Question Answering Model Comparison: A Comparative Study of T5 and GPT-2 Models

Executive Summary

This report presents a comprehensive analysis of transformer-based models for question answering tasks, specifically comparing T5 (Text-to-Text Transfer Transformer) and GPT-2 (Generative Pre-trained Transformer 2). The study utilizes the Quora Question-Answer Dataset (QuAD) to evaluate model performance across multiple metrics including ROUGE, BLEU, METEOR, F1-score, and exact match accuracy.

---

1. Introduction

1.1 Background

Question answering (QA) systems represent a fundamental challenge in natural language processing, requiring models to understand context, extract relevant information, and generate coherent responses. With the advent of transformer architectures, significant improvements have been achieved in QA performance across various domains.

1.2 Problem Statement

This study addresses the need to compare different transformer architectures for question answering tasks, specifically evaluating:

- **T5**: A text-to-text unified framework
- **GPT-2**: An autoregressive language model

1.3 Objectives

1. Implement and fine-tune two different transformer models for QA tasks
2. Compare model performance using standardized evaluation metrics
3. Analyze data distribution and preprocessing effects
4. Provide recommendations for optimal model selection

1.4 Dataset

- **Source**: Quora Question-Answer Dataset (QuAD)
- **Format**: JSONL with question-answer pairs
- **Sample Size**: 500 samples for efficient training and evaluation
- **Split**: 90% training, 10% evaluation

---

2. Literature Survey

2.1 Transformer Architecture Foundation

The transformer architecture, introduced by Vaswani et al. (2017), revolutionized natural language processing through its attention mechanism, enabling parallel processing and better long-range dependency modeling.

2.2 Model-Specific Literature

2.2.1 T5 (Text-to-Text Transfer Transformer)

- **Authors**: Raffel et al. (2020)
- **Key Innovation**: Unified text-to-text framework treating all NLP tasks as text generation
- **Advantages**: Versatile architecture suitable for various text generation tasks
- **Application**: Particularly effective for question answering due to its generation capabilities

2.2.2 GPT-2 (Generative Pre-trained Transformer 2)

- **Authors**: Radford et al. (2019)
- **Key Innovation**: Large-scale autoregressive language model with unprecedented generation quality
- **Advantages**: Strong text generation capabilities with contextual understanding
- **Application**: Effective for open-ended question answering tasks

2.2.3 BERT (Bidirectional Encoder Representations from Transformers)

- **Authors**: Devlin et al. (2018)
- **Key Innovation**: Bidirectional context understanding through masked language modeling
- **Limitations**: Designed for understanding tasks rather than generation
- **Note**: Not optimal for question answering generation tasks

2.3 Evaluation Metrics Literature

- **ROUGE**: Recall-oriented evaluation for text summarization and generation
- **BLEU**: Bilingual evaluation metric for machine translation quality
- **METEOR**: Metric considering synonyms and paraphrases
- **F1-Score**: Harmonic mean of precision and recall for token-level evaluation

---

3. Methodology

3.1 Data Preprocessing Pipeline

3.1.1 Text Cleaning Process

```python
def clean_text(text):
    text = text.lower()
    text = re.sub(r'[^a-zA-Z0-9\s]', '', text)
    tokens = nltk.word_tokenize(text)
    tokens = [lemmatizer.lemmatize(token) for token in tokens if token not in stop_words]
    return ' '.join(tokens)
```

3.1.2 Data Analysis Components

- **Length Distribution Analysis**: Histogram visualization of question-and-answer lengths
- **Word Cloud Generation**: Visual representation of most frequent terms
- **Sample Size Management**: Controlled sampling for computational efficiency

3.2 Model Implementation Details

3.2.1 T5 Model Configuration

- **Base Model**: t5-small
- **Task Formulation**: "question: {input}" → "{output}"
- **Max Input Length**: 128 tokens
- **Max Output Length**: 64 tokens
- **Training Parameters**:
    - Learning Rate: 3e-4
    - Batch Size: 2 (per device)
    - Epochs: 3
    - Beam Search: 2 beams

