Structure

GenerationOptions

Options that control how the model generates its response to a prompt.

iOS 26.0+ | iPadOS 26.0+ | Mac Catalyst 26.0+ | macOS 26.0+ | visionOS 26.0+

```
struct GenerationOptions
```

Mentioned in

Generating content and performing tasks with Foundation Models

Overview

Generation options determine the decoding strategy the framework uses to adjust the way the model chooses output tokens. When you interact with the model, it converts your input to a token sequence, and uses it to generate the response.

Only use `maximumResponseTokens` when you need to protect against unexpectedly verbose responses. Enforcing a strict token response limit can lead to the model producing malformed results or gramatically incorrect responses.

All input to the model contributes tokens to the context window of the `LanguageModelSession` — including the `Instructions`, `Prompt`, `Tool`, and `Generable` types, and the model's responses. If your session exceeds the available context size, it throws `LanguageModelSession.GenerationError.exceededContextWindowSize(_:)`. For more information on managing the context window size, see TN3193: Managing the on-device foundation model's context window.

Topics

Creating options

`init(sampling: GenerationOptions.SamplingMode?, temperature: Double?, maximumResponseTokens: Int?)`

Creates generation options that control token sampling behavior.

Configuring the response tokens

`var maximumResponseTokens: Int?`

The maximum number of tokens the model is allowed to produce in its response.

Configuring the sampling mode

`var sampling: GenerationOptions.SamplingMode?`

A sampling strategy for how the model picks tokens when generating a response.

`struct SamplingMode`

A type that defines how values are sampled from a probability distribution.

Configuring the temperature

`var temperature: Double?`

Temperature influences the confidence of the models response.

Relationships

Conforms To

`Equatable`, `Sendable`, `SendableMetatype`

See Also

Prompting

`class LanguageModelSession`

An object that represents a session that interacts with a language model.

`struct Instructions`

Details you provide that define the model's intended behavior on prompts.

`struct Prompt`

A prompt from a person to the model.

`struct Transcript`

A linear history of entries that reflect an interaction with a session.