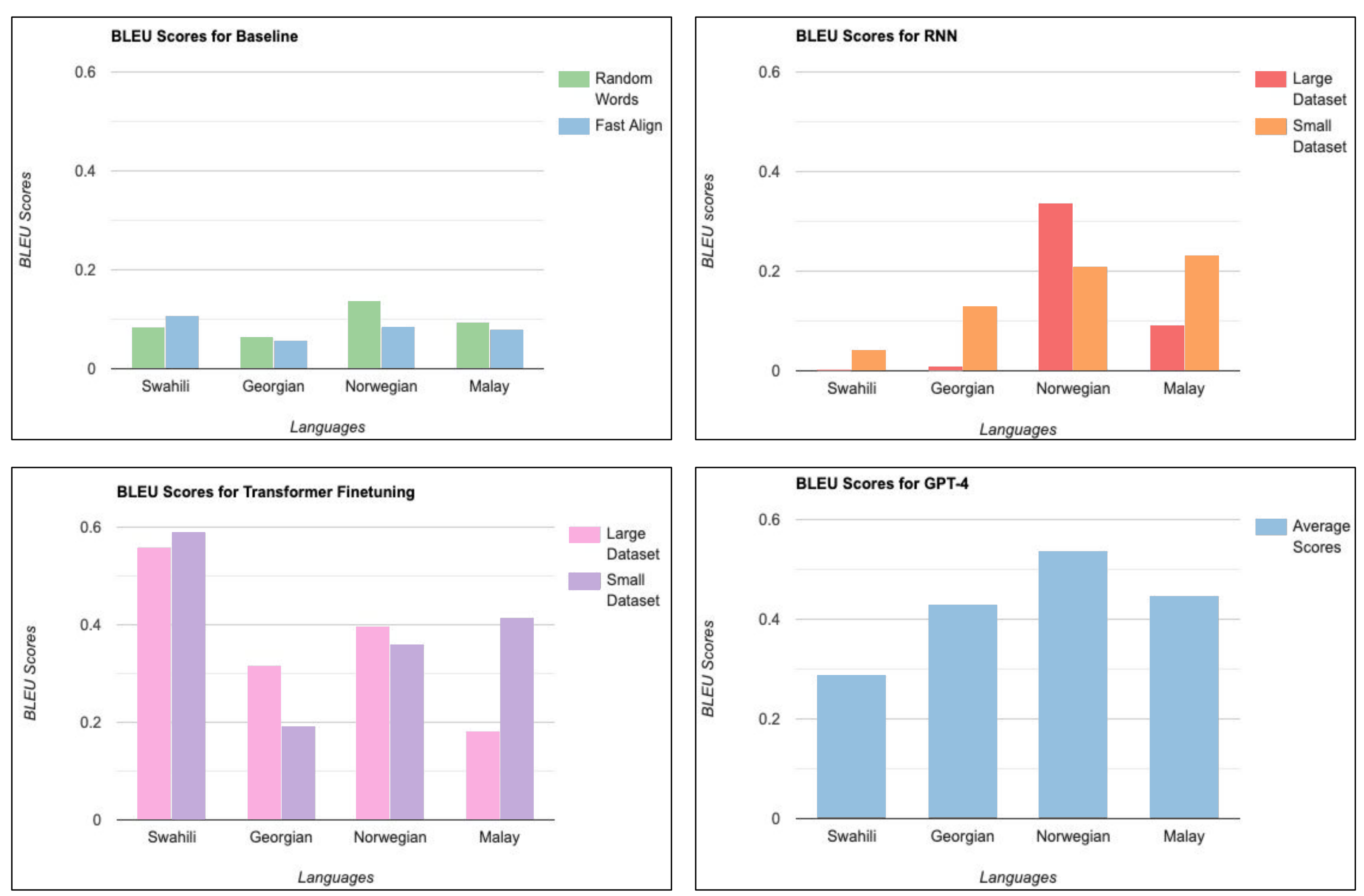
** fine-tuned on either TEDTalk external corpus or small set of solved linguistic puzzles*

## Experiments and Results

- Four RNN and four fine-tuned transformer models per language
- Trained for 16.5 hours on NVIDIA A100 GPUs with 80GB RAM