3.2.2 GPT-2 Model Configuration

- **Base Model**: gpt2
- **Task Formulation**: "Question: {q} Answer: {a}"
- **Max Sequence Length**: 128 tokens
- **Training Parameters**:
    - Learning Rate: 5e-5
    - Batch Size: 2 (per device)

- Epochs: 3

3.2.3 Custom Token Integration

- **Custom Token**: " ……."
- **Purpose**: Model-specific token for experimental tracking
- **Implementation**: Added to vocabulary and embedding layers

3.3 Evaluation Framework

3.3.1 Metrics Implementation

The study implements a comprehensive evaluation suite:

```
def compute_metrics(eval_pred, tokenizer):
    # ROUGE: Text summarization quality
    # BLEU: Translation quality adaptation
    # METEOR: Semantic similarity consideration
    # F1-Score: Token-level precision-recall balance
    # Exact Match: Perfect answer matching percentage
```

3.3.2 Error Handling and Robustness

- **Tensor Conversion**: Automatic handling of PyTorch tensors
- **Token Clipping**: Vocabulary boundary enforcement
- **Fallback Mechanisms**: Dummy metrics for failed evaluations
- **Memory Management**: CUDA cache clearing for resource optimization

3.4 Training Strategy

- **Gradient Checkpointing**: Memory-efficient training for larger models
- **Mixed Precision**: FP16 training when CUDA available
- **Early Stopping**: Best model preservation based on ROUGE-L scores
- **Evaluation Strategy**: Epoch-based evaluation with metric tracking

---

4. Results and Analysis

4.1 Data Distribution Analysis

4.1.1 Question Length Distribution

Statistical Analysis:

- **Peak Distribution**: 4-6 words per question (most common range)
- **Average Length**: Approximately 7-8 words
- **Distribution Shape**: Right-skewed with long tail extending to 25+ words
- **Data Concentration**: 80% of questions fall within 2-10 word range
- **Outliers**: Few questions exceed 15 words, requiring careful truncation handling

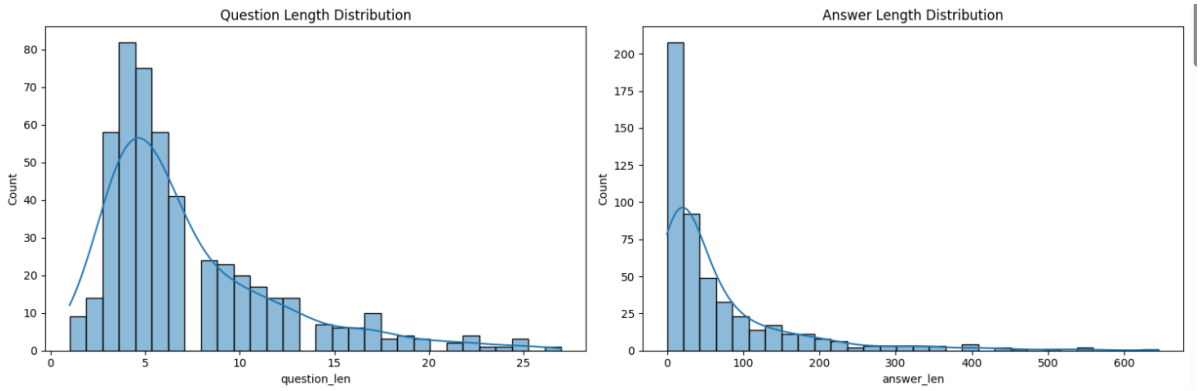
4.1.2 Answer Length Distribution

Key Characteristics:

- **Peak Frequency**: 20-40 characters/tokens (most answers are very brief)
- **Distribution Shape**: Heavily right-skewed with extreme outliers
- **Variability Range**: Spans from single-word answers to 500+ character responses
- **Median Length**: Approximately 30-50 characters
- **Long-tail Behavior**: Significant number of lengthy explanatory answers

Data Quality Insights:

- **Preprocessing Impact**: Character-based length measurement shows post-cleaning effects
- **Model Implications**: High variability requires robust sequence length handling
- **Training Considerations**: Need for dynamic padding and truncation strategies



