Model properties

In the Model properties dialogue several advanced properties of an OpenAI model can be set:

Model (display only):

- The name of the referenced OpenAI Model. Note, that with every training this name changes

Model is tunable / **Model is tuned** (display only):

- general capabilities of the underlying Model. Note that only some selected OpenAI Models can be tuned (the model is selected when adding an OpenAI language-resource)

Use default system-message when translating:

- defines, if the default system-message of a training (the topmost message in the training-window) is used when pretranslating with the model

Use user defined system-messages when translating:

- defines, if the other user-defined system-messages of a training are used when pretranslating with the model

Generation Sensitivity / Temperature:

- Temperature is a parameter that governs the randomness and thus the creativity of the responses. It is always a number between 0 and 1. A temperature of 0 means the responses will be very straightforward, almost deterministic (meaning you almost always get the same response to a given prompt) A temperature of 1 means the responses can vary wildly. It's advisable to adjust either the temperature or top_p, but not both.

Probability Threshold / Top P:

- The "top P" parameter, also known as nucleus sampling, is a nuanced alternative to temperature-based sampling. It is a "spotlight" that shines on the most probable words. At a default value of 1.0, the model considers all words. This parameter can help control the distribution of word choices, keeping the generated content relevant and coherent. It's advisable to adjust either the temperature or top_p, but not both.

Presence Penalty:

- This parameter is used to encourage the model to include a diverse range of tokens in the generated text. It is a value that is subtracted from the log-probability of a token each time it is generated. A higher presence_penalty value will result in the model being more likely to generate tokens that have not yet been included in the generated text.

Frequency Penalty:

- This parameter is used to discourage the model from repeating the same words or phrases too frequently within the generated text. It is a value that is added to the log-probability of a token each time it occurs in the generated text. A higher frequency_penalty value will result in the model being more conservative in its use of repeated tokens.

Max. target tokens (% of source tokens):

- A GPT Model always has a maximum size of tokens that can be used within a single request. This amount calculates as the sum of the sent tokens and the returned tokens. For a (pre)translation, this is the system message and the text or batch to translate plus the returned translations. Therefore a ratio is needed to leave "room" in a sent request for the generated translation. This is only relevant for batch-translations as used when pretranslating