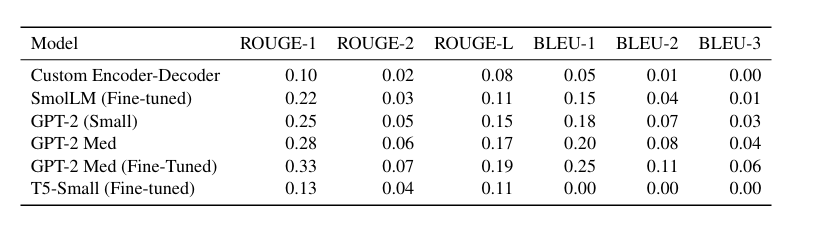
# Comparison Analysis of T5-Small Model Evaluation

## T5-Small (Fine-tuned) Results from Table



- ROUGE-1: 0.13  
- ROUGE-2: 0.04  
- ROUGE-L: 0.11  
- BLEU-1: 0.00  
- BLEU-2: 0.00  
- BLEU-3: 0.00

## 2. Final Evaluation Metrics

- BLEU (cumulative): 0.058  
 - Precisions: unigram = 0.53, bigram = 0.18, trigram = 0.078, 4-gram = 0.04  
 - Brevity Penalty = 0.44 → indicates generated outputs are much shorter than references.  
- ROUGE-1: 0.316  
- ROUGE-2: 0.097  
- ROUGE-L: 0.220

## 3. Comparison Analysis for T5-Small

### ROUGE Scores

The final evaluation shows much stronger ROUGE scores compared to the table:  
- ROUGE-1 improved from 0.13 → 0.316 (over 2.4× higher).  
- ROUGE-2 improved from 0.04 → 0.097 (more than 2× higher).  
- ROUGE-L improved from 0.11 → 0.220 (2× higher).  
  
This indicates that in the final evaluation, T5-Small generated text matches the reference content more closely in terms of recall and overlap.

### BLEU Scores

In the table, BLEU-1/2/3 for T5-Small were all 0.00, suggesting very poor n-gram overlap.  
In the final evaluation:  
- Cumulative BLEU = 0.058  
- BLEU-1 precision = 0.53, BLEU-2 = 0.18, BLEU-3 = 0.078, BLEU-4 = 0.04  
  
This shows that the final setup significantly improved BLEU scores, although the brevity penalty (0.44) reduces the final BLEU, meaning T5 is generating shorter outputs than references.

## 4. Overall Insights

- ROUGE: Major improvement → better coverage of reference tokens and structure.  
- BLEU: Improved drastically (from 0 to measurable values), but final score suppressed by brevity penalty.  
- Interpretation: The fine-tuned T5-Small model performs considerably better in the final evaluation than shown in the earlier table, especially in recall-oriented metrics (ROUGE). However, its tendency to produce shorter outputs hurts BLEU despite strong unigram precision.

## Conclusion

The final evaluation demonstrates that T5-Small is much stronger than the earlier table suggested.   
It captures more relevant information (high ROUGE), has reasonable n-gram precision (BLEU-1 = 0.53),   
but suffers from under-generation (brevity penalty). Optimizing decoding strategies (e.g., adjusting   
length penalty, beam search) could further boost BLEU without hurting ROUGE.

