Understanding Transformer Behaviour Through Decoding Controls

This tool is super interesting, It has helped me very much to visualise what is actually happening behind the hood, highlight how transformers organise information, assign probability mass to tokens and manage uncertainty during generation.

I used the prompt **“The future of AI depends on”** This short, open-ended phrase worked well. At low temperature or restrictive sampling settings, the model tended to complete it with predictable endings such as “technology” or “data.” When I increased temperature or expanded top-k/top-p, the model produced less conventional continuations like “our imagination,” “ethical choices,” or even abstract phrases. This made it easier to observe how different decoding settings shift the distribution of next-token probabilities and how attention influences which completions rise or fall in likelihood.

Temperature is the most intuitive control for observing shifts in output entropy. At low temperature values (close to 0), the model becomes highly deterministic: token probabilities are sharpened, so the model overwhelmingly selects the most likely next token. In practice, this leads to repetitive, predictable and sometimes overly literal continuations. As the temperature increases, the probability distribution flattens, allowing lower-ranked tokens to compete. The generated text becomes more surprising, occasionally more insightful, but also more prone to drifting off-topic. In the visualisation, this is evident as the probability bars for alternative tokens rise, meaning attention-derived representations allow a broader range of possibilities to be considered. The model’s entropy visibly increases, reflecting a more exploratory mode of generation.

Adjusting **top-k** sampling produces a different pattern of behaviour. By restricting the model to the *k* most probable tokens, the distribution is truncated, forcing selection from a small, predefined candidate set. With very small k values (e.g., 5), the output becomes conservative and highly controlled, often echoing expected phrasing. As k increases, the model regains flexibility and can choose from a wider pool, though still bound by the vocabulary topography defined by self-attention. Compared with temperature, top-k feels more rigid: it enforces a hard cutoff, even if token 6 is only slightly less likely than token 5. This sometimes leads to more coherent but less nuanced text.

Top-p (nucleus) sampling provides a more graded alternative. Instead of a fixed number of tokens, it includes only those whose cumulative probability reaches a threshold p. This means that when the model is confident, assigning high probability to a few tokens, the candidate set is small, improving coherence. When uncertainty is high, a larger set is admitted, enhancing creativity. In practice, top-p often yields more natural text than top-k because it adapts to the distribution's shape rather than imposing an arbitrary cutoff. The Explainer shows this clearly: the highlighted candidate set expands or contracts depending on how self-attention has shaped the probability distribution at that step.

These decoding strategies interact closely with **attention mechanisms**. Attention determines how the model integrates information from earlier tokens when constructing the hidden representation that drives next-token probabilities. In the visualisation, we see certain words receive stronger attention weights, meaning they exert more influence on the predicted distribution. When temperature is low, the model's focus on these high-salience tokens becomes even more dominant, pushing the output toward the most predictable continuation. Under higher temperature or broader sampling (higher k or p), tokens with weaker attention influence still enter consideration, which can introduce unexpected but occasionally meaningful deviations. Thus, decoding settings regulate *how much* the model adheres to the structure imposed by attention.

Together, temperature, top-k and top-p provide fine-grained control over the trade-off between randomness and precision. High randomness may be desirable for brainstorming, creative writing or exploring alternative phrasings. Precision matters in summarisation, factual question answering or safety-critical domains. Transformers, equipped with self-attention, are fundamentally more capable of supporting these variable decoding regimes than older RNN-based models. Unlike sequential architectures that relied heavily on previous tokens and suffered from vanishing gradients, transformers compute context holistically at each step. This allows them to adapt fluidly to different sampling strategies without losing coherence entirely.

Overall, the Transformer Explainer illustrates how decoding controls shape the interaction between probability distributions, attention patterns and token selection, making it clearer why transformers behave the way they do under different generation settings.