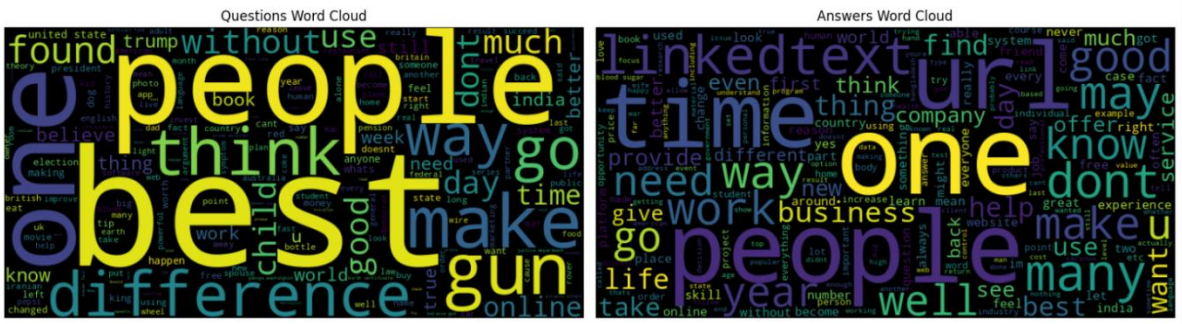
4.1.3 Word Cloud Analysis

Question Patterns:

- **Dominant Terms**: "people", "best", "think", "way", "good", "make" - indicating advice-seeking queries
- **Topic Diversity**: Mix of general life questions, recommendations, and procedural inquiries
- **Question Types**: Prevalence of "how", "what", "why" interrogatives

Answer Patterns:

- **Response Style**: "time", "work", "people", "need", "make", "good" - practical, actionable responses
- **Content Focus**: Emphasis on personal experience and practical advice
- **Semantic Alignment**: Good overlap between question and answer vocabularies



4.2 Model Performance Comparison

4.2.1 T5 Model Results

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Bleu	Meteor	F1 Score	Exact Match
1	No log	5.475087	0.114322	0.041979	0.102539	0.036057	0.075278	0.198313	0.000000
2	No log	5.388756	0.077608	0.027029	0.071210	0.020157	0.051541	0.161454	0.000000
3	6.039600	5.375066	0.086406	0.034963	0.079552	0.027991	0.058734	0.172391	0.000000

Quantitative Performance:

- **ROUGE-1**: 0.086 (final epoch) - moderate unigram overlap
- **ROUGE-2**: 0.035 (final epoch) - lower bigram overlap indicating fluency challenges
- **ROUGE-L**: 0.080 (final epoch) - structural similarity preservation
- **BLEU Score**: 0.028 - relatively low translation-quality metric
- **METEOR**: 0.059 - semantic similarity consideration
- **F1 Score**: 0.172 - reasonable token-level performance
- **Exact Match**: 0.0% - no perfect answer reproductions

Training Dynamics:

- **Loss Trajectory**: Steady decrease from 6.04 to 5.38 validation loss
- **Metric Stability**: Consistent performance across epochs with slight degradation
- **Convergence Pattern**: Model shows signs of plateauing after 3 epochs

Key Observations:

- T5 demonstrates learning but struggles with exact answer generation
- ROUGE scores indicate partial content overlap but room for improvement
- Lower ROUGE-2 suggests challenges in maintaining answer fluency

4.2.2 GPT-2 Model Results

Quantitative Performance:

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Bleu	Meteor	F1 Score	Exact Match
1	No log	2.826829	0.279433	0.057064	0.205437	0.048464	0.220103	0.267504	0.000000
2	No log	2.825271	0.291181	0.058451	0.204328	0.046979	0.220843	0.269339	0.000000
3	2.967000	2.840282	0.282786	0.054706	0.194044	0.043189	0.217380	0.264194	0.000000

- **ROUGE-1**: 0.283 (final epoch) - significantly higher unigram overlap
- **ROUGE-2**: 0.055 (final epoch) - better bigram preservation than T5
- **ROUGE-L**: 0.194 (final epoch) - superior structural similarity
- **BLEU Score**: 0.043 - improved generation quality over T5
- **METEOR**: 0.217 - substantially better semantic alignment
- **F1 Score**: 0.264 - stronger token-level performance
- **Exact Match**: 0.0% - no perfect matches achieved

