

Gleicher Prompt, anderer Zeitpunkt – anderes Ergebnis?



Periodische Schwankungen in der Leistungsfähigkeit großer Sprachmodelle und ihre Bedeutung für die naturwissenschaftsdidaktische Forschung



Paul Tschisgale & Peter Wulff

tschisgale@leibniz-ipn.de



Take a picture to download the full preprint



Motivation

r/ChatGPTPro • vor 3 Jahren
BeConvincible

Does ChatGPT get better / worse at different times of the day? (When under load.)

I've noticed that, for exactly the same task, ChatGPT performs very well sometimes and other times commits tons of basic mistakes.

I think I've noticed that it does better late in the evening, when presumably fewer people are using it.

Is this just my imagination, or is the model adjusted under load to be stupider?

37 upvotes · 44 comments

Falls das zutrifft, so wären bei **Forschung mit LLMs** (als Gegenstand oder Werkzeug) zentrale **Gütekriterien gefährdet**.

Reliabilität · Reproduzierbarkeit · Validität

Inwieweit weist die Performanz von GPT-4o beim Lösen eines Physikproblems unter konstanten Bedingungen über einen längeren Zeitraum hinweg zeitabhängige Muster auf?

Methodik

MC-Aufgabe aus der PhysikOlympiade, wobei jede Antwortoption separat bewertet werden muss

Eine einzelne Batterie kann eine Glühlampe für eine Zeit t zum Leuchten bringen. Nimm vereinfachend an, dass die Lampe mit konstanter Helligkeit leuchtet, bis die Batterie leer ist, und dass der Widerstand der Glühlampe konstant ist.

Welche der Aussagen ist korrekt, wenn zwei dieser Batterien zum Betreiben von zwei der Glühlampen verwendet werden?

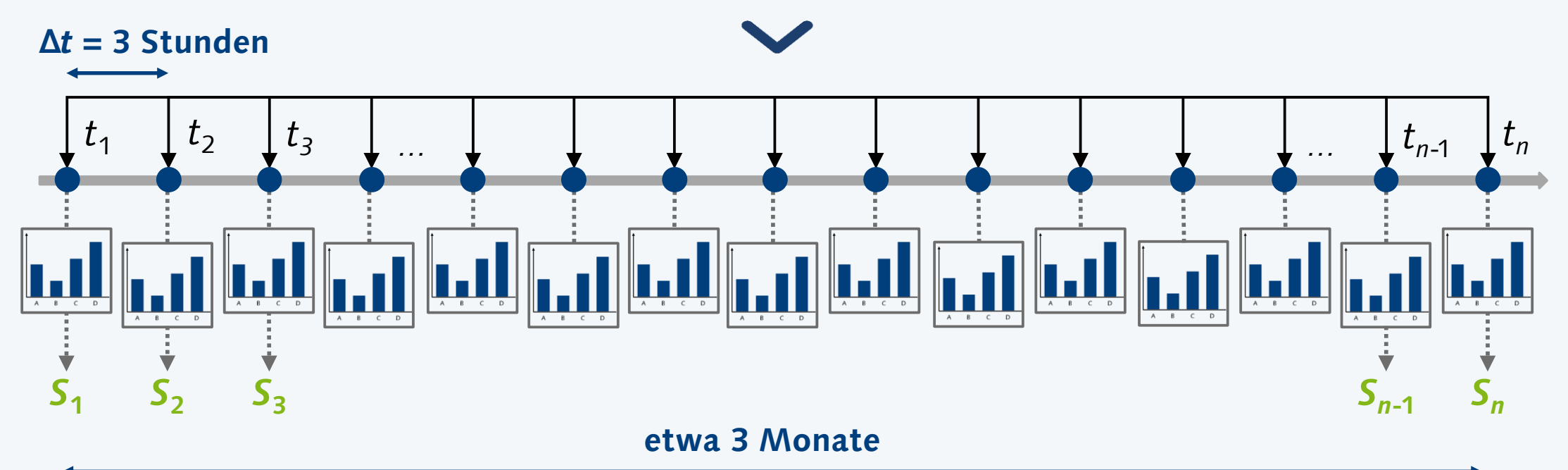
A Wenn die Batterien in Serie und die Glühlampen in Serie geschaltet sind, können die Glühlampen etwa eine Zeit $t/4$ betrieben werden.
B Wenn die Batterien in Serie und die Glühlampen parallel geschaltet sind, können die Glühlampen etwa eine Zeit $t/2$ betrieben werden.
C Wenn die Batterien parallel und die Glühlampen in Serie geschaltet sind, können die Glühlampen etwa eine Zeit $2t$ betrieben werden.
D Wenn die Batterien parallel und die Glühlampen ebenfalls parallel geschaltet sind, können die Glühlampen etwa eine Zeit t betrieben werden.

