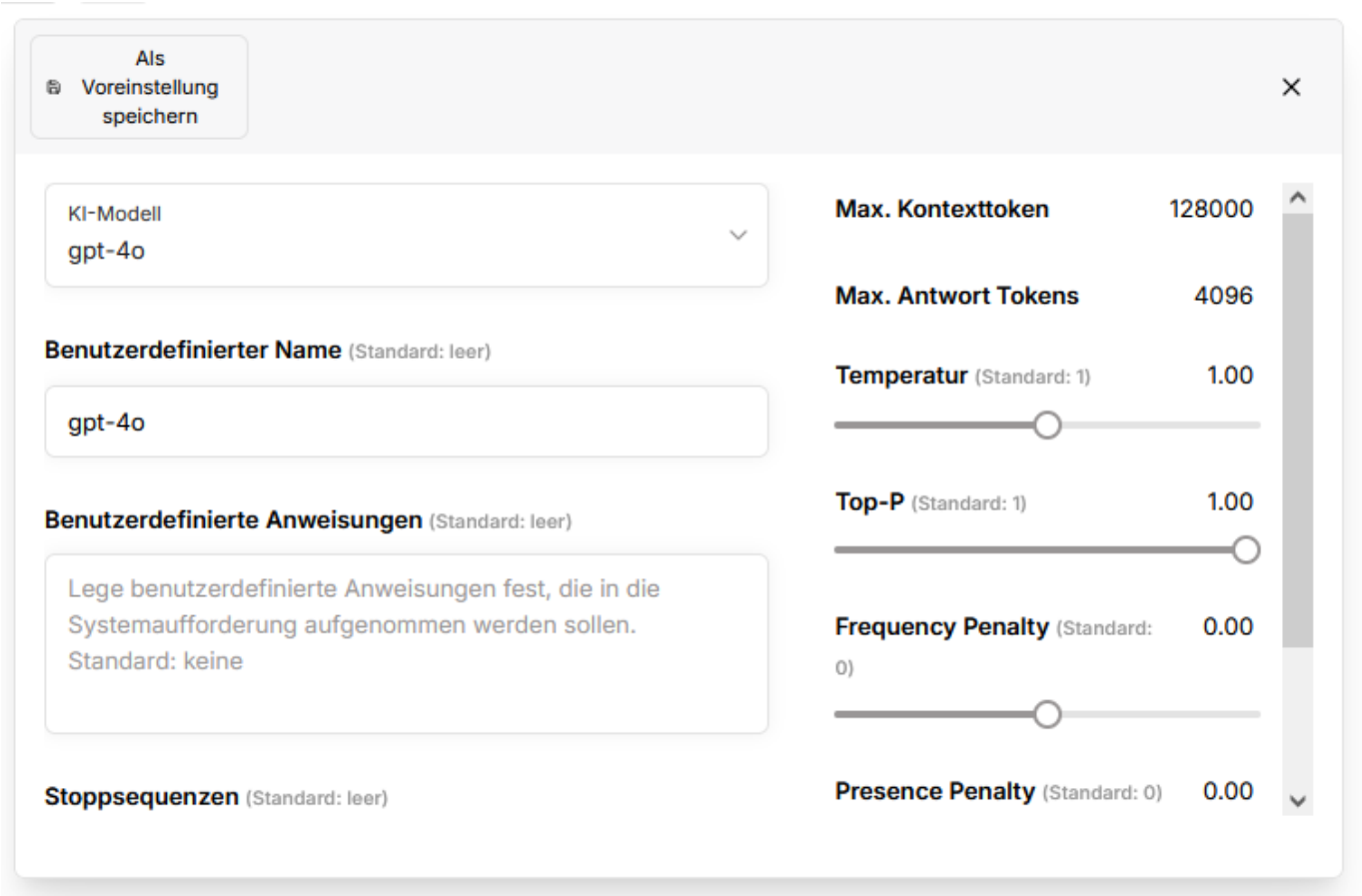
# Einstellung der Modellparameter

Diese Parameter helfen dabei, die Vielfalt, Relevanz und Einzigartigkeit der generierten Antworten zu kontrollieren. Anfängern wird jedoch empfohlen die Standardwerte beizubehalten.




Erläuterung der Modellparameter

- **Kontexttokens**: Dies bezieht sich auf die maximale Anzahl von Token (Einheiten von Wörtern oder Zeichen), die das Modell auf einmal „verstehen“ oder verarbeiten kann. Bei vielen Sprachmodellen gibt es eine obere Grenze für die Anzahl von Token, die im Eingabekontext enthalten sein können. Wenn diese Grenze überschritten wird, muss der Text gekürzt oder geteilt werden.
- **Antwort Tokens**: Dies bezieht sich auf die Tokens, die im generierten Output oder in der Antwort des Modells enthalten sind. Wenn das Modell eine Antwort erzeugt, besteht diese Antwort aus einer bestimmten Anzahl von Token. Es gibt oft auch eine maximale Begrenzung dafür, wie viele Token die Antwort des Modells haben kann, abhängig von der Implementierung oder den Nutzerpräferenzen.

- **Temperatur** (Temperature): Dieser Parameter steuert den Zufallsgrad der Antworten. Ein niedrigerer Wert (z.B. 0.1) führt zu vorhersehbareren und konsistenteren Antworten. Ein höherer Wert (z.B. 1.0) erhöht die Vielfalt und macht die Antworten weniger vorhersagbar. Eine Temperatur von 0 würde immer dieselbe Antwort auf dieselbe Eingabeaufforderung geben (sofern der Rest der Einstellungen identisch ist).
- **Top-P** (Topical Penalization, auch bekannt als Nucleus Sampling): Dieser Wert bestimmt, wie das Modell aus einer Gruppe von wahrscheinlichen nächsten Worten auswählt. Ein niedrigerer Top-P-Wert (z.B. 0.5) bedeutet, dass nur die wahrscheinlichsten Worte berücksichtigt werden. Ein höherer Wert (z.B. 0.95) erlaubt eine größere Vielfalt potenzieller Worte. Dieser Mechanismus funktioniert durch das Beschränken der Auswahl an Wörtern auf ein kumulatives Wahrscheinlichkeitsintervall („P“).
- **Frequency Penalty** (Häufigkeitsstrafe): Dieser Parameter verhindert, dass das Modell dieselben Wörter oder Phrasen wiederholt. Je höher der Wert, desto größer die Strafe für die Wiederholung von Wörtern, was zu einer variableren und diversifizierteren Antwort führt.
- **Presence Penalty** (Anwesenheitsstrafe): Ähnlich wie die Frequency Penalty, aber dieser Parameter bestraft Wörter basierend darauf, ob sie bereits im Text erschienen sind, unabhängig von ihrer Häufigkeit. Ein höherer Wert fördert die Einführung neuer, bisher nicht verwendeter Wörter oder Phrasen in den Antworten.