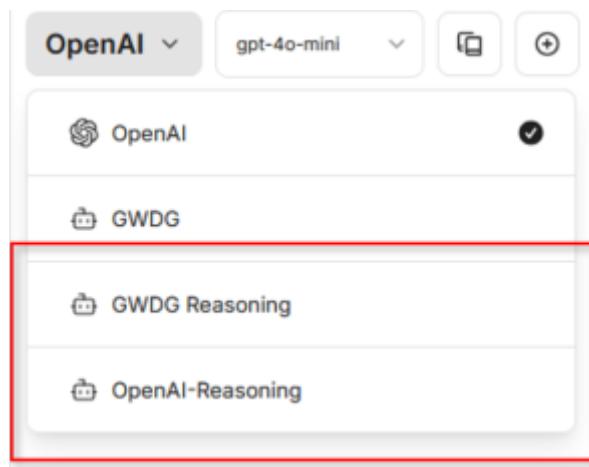


Reasoning Modelle

Reasoning-Modelle sind Sprachmodelle, die mit Reinforcement Learning trainiert wurden, um komplexe Schlussfolgerungen zu ziehen. Reasoningmodelle erstellen eine interne Gedankenkette, bevor sie dem Benutzer antworten. Die Idee hinter diesen Modellen ist es das logische Denken und den Problemlösungsmechanismus nachzubilden, wie etwa Usache-Wirkungs-Beziehungen. Sie zeichnen sich durch eine höhere Genauigkeit und Zuverlässigkeit aus. Dies geht zu Lasten der Geschwindigkeit. Beispiele für diese Art von Modellen sind: o1, o1-mini oder o3-mini.

Derzeit bieten wir sowohl Reasoning-Modelle von OpenAI, also auch von der GWDG, an, um diese Modelle zu benutzen wählen Sie einfach den entsprechenden Endpoint aus:



In den Parametereinstellungen kann der Reasoning-Effort (derzeit nur für OpenAI-o1 Modelle) eingestellt werden. Eine Verringerung des Wertes kann zu schnelleren Antworten führen.

The screenshot shows the configuration interface for the GDWG-Reasoning service. At the top, it displays the model selected: 'deepseek-r1-distill-lama-70b'. Below this are sections for 'Prompts' and 'Parameters'. A 'Custom Name' input field is present, followed by a 'Custom Instructions' section with a note about including them in the System Message. The 'Max Context Tokens' and 'Max Output Tokens' are both set to 'System'. Under the 'Temperature' and 'Top P' sliders, the values are both set to 1.00. The 'Frequency Penalty' and 'Presence Penalty' are both set to 0.00. A 'Stop Sequences' input field is available for separating values by pressing 'Enter'. Below these are 'Resend Files' and 'Image Detail' settings, with 'Image Detail' set to 'auto'. A 'Reasoning Effort' slider is highlighted with a red box and is set to 'medium'. A green 'Save As Preset' button is located below the sliders. At the bottom of the interface are buttons for 'Attach Files', 'Bookmarks', and 'Hide Panel'.

Direkt-Link:

https://doku.tu-clausthal.de/doku.php?id=sonstige_dienste:ki-dienste:librechat:reasoningmodelle&rev=1739973706

Letzte Aktualisierung: 14:01 19. February 2025