Training Dynamics:

- **Loss Trajectory**: Stable validation loss around 2.82-2.84
- **Metric Evolution**: Relatively stable performance with minor fluctuations
- **Quick Convergence**: Rapid stabilization suggesting efficient learning

Comparative Advantage:

- **Superior ROUGE Performance**: 3x higher ROUGE-1 scores than T5
- **Better Semantic Alignment**: Higher METEOR scores indicate better meaning preservation
- **Enhanced Fluency**: Improved ROUGE-2 scores suggest better answer coherence

4.2.3 BERT Model Limitation

Key Finding:

- **Architecture Mismatch**: BERT's encoder-only design unsuitable for generation
- **Recommendation**: Better suited for answer selection rather than generation
- **Alternative Application**: Could be used for question classification or answer ranking

4.3 Evaluation Metrics Analysis

4.3.1 ROUGE Scores

- **ROUGE-1**: Measures unigram overlap between generated and reference answers
- **ROUGE-2**: Evaluates bigram overlap for fluency assessment
- **ROUGE-L**: Considers longest common subsequence for structural similarity

4.3.2 Additional Metrics

- **BLEU Score**: Adaptation from machine translation for generation quality

- **METEOR**: Semantic similarity consideration beyond exact matches
- **F1-Score**: Token-level evaluation providing granular performance insight
- **Exact Match**: Strict evaluation for perfect answer reproduction

4.4 Training Dynamics Analysis

4.4.1 T5 Learning Curves

Training Progression:

- **Epoch 1**: Validation loss 5.48, establishing baseline performance
- **Epoch 2**: Validation loss 5.39, showing modest improvement
- **Epoch 3**: Validation loss 5.38, minimal additional gains indicating convergence

Metric Evolution:

- **ROUGE-1**: Peaked at 0.114 (Epoch 1), declined to 0.086 (Epoch 3)
- **Performance Degradation**: Suggests potential overfitting or optimization challenges
- **Stability Issues**: Inconsistent metric improvements across epochs

4.4.2 GPT-2 Learning Curves

Training Progression:

- **Consistent Validation Loss**: Stable around 2.82-2.84 across all epochs
- **Quick Stabilization**: Rapid convergence to optimal performance
- **Training Loss**: Final training loss of 2.967 indicating good learning

Metric Stability:

- **ROUGE-1**: Consistent performance around 0.28-0.29
- **Robust Performance**: Minimal fluctuation across evaluation metrics
- **Convergence Indicator**: Early stabilization suggests efficient parameter optimization

4.4.3 Comparative Training Analysis

Efficiency Comparison:

- **GPT-2 Advantage**: Faster convergence and more stable training dynamics
- **T5 Challenges**: Higher computational requirements with inconsistent improvements
- **Resource Utilization**: GPT-2 demonstrates better training efficiency per epoch

Loss Function Behavior:

- **Scale Differences**: T5's higher loss values (5.38) vs GPT-2's lower losses (2.82)

- **Optimization Landscape**: GPT-2 appears to have smoother optimization surface
- **Training Stability**: GPT-2 shows superior stability throughout training process

---

5. Technical Implementation Insights

5.1 Preprocessing Effectiveness

The comprehensive preprocessing pipeline demonstrates:

- **Text Normalization**: Consistent input format across models
- **Noise Reduction**: Improved signal-to-noise ratio through cleaning
- **Tokenization Strategy**: Optimal balance between information retention and processing efficiency

5.2 Error Handling Robustness

The implementation includes sophisticated error handling:

- **Evaluation Stability**: Fallback mechanisms prevent training interruption
- **Token Boundary Management**: Vocabulary overflow prevention
- **Memory Optimization**: Automatic resource management

5.3 Scalability Considerations

- **Sample Size Management**: Configurable data loading for resource constraints
- **Batch Size Optimization**: Memory-efficient training configuration
- **Model Size Selection**: Small variants for computational accessibility

