**Explanations of how temperature and top_p affect AI responses**

Both Temprature and top_p are hyper-parameters that can be used to tune the response of the Large Language Model (LLM).

**Temprature (T)** influences the probabilites of the model outputs, such that a lower temprature makes the model adopt a more confident, deterministic-like behaviour, while a higher temprature makes its predictive behaviour more random. It's value ranges from 0 to 1. When T is closer to 1, next-word probabilities are distributed more evenly, introducing the likelihood that less "expected" words might be chosen. Meanwhile, when T is closer to 0, the model favors the highest-probability words, which narrows down options for the next word to predict.

**Top-P** (also called nucleus sampling) is a way to make word choices. Instead of limiting the model to a fixed number of words as in Top-k, Top-P looks at the most likely options and selects the smallest group of words whose total probability is at least *P*. At High P (e.g., 0.9 or 1.0),  the model has more freedom to choose from a wide range of words, allowing for creative responses. At low P (e.g., 0.3 or 0.6), the model narrows its focus to just the most probable next words, making the output more predictable.

While temprature scales logits before softmax to adjust the overall randomness of token selection, top-p Limits to tokens within cumulative probability p to adapt the vocabulary size based on confidence.