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:

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 ao momento.

A **Amostragem (Sampling)** é o conjunto de regras que decide como vamos selecionar o vencedor desta folha.

2. Temperatura: A Vibração da Folha

A **Temperatura** é o primeiro ajuste. Ela não corta a folha, mas altera as probabilidades antes da escolha. Imagine que a Temperatura é o quanto a folha "vibra".

- **Temperatura Baixa (0.1 - 0.3):** A folha está parada. A palavra do topo torna-se muito "pesada" e as outras quase desaparecem. A IA fica ultra-focada.
- **Temperatura Alta (0.8 - 1.5):** A folha vibra intensamente. As palavras do fundo sobem um pouco e a do topo perde força. As probabilidades ficam mais parecidas entre si.

Valor	Estado da Folha	Resultado no Texto
0.2	Congelada no topo.	Previsível, ideal para factos e código.
0.7	Vibração natural.	Equilíbrio entre lógica e fluidez humana.
1.2	Vibração caótica.	Criatividade selvagem, mas pode perder o sentido.

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

Top-K: O Corte da Tesoura (Quantidade)

O **Top-K** limita a escolha às primeiras **K** palavras da folha, ignorando o resto.

- **K = 1:** "O sol nasce no **leste**." (Corte imediato após a 1ª opção).
- **K = 50:** "O dia hoje está **agradável**." (Corte que permite 50 variações seguras).

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

O **Top-P** seleciona uma área no topo da folha que deve somar **P%** de probabilidade total.

- **P = 0.1:** A zona de seleção é minúscula, focando apenas na certeza absoluta.
- **P = 0.9:** A zona de seleção é ampla, incluindo até opções menos prováveis, desde que façam parte do "núcleo" de sentido.

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, forçando a IA a procurar opções novas na folha.

Max Tokens: O Tamanho da Folha

O **Max Tokens** define o limite máximo de "pedaços de palavras" que a IA pode escrever. É a margem final do papel.

- **Tokens Curtos (50-100):** Ideal para títulos, metas descriptions ou respostas rápidas.
- **Tokens Longos (1000+):** Necessário para ensaios, artigos ou roteiros longos.

Atenção: Se o Max Tokens for muito baixo, a IA pode parar de escrever a meio de uma frase.

5. Análise Comparativa de Prós e Contras

Parâmetro	Quando Brilha (Prós)	Onde Falha (Contras)
Temperatura	Controla o "nível de ousadia" global.	Se for muito alta, gera alucinações sem nexo.
Top-P/K	Filtra o vocabulário para manter a qualidade.	Se for muito baixo, o texto fica repetitivo.
Max Tokens	Controla custos e tempo de resposta.	Pode cortar a resposta de forma abrupta.
Penalties	Elimina vícios de linguagem e ecos.	Pode tornar a gramática confusa se for exagerado.

6. Cheat Sheet: Quando ajustar o quê?

Objetivo	Temperatura	Top-P	Penalidades	Max Tokens
Factos / 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 e o Top-P influenciam-se. Normalmente, recomenda-se ajustar apenas um deles de cada vez para manter o controlo sobre o comportamento do modelo.