Guia Técnico: Dominando os Parâmetros de IA Generativa

Assunto: Amostragem e Controle de Variabilidade (Temperatura, Top-K/P, Penalidades e Limites)
Público-alvo: Desenvolvedores, Designers de Prompt e Entusiastas de I.A.

1. O que é Amostragem e a "Folha de Candidatas"

Para entender como a IA escreve, imagine que, para cada nova palavra, ela gera uma **Folha de Candidatas**.

De onde vem essa folha?

A IA analisa três pilares para preencher esta folha:

1. **O seu Prompt:** Suas instruções e o contexto.
2. **Cálculos Internos:** O conhecimento de mundo absorvido no treino.
3. **O que ela já escreveu:** A memória do texto gerado até o momento.

Ordenação e Pesos (Probabilidades)

Assim que a folha é preenchida, a IA **ordena** as palavras e atribui um **peso** (probabilidade) a cada uma.

Simulação da Folha para a frase: "O céu hoje está..."

Palavra (Candidata)	Peso (Chance)	Probabilidade Acumulada
1. azul	45%	45%
2. nublado	20%	65%
3. limpo	15%	80%
4. cinza	10%	90%
5. escuro	5%	95%
6. lindo	3%	98%
7. infinito	1.5%	99.5%
8. verde	0.5%	100%

2. Temperatura: A Vibração da Folha

A **Temperatura** altera os pesos antes da escolha final. Imagine que a Temperatura é o quanto a folha "vibra".

- **Baixa (0.1 - 0.3):** A palavra "azul" (45%) ficaria com quase 99% de peso. A IA ignora o resto.
- **Alta (0.8 - 1.5):** A folha vibra tanto que "azul" cai para 25%, e palavras improváveis como "verde" ou "infinito" ganham força (sobem para 5% ou 10%).

Valor	Estado da Folha	Resultado Típico
0.2	Congelada no topo.	Previsível: escolhe sempre **azul**.
0.7	Vibração natural.	Equilibrado: varia entre **azul**, **nublado** ou **limpo**.
1.2	Vibração caótica.	Inusitado: pode escolher **infinito** ou **verde**.

3. Filtrando a Folha: Top-K vs. Top-P

Top-K: O Corte da Tesoura (Quantidade)

O **Top-K** corta a folha num número fixo de linhas, de cima para baixo.

- **Exemplo 1 (K = 1):** "O céu hoje está **azul**." (Corta na 1ª linha; só permite a opção nº 1).
- **Exemplo 2 (K = 3):** "O céu hoje está **limpo**." (Corta na 3ª linha; a IA sorteia apenas entre azul, nublado e limpo).
- **Exemplo 3 (K = 8):** "O céu hoje está **verde**." (Não há corte; a IA pode escolher qualquer palavra da folha, até a última).

Top-P (Nucleus Sampling): A Zona de Seleção (Qualidade)

O **Top-P** seleciona uma área que deve somar **P%** da probabilidade total (coluna "Acumulada").

- **Exemplo 1 (P = 0.4):** "O céu hoje está **azul**." (A zona de seleção fecha antes mesmo da 2ª palavra, pois 45% já ultrapassa o limite de 40%).
- **Exemplo 2 (P = 0.8):** "O céu hoje está **limpo**." (A zona desce até a 3ª linha para conseguir somar os 80% necessários).
- **Exemplo 3 (P = 0.99):** "O céu hoje está **infinito**." (A zona percorre quase a folha toda até atingir 99%, permitindo a escolha da 7ª palavra).

4. Ajustando a Folha: As Penalidades e o Tamanho

Presence & Frequency Penalty (A Caneta Corretora)

Funcionam como uma correção que empurra para baixo palavras que já foram usadas.

- **Presence Penalty:** Evita repetir o **assunto**. Se já descreveu o céu, a caneta risca "azul" e "nublado" da folha para forçar a IA a falar de outro tema (ex: o clima ou a temperatura).
- **Frequency Penalty:** Evita o "eco". Se a IA já usou "azul" três vezes, essa palavra é jogada para o fim da folha na próxima rodada.

Max Tokens: O Tamanho da Folha

Define o limite máximo de tokens (pedaços de palavras) que a IA pode escrever. É a margem

final do papel. Se o limite for 5 tokens, a IA parará de escrever abruptamente após as primeiras palavras.

5. Análise Comparativa de Prós e Contras

Parâmetro	Quando Brilha (Prós)	Onde Falha (Contras)
Temperatura	Controla a ousadia global.	Se alta demais, gera erros sem sentido.
Top-P/K	Garante qualidade no vocabulário.	Se baixo demais, o texto fica repetitivo.
Max Tokens	Controla custos e tempo.	Pode cortar a resposta no meio de uma frase.
Penalties	Elimina vícios e palavras repetidas.	Pode tornar a gramática estranha se exagerado.

6. Cheat Sheet: Quando ajustar o quê?

Objetivo	Temperatura	Top-P	Penalidades	Max Tokens
Fatos / Dados	0.1	0.2	0.0	Baixo
Chatbot Padrão	0.7	0.8	0.2	Médio
Escrita Criativa	0.9	0.95	0.6	Alto
Poesia / Humor	1.2	1.0	0.8	Médio

Dica de especialista: A Temperatura mexe no peso das palavras (na vibração), enquanto o Top-P/K define quem entra no sorteio (o corte). Ajuste um de cada vez!