Dream Interpreter: A Comparative Study of GPT-2 and T5 for Automated Dream Interpretation

Dr. Sharon Yalov-Handzel, Aviv Salomon, Son Levi, Ben Gornizky

Afeka College of Engineering, Tel Aviv

sharony@afeka.ac.il, avivsalo@gmail.com, son.xoxo@gmail.com, Bengornizky@gmail.com

https://github.com/Soncity2/NLP-project-dream-interpretation/tree/main

March 2025

Abstract

This paper presents a study on generating dream interpretations using two large language models (LLMs): GPT-2 and T5. Both models were fine-tuned on a merged dataset comprising the *Dictionary of Dreams* dataset (from Kaggle) and a CSV dataset derived from Freud's *The Interpretation of Dreams*. The generated outputs were evaluated using SacreBLEU, ROUGE-L, BERTScore-F1, and Perplexity. Our results reveal that while GPT-2 produces verbose and less structured interpretations, T5 yields more concise and semantically coherent outputs. Quantitatively, T5 outperforms GPT-2 across most metrics, suggesting that its text-to-text framework is better suited for the dream interpretation task.

1 Introduction

Dream interpretation has long been an area of interest in psychology and creative writing. In this project, we explore the use of LLMs to generate automated dream interpretations. By leveraging pre-trained models and fine-tuning them on a dataset specifically curated for dream symbolism, our aim is to develop a *Dream Interpreter* capable of providing meaningful and structured interpretations for various dream symbols.

2 Methodology

2.1 Data Preparation

We began with Sigmund Freud's seminal work, *The Interpretation of Dreams* [2], which describes human behavior in relation to dreams. Using GPT-4, we extracted 200 common dreams from the text and sorted them into subject categories. For example, the extracted categories include:

- **200 Common Dreams According to Psychology**
- **Falling & Movement Dreams:**
  - Falling through space
  - Tripping or stumbling
  - Flying effortlessly

  - Flying with difficulty
  - Running in slow motion
  - Running but getting nowhere
  - Paralysis/inability to move
  - Walking through molasses
  - Floating or levitating
  - Swimming with ease
  - Drowning or struggling in water
  - Driving out of control
  - Missing transportation (bus, train, plane)
  - Car brakes failing
  - Being chased but moving slowly

- **Performance & Social Anxiety Dreams:**
  - Being naked in public
  - Being underdressed for an occasion
  - Public speaking failure
  - Forgetting lines in a performance
  - Being unprepared for an exam
  - Missing an important deadline
  - Arriving late to an important event
  - Getting lost in a familiar place
  - Finding yourself in the wrong classroom
  - Teeth falling out
  - Hair falling out
  - Being unable to find a bathroom
  - Being unable to close/lock a door
  - Embarrassing yourself in front of a crowd
  - Being judged by others

- **Pursuit & Escape Dreams:**
  - Being chased by an unknown entity
  - Being chased by an animal
  - Being chased by a monster
  - Being pursued by an authority figure
  - Hiding from a threat
  - Being trapped in a small space
  - Being cornered with no escape

- Running from natural disasters
- Escaping from prison or confinement
- Being stalked by someone
- Trying to scream but no sound comes out
- Unable to call for help
- Phone not working in emergency
- Being hunted
- Escaping from a burning building

The resulting data was compiled into a CSV file containing each dream symbol alongside its corresponding interpretation as derived from Freud's text.

2.2 Dataset

We merged two datasets for fine-tuning: the *Dictionary of Dreams* dataset (available on Kaggle) and the CSV dataset derived from Freud's *The Interpretation of Dreams*. This merged dataset provides a rich and diverse source of symbolic and metaphorical language ideal for training our generative models.