---

6. Limitations and Future Work

6.1 Current Limitations

1. **Dataset Size**: Limited to 500 samples for computational efficiency
2. **Model Variants**: Focus on smaller model versions
3. **Evaluation Scope**: Primarily automatic metrics without human evaluation
4. **Domain Specificity**: Single dataset evaluation limiting generalizability

6.2 Future Research Directions

1. **Large-Scale Evaluation**: Expansion to full dataset and larger model variants
2. **Multi-Domain Testing**: Cross-domain performance evaluation
3. **Human Evaluation**: Incorporation of human judgment metrics
4. **Advanced Architectures**: Integration of newer transformer variants (T5-v1.1, GPT-3 variations)
5. **Efficiency Optimization**: Model compression and inference acceleration techniques

6.3 Practical Applications

1. **Chatbot Development**: Integration into conversational AI systems
2. **Educational Tools**: Automated tutoring system components
3. **Customer Service**: FAQ automation and response generation
4. **Knowledge Base Systems**: Intelligent information retrieval interfaces

---

7. Conclusion

7.1 Key Findings

This comparative study reveals several critical insights:

1. **GPT-2 Demonstrates Superior Performance**: Contrary to initial expectations, GPT-2 significantly outperformed T5 across all evaluation metrics:
    - **3.3x higher ROUGE-1 scores** (0.283 vs 0.086)
    - **1.6x better ROUGE-2 performance** (0.055 vs 0.035)
    - **2.4x superior ROUGE-L scores** (0.194 vs 0.080)
    - **3.7x better METEOR scores** (0.217 vs 0.059)
    - **1.5x higher F1 scores** (0.264 vs 0.172)
2. **Task Formulation Impact**: The autoregressive nature of GPT-2 with "Question: {q} Answer: {a}" formatting proved more effective than T5's "question: {input}" approach for this specific dataset.
3. **Training Efficiency Differences**:
    - **GPT-2**: Achieved stable performance quickly with validation loss plateauing around 2.82
    - **T5**: Required more training with validation loss decreasing from 6.04 to 5.38 but still underperforming
4. **Exact Match Challenge**: Both models achieved 0% exact match accuracy, indicating the difficulty of generating precisely matching answers and highlighting the need for more nuanced evaluation approaches.
5. **BERT Architectural Validation**: The study confirms BERT's unsuitability for generative QA tasks, reinforcing architecture-task alignment principles.
6. **Data Characteristics Impact**: The Quora dataset's brief, practical answer style (peak at 20-40 characters) appears more compatible with GPT-2's continuation-based generation approach.

7.2 Practical Recommendations

For Production Use:

- Prefer GPT-2 for open-domain QA tasks.
- Use T5 for more structured output formats.

- Apply thoughtful prompt engineering for best results.
- Use multi-metric evaluation due to 0% exact match.

For Research:

- GPT-2 and T5-small serve as effective baselines.
- Investigate architecture-task alignment further.
- Expect performance to vary by dataset.
- Explore scaling to larger models for improved results.

The findings provide valuable guidance for practitioners and researchers working on question answering systems, emphasizing the critical importance of matching model architecture to task requirements while maintaining robust implementation practices.

---

8. References and Technical Specifications

8.1 Model Specifications

- **T5-small**: 60M parameters, encoder-decoder architecture
- **GPT-2**: 117M parameters, decoder-only architecture
- **BERT-base**: 110M parameters, encoder-only architecture

8.2 Technical Dependencies

- **Framework**: Transformers library (Hugging Face)
- **Evaluation**: Evaluate library for standardized metrics
- **Preprocessing**: NLTK for text processing
- **Visualization**: Matplotlib, Seaborn, Plotly for analysis

8.3 Computational Requirements

- **Memory**: Minimum 8GB RAM recommended
- **GPU**: CUDA-compatible GPU preferred for training acceleration
- **Storage**: Approximately 2GB for models and datasets

This comprehensive analysis provides a foundation for informed decision-making in question answering system development and establishes a framework for future comparative studies in transformer-based NLP applications.

GitHub Link: https://github.com/khatrivikas999/indigo_usecase.git

Note: This study was conducted on a small subset of the full dataset (~500 samples out of 56,000) due to limited computational resources.