Understanding Transformer Decoding Through the Lens of EEG Signal Analysis

Exploring the Transformer Explainer tool provides a valuable opportunity to relate the behaviour of language models to concepts familiar in EEG analysis—particularly how models handle uncertainty, prioritise information and integrate long-range dependencies. Although decoding sliders such as temperature, top-k and top-p may seem like superficial sampling controls, their effects mirror many phenomena we observe in neural signal interpretation, where noise levels, signal weighting and thresholding decisions fundamentally alter observed patterns.

Adjusting the **temperature** parameter is a clear parallel to manipulating noise levels or filtering thresholds in EEG preprocessing. At low temperatures, the model behaves like a highly deterministic filtering pipeline: it suppresses "noise" in the probability distribution and consistently selects the most likely next token. This produces repetitive, predictable outputs—much like overly aggressive smoothing in EEG can flatten meaningful variability. As temperature increases, entropy rises and the model becomes more sensitive to lower-probability alternatives. In EEG terms, this resembles examining higher-frequency components or tolerating more variability in sensor space, potentially revealing weaker but behaviourally relevant signals. The visualisation confirms this shift: probability bars widen, meaning the attention-constructed representation supports multiple plausible continuations rather than a single dominant one.

Top-k sampling functions similarly to applying a strict threshold on EEG signal features. Restricting the model to the top k tokens forces it to ignore everything outside a narrow band of likelihood—analogous to keeping only the strongest ERPs or highest-power spectral peaks. When k is small, the output becomes highly constrained and often lacks nuance. Increasing k widens the "feature space" the model is allowed to draw from, but the truncation remains absolute: even a token that is only marginally below the cutoff is unavailable. This resembles the risk of losing subtle neurophysiological signals when thresholding EEG too aggressively.

Top-p (nucleus) sampling aligns more closely with adaptive thresholding strategies in EEG, where decisions are based on cumulative evidence rather than fixed cutoffs. By including only tokens that collectively account for probability mass p, the model adapts its sampling window to the relative certainty of its predictions. When the model is confident, the candidate set is small—similar to moments in EEG where dominant rhythms overshadow weaker activity. When uncertainty is higher, the candidate pool expands, mirroring how broader frequency or spatial windows are sometimes

necessary in EEG to capture distributed or subtle cognitive patterns. In practice, this leads to a balance of coherence and creativity that often produces more natural language.

These decoding behaviours interact deeply with **self-attention**, which integrates contextual information across the entire sequence. Attention determines which prior “signals” (tokens) receive higher weighting—much like how EEG analysis identifies the channels, time points or frequency bands most relevant for a cognitive task. In the Explainer, certain words clearly attract stronger attention, shaping the resulting token distribution. Under low temperature settings, these high-attention signals dominate even more, producing highly predictable outputs. Under higher temperature or broader sampling (higher k or p), tokens with weaker attention relevance still enter consideration, creating opportunities for varied or innovative continuations. This dynamic illustrates how transformers, like EEG-based models, balance dominant and subtle signals during interpretation.

Taken together, the decoding settings—temperature, top-k and top-p—act as tools for modulating the model's exploratory behaviour. In research contexts, higher randomness might be useful when generating hypotheses, exploring linguistic variability or simulating alternative interpretations of ambiguous neural data. More constrained settings are essential in summarisation, clinical reporting or real-time BCI systems where stability and precision matter. The transformer’s capacity for flexible decoding stands in contrast to older RNN architectures, which processed information sequentially and struggled with long-range dependencies—a limitation also seen in classical signal-processing approaches to EEG before the adoption of attention-based neural models.

Ultimately, the Transformer Explainer makes visible how decoding strategies interact with attention to guide token selection. This connection between transformer behaviour and EEG signal interpretation highlights why attention-based architectures have become so powerful in time-series neuroscience and why controlling sampling parameters is essential for shaping coherent, reliable and contextually meaningful outputs.