Averaged results by model type

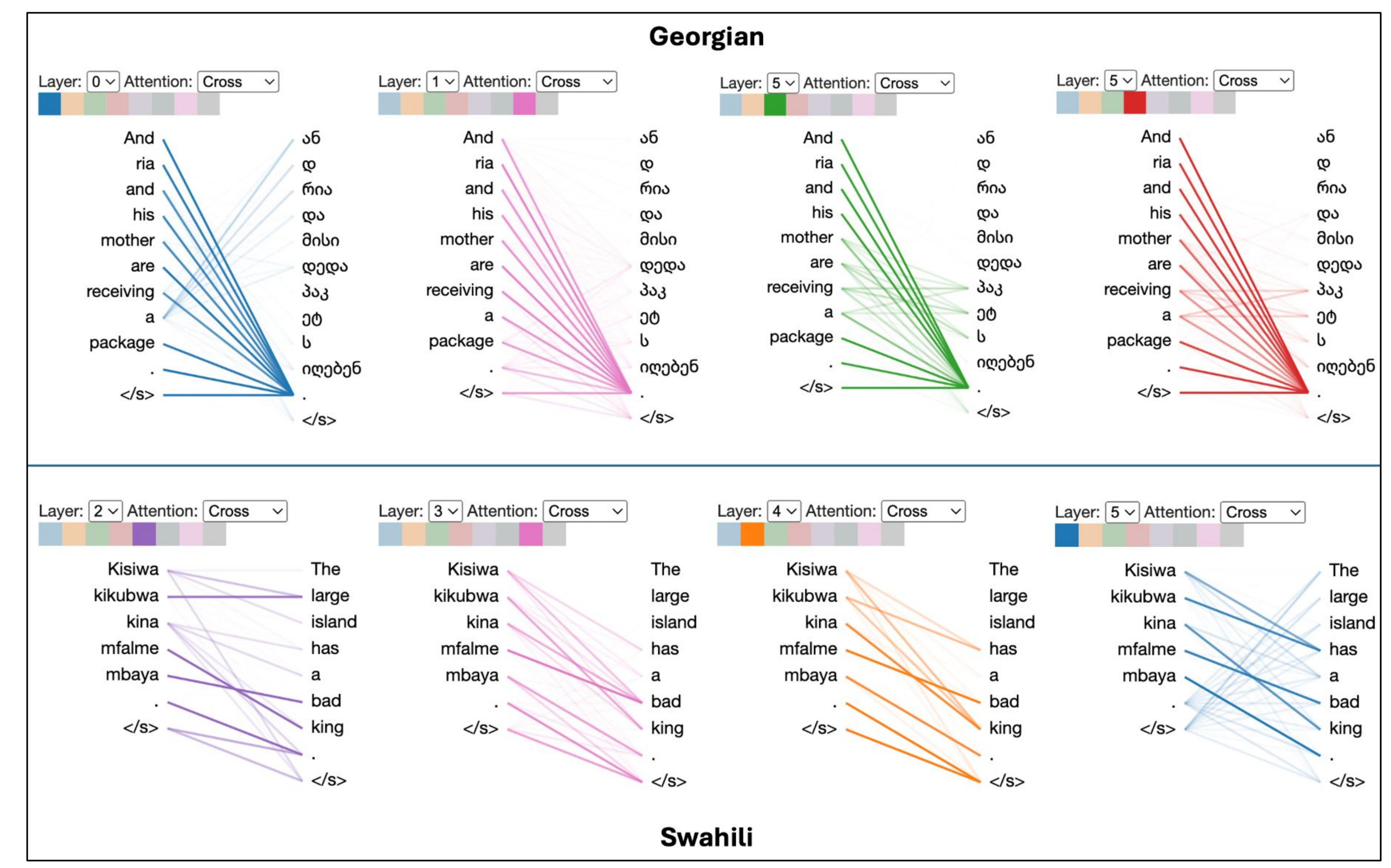
| Model | BLEU | CHRF | CTER | EM |
|---|---|---|---|---|
| Random Words | 0.083 | 0.271 | 0.188 | 0.000 |
| Fast Align | 0.096 | 0.295 | 0.237 | 0.000 |
| RNN | 0.132 | 0.295 | 0.059 | 0.000 |
| Transformer Finetuning | 0.387 | 1.710 | 11.142 | 0.000 |
| GPT-4 | 0.420 | 0.793 | 0.243 | 0.000 |




BLEU score comparisons by language across model types

## Analysis

- RNN and GPT-4 perform best on Norwegian plausibly due to close proximity to English (both Germanic languages with Latin alphabet)
- Performance varies with language *i.e.*, Georgian's subject-object-verb grammar vs. subject-verb-object for English
  - Shows importance of model adaptability to infer linguistic patterns
- Most Georgian model attention heads focus on "." punctuation mark
- Swahili model attention shows strong links between correct translation pairs such as "mfaime"—"king" and "mbaya"—"bad"




Attention for Georgian and Swahili fine-tuned transformers

## Conclusions and Future Work

- Best performances achieved by transformer fine-tuning and GPT-4, with BLEU scores of 0.387 and 0.420, respectively
- **Research limitation**: variability in puzzle difficulty across different languages. We suggest standardization using human benchmarks
- **Future work**: propose meta-learning techniques to surpass the limitations of current deep-learning models and LLMs

## References

1) Sahin *et. al*. 2020. PuzzLing Machines: A Challenge on Learning From Small Data.
2) https://opus.nlpl.eu/NeuLab-TedTalks/corpus/version/NeuLab-TedTalks.