2.3 Model Selection and Training

Two distinct LLM architectures were chosen for this study:

- **GPT-2 [5]:** An autoregressive transformer model designed primarily for generating coherent long-form text. GPT-2 was trained on the WebText dataset (approximately 40GB of curated internet text) using a next-word prediction objective. This unsupervised training enabled GPT-2 to learn a wide range of linguistic patterns and generate creative outputs. For our task, it was fine-tuned on the merged dream dataset to learn the mapping between a dream symbol and its interpretation.
- **T5 [6]:** A text-to-text transformer model that frames every NLP problem as a text transformation task. T5 was pre-trained on the cleaned Colossal Clean Crawled Corpus (C4) using a denoising objective that requires the model to reconstruct corrupted text. This approach allows T5 to excel at tasks such as summarization and translation. In our work, T5 was fine-tuned on the same merged dream dataset, leveraging its encoder-decoder architecture to generate concise and structured interpretations.

Fine-Tuning Procedure and Hyperparameters

Both models underwent a structured fine-tuning process over 5 epochs with similar training configurations, although some hyperparameters were model-specific.

GPT-2 Fine-Tuning:

- **Model Name:** `openai-community/gpt2`
- **Output Directory:** `./models/fine_tuned_gpt2`
- **Logging Directory:** `./logs`
- **Batch Size:** 4 (for both training and evaluation)

- **Epochs:** 5
- **Save Strategy:** "epoch" with a total save limit of 2
- **Evaluation Strategy:** "epoch"
- **Logging Frequency:** Every 50 steps
- **Learning Rate:** 5×10^{-5}
- **Weight Decay:** 0.01
- **Warmup Steps:** 100
- **Gradient Clipping:** Maximum gradient norm set to 1.0

T5 Fine-Tuning:

- **Output Directory:** `./models/fine_tuned_t5`
- **Logging Directory:** `./logs/t5`
- **Batch Size:** 4 (for both training and evaluation)
- **Epochs:** 5
- **Save Strategy:** "epoch" with a total save limit of 2
- **Evaluation Strategy:** "epoch"
- **Logging Frequency:** Every 50 steps
- **Learning Rate:** 5×10^{-5}
- **Weight Decay:** 0.01
- **Gradient Clipping:** Maximum gradient norm set to 1.0

These settings ensured that both models were trained under similar conditions, allowing for a direct comparison of their inherent capabilities and architectural strengths.

2.4 Evaluation Metrics

To measure the quality of the generated interpretations, the following metrics were used:

- **SacreBLEU [4]:** Assesses the n-gram overlap between generated outputs and reference interpretations.
- **ROUGE-L [3]:** Evaluates the longest common subsequence between the generated and reference texts.
- **BERTScore-F1 [7]:** Measures semantic similarity by comparing contextual embeddings of generated and reference texts.
- **Perplexity [1]:** Indicates the confidence of the language model in generating the sequence (applicable for GPT-2; note that perplexity was not reported for T5).

Before presenting the metric formulas, we now provide the mathematical definitions along with the corresponding references.

BLEU Metric: As defined in [4], the BLEU score is computed as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{1}$$

where BP is the brevity penalty, and w_n are the weights for n-gram precision p_n.

ROUGE-L Metric: As defined in [3], the ROUGE-L score is computed as:

$$\text{ROUGE-L} = \frac{(1+\beta^2) \cdot \text{LCS}(X,Y)}{\text{len}(X) + \beta^2 \cdot \text{len}(Y)}, \tag{2}$$

where $\text{LCS}(X,Y)$ is the longest common subsequence between the generated text X and the reference Y, and β is a parameter balancing recall and precision.

BERTScore-F1: As defined in [7], BERTScore-F1 is computed as:

$$\text{BERTScore-F1} = \frac{2 \cdot P \cdot R}{P + R}, \tag{3}$$

where P and R denote the precision and recall based on contextual embeddings.

Perplexity: According to [1], perplexity is computed as:

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N} \log p(x_i)\right), \tag{4}$$

where $p(x_i)$ is the probability assigned by the language model to token x_i, and N is the total number of tokens.

