Exploring Transformers

1. Transformers and Their Importance in Modern AI

Before transformers, AI models like Recurrent Neural Networks (RNNs) were constrained by a fundamental limitation: long sequences of text needed to be processed sequentially, token by token. This linear approach created some significant issues. As text sequences increased in length, models would lose information from earlier on in the sequence making it difficult to grasp context and understand relationships between distant words in a paragraph. This sequential processing was also inherently computationally expensive and slow.

The transformer architecture, introduced in 2017, provided a solution to these issues with a mechanism called self-attention. Self-attention allows the model to look at the entire sequence at once and assign weights of importance to every word in relation to every other word, enabling the model to quickly identify the most relevant pieces of information, even for distant words. By processing data in parallel rather than sequentially, transformers are not only more contextually aware but also far more efficient to train than prior architectures. this breakthrough in efficiency and understanding is what directly enabled the development of the advanced, large-scale AI models that we use today.

2. Methodology

Four distinct prompt types were designed to test the model's capabilities in different domains:

- **News Headline Prompt:** To test factual and formal text generation.
- **Short Story Opener Prompt:** To evaluate creative and narrative generation.
- **Dialogue Prompt:** To assess the ability to generate conversational and character-driven text.
- **Factual Prompt:** To measure adherence to a given fact or the ability to answer a question.

Each of these four prompts was tested across three temperature settings: **0.9**, **0.5**, and **0.1**. This setup allowed for a direct, prompt-for-prompt comparison, isolating the variable of temperature to clearly demonstrate its effect on the output for each specific task. The high temperature (0.9) was expected to produce more creative and random text, while the low temperature (0.1) was expected to be more deterministic and focused.

3. Observations

Analysis of the GPT-2 model's output across the four prompt types and three temperature settings revealed clear and consistent patterns. The temperature setting was the single most significant factor influencing the coherence, creativity, and structural integrity of the generated text. A clear trade-off emerged between logical coherence and creativity, with higher temperatures producing more varied but nonsensical text, and lower temperatures resulting in severe repetitive degradation.

High Temperature (0.9): Superficial Coherence:

At a temperature of 0.9, the model consistently demonstrated an impressive ability to recognize the context and format of the prompt. However, while grammatically correct and stylistically appropriate, the generated content was logically unsound upon closer inspection. This "superficial coherence" was a recurring theme across all high-temperature tests.

A prime example of this was the output for the "News Headline" prompt, which adopted the form of a news article but was filled with illogical statements and fabricated details:

> **Prompt:** `Why our Ancestors Drilled Holes in Each Other's Skulls`
>
> **Output (temp=0.9):** "(SACRAMENTO) In the late 1990s, in response to the growing body of evidence linking paleontology research to the creation and spread of modern humans, the Grand Canyon National Park Conservancy took action by requiring paleontologists to examine their skulls and skulls to determine whether they represent human ancestors. For those looking to see if the skulls have been reconstructed by modern humans, paleoclimates were made to be placed in an oak or oak grove and placed into a pit dug in order to study the evolution of humans. They were then turned over to researchers who then examined them and had an opportunity to test the skull quality and analysis using an X-ray technique. Some of the skulls were altered in the process of development, such as those of African American, Asian, and possibly white (SACRAMENTO, p. 35). These skulls then were analyzed using a series of x-ray tomography scans (Klebs and Bones, 2002). The results showed that the DNA of those individuals showed a strong correlation with ancestry. Some of the bones of Native American Americans in the areas affected by the hominid, like the Upper Miowa area, were later found to be much older than their ancestral ancestors."

This text is stylistically convincing, yet contains nonsensical phrases (`...examine their skulls and skulls...`), irrelevant details (the Grand Canyon), and invented citations (`Klebs and Bones, 2002`), illustrating the model's ability to mimic form without genuine comprehension.

Mid-to-Low Temperatures (0.5 and 0.1): Repetitive Collapse

As the temperature was lowered, the model's tendency to get stuck in repetitive loops increased dramatically. At a temperature of 0.5, the model would often begin with a coherent response before quickly devolving. The Short Story prompt demonstrated this gradual breakdown, maintaining a narrative for several sentences before getting stuck:

> **Prompt:** `The air turned black all around me`
>
> **Output (temp=0.5):** "The air turned black all around me. The sun was setting. The air was hot and the cold. It was cold, and it was cold like a cold day. I was on the ground. I was sitting on the ground, and my eyes were closed. My whole body was cold. My whole body was cold. I was

> not happy. I was not happy. I was not happy. I was not happy. I was not happy. I was not happy..."

At a temperature of 0.1, this degradation was typically immediate and absolute. The output for the Dialogue prompt exemplifies this complete collapse, where the model was unable to generate any new text beyond the prompt itself:

> **Prompt:** `"I'm not buying that umbrella," Sophie said indignantly, "It's too expensive!"`
>
> **Output (temp=0.1):** `"I'm not buying that umbrella," Sophie said indignantly, "It's too expensive!"`
>
> `"I'm not buying that umbrella," Sophie said indignantly, "It's too expensive!"`
>
> `"I'm not buying that umbrella," Sophie said indignantly, "It's too expensive!"`
>
> `"I'm not buying that umbrella," Sophie said indignantly, "It's too expensive!"...`

An interesting exception was the Factual Statement prompt, which at lower temperatures began looping entire sentences rather than single phrases, suggesting that the prompt's structure can influence the specific pattern of repetitive failure.

4. Reflection

This investigation provided me with appreciation for the revolutionary impact of the transformer architecture, as introduced in "Attention Is All You Need." The conceptual leap from slow, sequential computation to a parallelized mechanism capable of weighing the importance of all parts of a sequence at once is both clever and impressive. It illustrates an important shift in AI development, and is the clear prerequisite for the powerful models that have followed.

Conducting this experiment with GPT-2, an early and influential transformer model, served as a practical and illustrative lesson. The limitations of this architecture are very clear when compared to the capabilities of modern AI. The outputs, particularly at a high temperature, were often a "hallucination" of coherent text, stylistically correct but logically unsound. This tendency to invent details, such as the fabricated "(Klebs and Bones, 2002)" citation, is a clear and direct ancestor to the more subtle hallucination issues that still persist in today's state-of-the-art models. While modern systems are vastly more capable, this experiment shows that the fundamental challenge of ensuring factual accuracy has been a part of the transformer's legacy from the beginning.

Perhaps the most surprising observation was the model's rapid degeneration into repetition, especially at lower temperature settings. This behaviour, where the model becomes trapped in a loop of repeating the same phrase or sentence, is not something commonly seen in contemporary models like GPT-4 or other modern LLMs. Witnessing this failure mode firsthand highlights the significant advancements that have been made in model training, architecture, and fine-tuning to improve output stability and coherence.