Analyzing the Transformer Architecture
In this exercise, you will analyze the Transformer architecture introduced in "Attention Is All You Need".
Your response must be structured, formal, and academically written.
Use clear headings and concise technical explanations

Part A — Architecture Decomposition
Provide a structured explanation of the main components of the Transformer encoder block.
Your answer must include:

1. input embeddings
2. positional encoding
3. multi-head self-attention
4. feed-forward network
5. residual connections
6. layer normalization

For each component, explain:

- its function
- why it is necessary
- what problem it solves

You may include a diagram.

Part B — Self-Attention Calculation
Consider the following token sequence:
`"The model learns patterns"`
Assume the model is processing the token:
`learns`
Task: Explain conceptually how self-attention computes the contextual representation of this token.
Your answer must explicitly refer to:

- query
- key
- value
- attention scores
- weighted sum

You do not need to compute numerical values.
A conceptual matrix illustration is encouraged.

Part C — Why Transformers Replaced RNNs
Write a short analytical comparison between:

- RNN / LSTM architectures
- Transformer architectures

Discuss at least:

- parallelization

- long-range dependencies
- computational complexity
- training efficiency

Conclude with a brief explanation of why Transformers became dominant in modern NLP.

Part D — Critical Reflection

In 1–2 paragraphs, discuss one limitation of Transformer architectures.

Possible topics include:

- quadratic attention cost
- context window limits
- interpretability
- hallucinations in large models
- dependence on large-scale data and compute

Support your reasoning with technical arguments.

Part E — Bonus Research Question

Optional for extra credit.

Discuss one recent extension of the Transformer architecture, such as:

- sparse attention
- linear attention
- retrieval-augmented transformers
- multimodal transformers
- vision transformers

Explain how it addresses a limitation of the original architecture.

Submission Requirements

- 2–4 pages
- formal academic writing
- references required if external sources are used
- file format: PDF

Python script to run BertViz transformer visualizer is available in parent directory.
To run it, you need Google account.