3 Experimental Results

3.1 Quantitative Comparison

The evaluation results for each model are as follows:

GPT-2 Results [**?**]:

- SacreBLEU: 0.0186
- ROUGE-L: 0.0096
- BERTScore-F1: 0.7958
- Perplexity: 59.89

T5 Results [**?**]:

- SacreBLEU: 0.2029
- ROUGE-L: 0.0495
- BERTScore-F1: 0.8307
- Perplexity: Not reported

The substantial improvements in SacreBLEU and ROUGE-L for T5 indicate a higher n-gram and sequence overlap with the reference interpretations. The elevated BERTScore-F1 for T5 also reflects a closer semantic match to the ground truth. Although perplexity was only available for GPT-2 (approximately 60), its moderate value suggests some uncertainty in its generated sequences.

3.2 Qualitative Analysis

A review of sample outputs reveals that:

- **GPT-2** tends to produce lengthy and sometimes erratic interpretations, with occasional repetition and less structured responses.
- **T5** consistently generates more succinct and coherent interpretations. While some outputs include repetitive phrases, the overall quality is more aligned with concise dream symbolism interpretation.

3.2.1 Sample Output Comparison for the Dream Symbol "Falling"

For the dream symbol **Falling**, the generated outputs from both models are as follows:

GPT-2 Output:

```
Dream: Falling
Interpretation: To dream that you are falling represents feelings of helplessness and
    hopelessness.
```

T5 Output:

```
To dream that you are falling represents your desire to be more open and to feel more at
    home in your own life.
```

The comparison shows that GPT-2 produces an output emphasizing feelings of helplessness and hopelessness, while T5 generates a concise interpretation focusing on openness and the desire for a sense of belonging.

3.2.2 Sample Output Comparison for the Dream Symbol "Flying"

For the dream symbol **Flying**, the generated outputs are:

GPT-2 Output:

```
Dream: Flying
Interpretation: To dream that you are flying signifies your ability to handle life's ups
    and downs.
```

T5 Output:

```
To dream that you are flying represents your ability to travel in a different direction.
```

Difference: GPT-2 frames flying as a metaphor for managing life's fluctuations, while T5 emphasizes the potential for exploring new directions and opportunities.

3.2.3 Sample Output Comparison for the Dream Symbol "Teeth falling out"

For the dream symbol **Teeth falling out**, the generated outputs are:

GPT-2 Output:

```
Dream: Teeth falling out
Interpretation: To see or dream that you have a leaking tooth indicates your desire to
    rid yourself of unwanted possessions.
```

T5 Output:

```
To see teeth falling out in your dream represents your inner turmoil. You are trying to
    protect yourself from being held captive.
```

Difference: GPT-2 interprets the dream as a desire to discard unwanted elements, whereas T5 highlights internal conflict and the struggle against feeling trapped.

3.2.4 Sample Output Comparison for the Dream Symbol "Being chased"

For the dream symbol **Being chased**, the generated outputs are:

GPT-2 Output:

```
Dream: Being chased
Interpretation: To dream that you are chasing something indicates fear, guilt or other
    negative emotions.
```

T5 Output:

```
To dream that you are being chased indicates that you are feeling a little apathetic
    towards others.
```

Difference: GPT-2 associates the act of chasing with intense negative emotions such as fear and guilt, while T5's interpretation suggests a more subdued, detached emotional state.

4 Discussion

4.1 Model Architecture Comparison

Table 1 summarizes some key architectural differences between GPT-2 and T5.

Model	Architecture	Layers	Parameters
GPT-2 (small)	Decoder-only Transformer	12	117M
T5 (base)	Encoder-Decoder Transformer	12 (encoder) + 12 (decoder)	220M

Table 1: Comparison of GPT-2 and T5 model architectures.

4.2 Architectural Analysis: GPT-2 vs. T5

Both GPT-2 [5] and T5 [6] are built upon the transformer architecture, but they differ significantly in design, training objectives, and training data.

