A screen shot of a graph

AI-generated content may be incorrect.

Source link: <https://ourworldindata.org/artificial-intelligence>

The graphic visualization above depicts a scatterplot that compares the performance of different artificial intelligence models developed between 2019 and 2023 (such as GPT-4, GPT-3, GPT-2, PaLM-2, and Chinchilla among others) on knowledge tests (based on an MMLU benchmark measurement) with their corresponding training computation (measured in petaFLOPs). It should be noted that the x-axis uses a logarithmic scale to accommodate the vast range in values from 100,000 to more than 10 billion petaFLOPs. Each point of data represents the training computation vs. knowledge performance of a specific artificial intelligence model that is color-coded with multiple colors representing the developer of the model. There is also a dotted line around 90% MMLU to reference expert human performance against the models.

The main story of the chart is that performance on knowledge tests generally improves with increased training computational, however this relationship is nonlinear and may be influenced further by how the models are developed aside from training computation. Thus, a secondary story is that training computation does not explain all of the variability in knowledge test performance of models, as some computationally less expensive models may outperform more expensive models (for example GPT-2 (finetuned) performs better than Gopher (0.4B) despite being less computationally expensive). A tertiary story is that the most computationally-expensive models GPT-4 and PaLM-2 have surpassed other models to achieve similar knowledge test performance as expert humans (though not surpassing expert humans).

The chart aligns with Pre-attentive Processing by using 2-D position on a common scale for both axes to encode information on comparing numerical variables (knowledge test performance vs. training computation), which is further enhanced by using gridlines as visual aids. Therefore, the chart also aligns with attributes of form since 2-D position is an attribute that is easily quantitatively perceived. However, the overlap of many points and inconsistent labelling (not all points are labelled, some points have labels that cover other points) makes it slightly bothersome to discern the position of each point on the common scale.

For Visual Perception techniques, inconsistent labelling of points show lack of visual order. Many points with similar test performance vs. training computation values are placed overlap each other despite having distinct difference in values, meaning there is misalignment that promotes visual clutter. Furthermore, there are many points concentrated in the center of the chart with whitespace existing in the edges, contributing to lack of visual order. Lastly, since some points are labelled while others are not without any explanation, non-strategic use of contrast exists which does not contribute to the main story of comparing the effect of training computation on knowledge test performance for AI models.

The Gestalt Principle of Proximity is supported by the chart since points are clustered together based on closeness in test performance vs. training computation, and points that are more different in test performance or training computational are placed further away. The Gestalt Principle of Similarity is somewhat supported by points with same colors being attributed to the same developer, though this distracts from the overall main story (effect of training computation on test performance).