GPT-4o (gpt-4o-2024-08-06) unter festen Bedingungen (gleicher Prompt, fixierte Hyperparameter) angesteuert über OpenAI-API

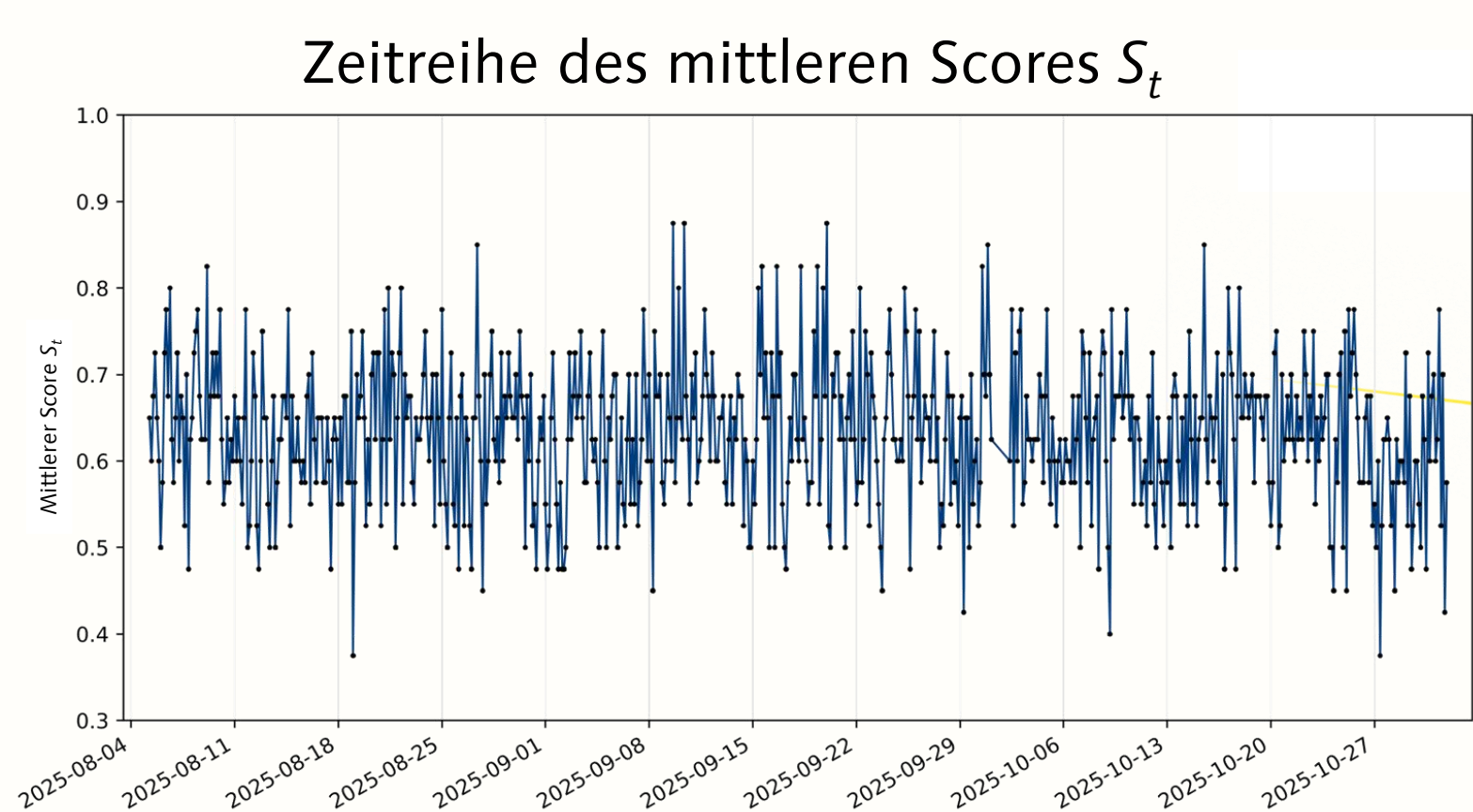
$N=10$ Lösungen alle 3 Stunden für ca. 3 Monate