GPT-2:

- **Design and Training:** GPT-2 is a decoder-only transformer model trained on the WebText dataset, which consists of approximately 40GB of curated internet text using a next-word prediction objective. This unsupervised training enabled GPT-2 to capture diverse linguistic patterns and generate creative outputs.
- **Training Characteristics:** The model learns to generate coherent long-form text by predicting the next word in a sequence. Its training objective, however, can lead to overly verbose or less focused outputs, as the model is encouraged to continue the context as long as possible.

- **Strengths and Weaknesses:** GPT-2 excels at generating fluent and imaginative text but may sometimes lack precision due to its unidirectional (left-to-right) context.

T5:

- **Design and Training:** T5 employs an encoder-decoder transformer architecture and is pre-trained on the cleaned Colossal Clean Crawled Corpus (C4) using a denoising objective. The denoising task forces T5 to reconstruct corrupted text, thereby learning robust representations for text-to-text transformations.
- **Training Characteristics:** The text-to-text framework allows T5 to excel at a variety of NLP tasks by converting all tasks into a unified text generation problem. This approach enhances its ability to generate concise and well-structured outputs.
- **Strengths and Weaknesses:** T5's bidirectional context from the encoder and its explicit reconstruction objective enable more focused and semantically accurate outputs. However, its encoder-decoder structure generally requires more computational resources during training and inference.

4.3 Comparative Overview: Similarities and Differences

Similarities:

- Both models are built on the transformer architecture and utilize self-attention mechanisms to capture long-range dependencies.
- They benefit from large-scale unsupervised pre-training on diverse corpora, capturing extensive linguistic and semantic patterns.
- Both models can be fine-tuned on domain-specific data, as demonstrated in this study on dream interpretation.

Differences:

- **Architecture:** GPT-2 is a decoder-only model optimized for autoregressive text generation, while T5 uses an encoder-decoder framework that supports explicit text-to-text transformations.
- **Training Objectives:** GPT-2's next-word prediction encourages continuous text generation, which can lead to verbosity. In contrast, T5's denoising objective enhances its ability to understand context and generate concise, structured text.
- **Output Characteristics:** GPT-2 often produces longer, more creative outputs that may lack focus, whereas T5 yields concise and semantically precise outputs.
- **Resource Requirements:** Due to its encoder-decoder architecture, T5 typically requires more computational resources during both training and inference compared to GPT-2.

4.4 Implications for Dream Interpretation Tasks

For applications such as dream interpretation, where clarity and semantic precision are crucial, T5's architecture demonstrates clear advantages. Its ability to generate concise, contextually rich interpretations makes it more suitable for extracting meaningful insights from symbolic and metaphorical data. In contrast, GPT-2's tendency to generate lengthy outputs can result in interpretations that, while creative, may lack the focus and clarity required for effective analysis.

5 Conclusion

This study compares GPT-2 and T5 for the task of automated dream interpretation. Both models were fine-tuned under similar conditions on the merged dream dataset. Quantitative evaluations show that T5 outperforms GPT-2 in generating coherent and semantically accurate interpretations. The detailed architectural analysis highlights that while both models share a transformer foundation, the encoder-decoder design and denoising pre-training of T5 lend it significant advantages for text transformation tasks. Future work may extend this comparison to other domains of symbolic text generation and explore additional hyperparameter optimization to further refine model outputs.

Acknowledgment

We would like to express our sincere gratitude to Dr. Sharon Yalov-Handzel for her invaluable guidance, dedication, and the generous amount of time and patience she devoted to teaching and supporting us throughout this project.

References

[1] Stanley F. Chen, Doug Beeferman, and Ron Rosenfeld. Evaluation metrics for language models. Technical report, Carnegie Mellon University, 1998. `https://kilthub.cmu.edu/articles/journal_contribution/Evaluation_Metrics_For_Language_Models/6605324/files/12095765.pdf`.

[2] Sigmund Freud. *The Interpretation of Dreams*. Basic Books, 1900. Available at: `https://psychclassics.yorku.ca/Freud/Dreams/dreams.pdf`.

[3] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. OpenAI Blog, `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. arXiv:1910.10683.

[7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*, 2020.