Verteilung über die Antwortoptionen für $N=10$ Lösungen

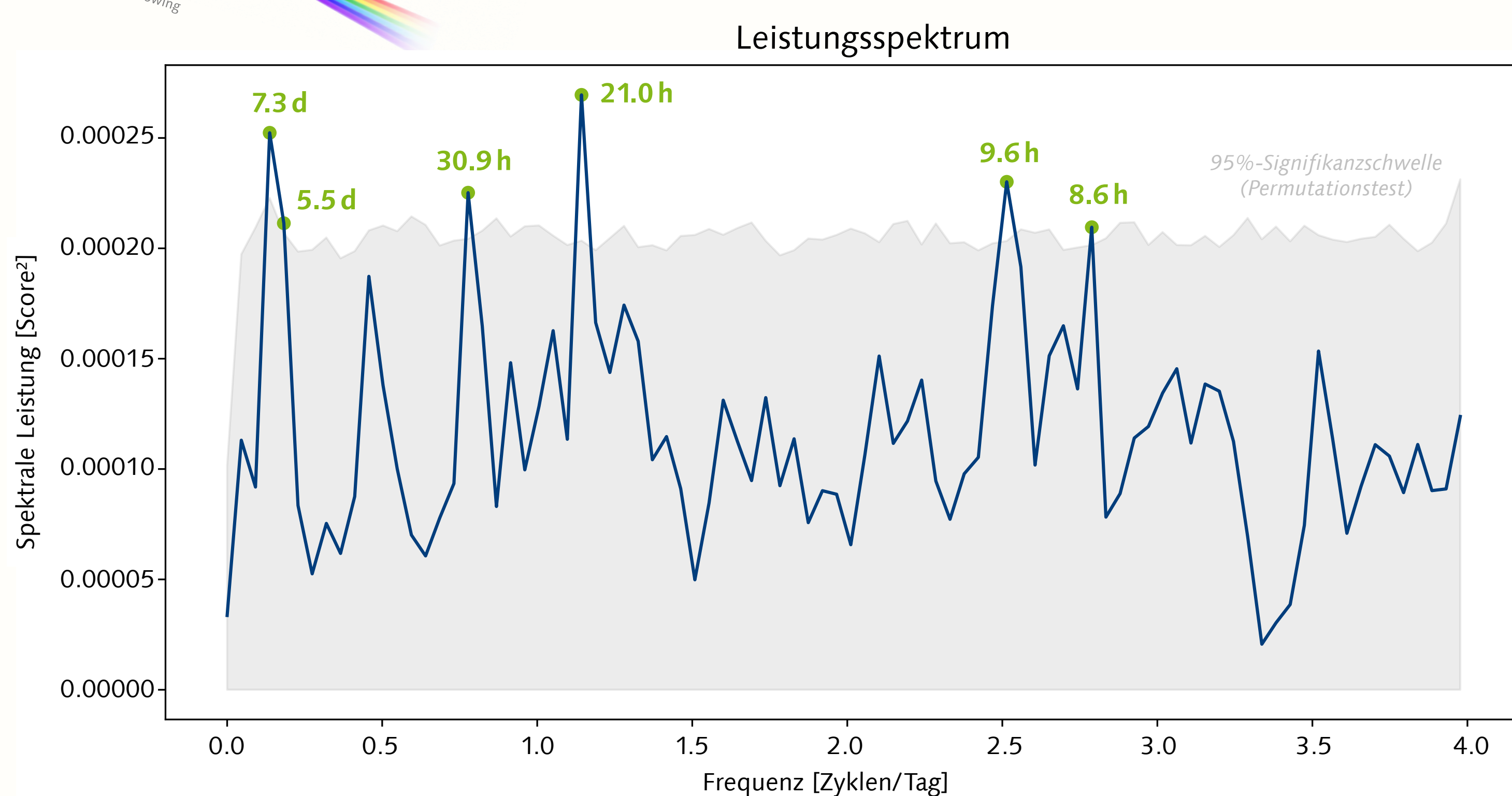
Zeitreihe des mittleren Scores $S_t \in [0,1]$



Ergebnisse



Fast Fourier Transform
Welch's Methode
Hann-Windowing



20.3% Varianz

der Zeitreihe werden durch die signifikanten periodischen Komponenten erklärt

0.139 Spannweite

in der Zeitreihe des Scores (14 % der 0–1-Skala) werden durch die signifikanten periodischen Komponenten erklärt

Es zeigt sich Evidenz, dass die Performanz von GPT-4o trotz konstanter Bedingungen einem **Tages- und Wochenzyklus** folgt.

Warum kein 24-h-Peak?

Der Tagesrhythmus fällt an Werktagen anders aus als am Wochenende. Der Tagesrhythmus (f_d) wird also vom Wochenzyklus (f_w) multiplikativ moduliert (**Amplitudenmodulation**). Das erzeugt keine reine 24-h-Spitze, sondern Peaks bei

$$f = k f_d \pm m f_w$$

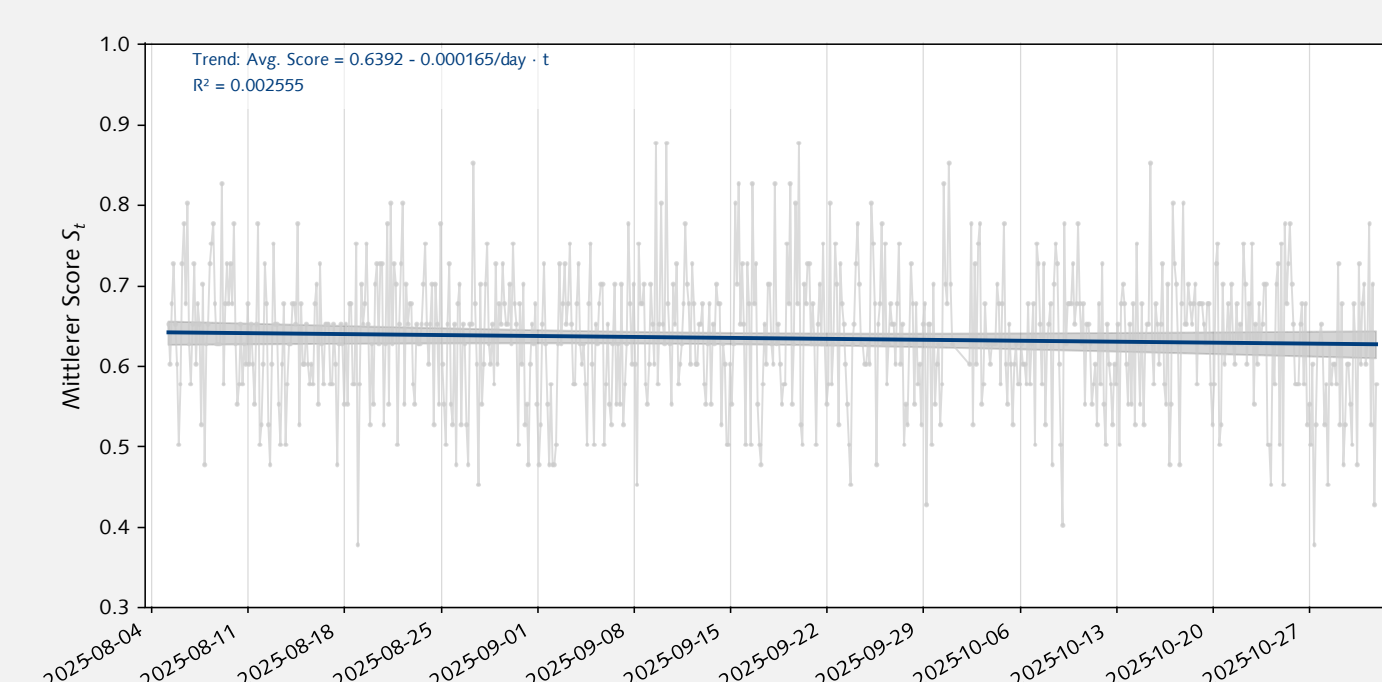
Für $k = m = 1$ enthält man Seitenbänder bei etwa **20.6 h** und **28.4 h**. Diese Bänder zeigen sich annähernd im Spektrum.

$$\begin{aligned} & \cos(2\pi f_d t) \cdot \cos(2\pi f_w t) \\ &= \frac{1}{2} (\cos(2\pi (f_d - f_w) t) + \cos(2\pi (f_d + f_w) t)) \end{aligned}$$

24h 7d
28.4h 20.6h

Weitere Ergebnisse

Kein signifikanter Drift in der Performanz. OLS Regression mit HAC Standardfehlern liefert nicht-signifikanten Drift-Koeffizienten ($t = -1.01$, $p = 0.303$). Die mittlere Korrektheit beträgt $M = 0.63$ ($SD = 0.26$).



Implikation für LLM-basierte Forschung

Erhebungsfenster ist relevant

da ein enges Zeitfenster Schätzungen von LLM-Fähigkeiten verzerren kann — und es dürfte sich auch auf LLM-gestützte Kodierungen auswirken.

Bestehende Befunde zu LLM-Fähigkeiten

Könnten mit größerer Unsicherheit behaftet sein als bislang üblicherweise berücksichtigt.

Empfohlene Vorgehensweise

Daten möglichst über ≥ 1 volle Woche erheben, gleichmäßig verteilt mit mindestens täglichem (idealerweise stündlichem) Sampling.

Mittlere Performanz nach Wochentag $\times$ Uhrzeit. Die Performanz im Tagesverlauf unterscheidet sich je nach Wochentag. Das spricht dafür, dass der Wochentag den täglichen Leistungsrhythmus moduliert.

