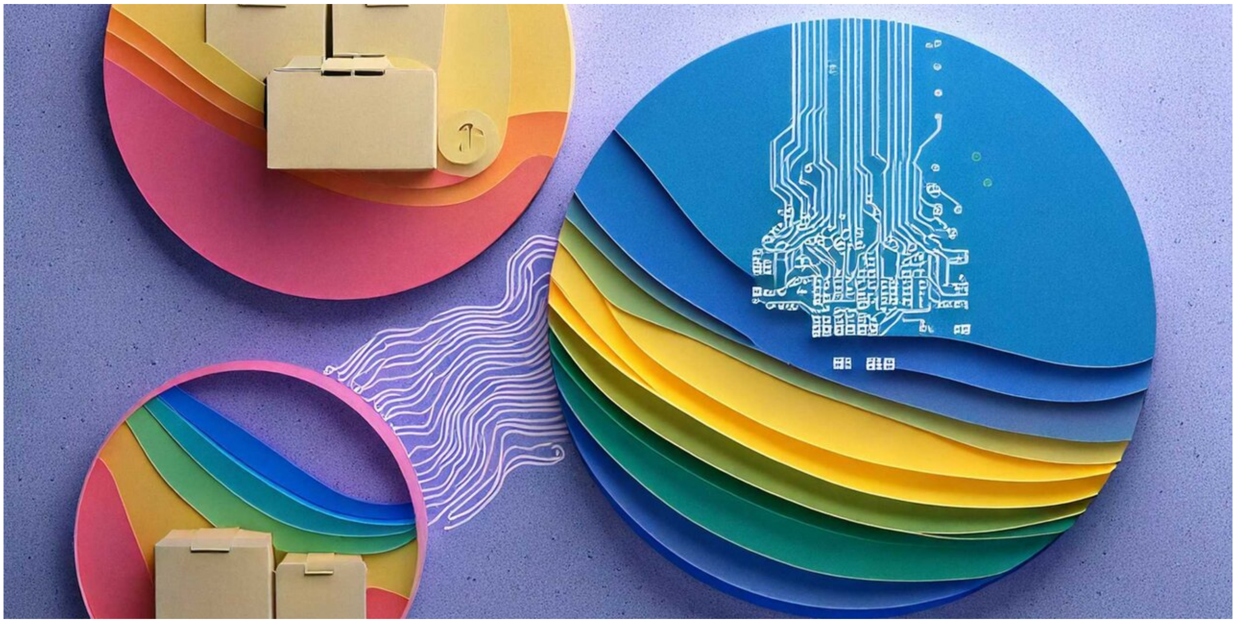
Algorithm can make AI responses increasingly reliable with less computational overhead

April 24 2025, by Daniel Meierhans



A new algorithm from ETH researchers improves large language models (LLMs) so that the selected answers are more accurate and relevant. Credit: AI generated/ ETH Zurich

ChatGPT and alike often amaze us with the accuracy of their answers, but unfortunately, they also repeatedly give us cause for doubt. The main issue with powerful AI response engines (artificial intelligence) is that they provide us with perfect answers and obvious nonsense with the same ease. One of the major challenges lies in how the large language models (LLMs) underlying AI deal with uncertainty.

Until now, it has been very difficult to assess whether LLMs designed for text processing and generation base their responses on a solid foundation of data or whether they are operating on uncertain ground.

Researchers at the Institute for Machine Learning at the Department of Computer Science at ETH Zurich have now developed a method that can be used to specifically reduce the uncertainty of AI. The work is published on the *arXiv* preprint server.

"Our algorithm can enrich the general language model of the AI with additional data from the relevant subject area of a question. In combination with the specific question, we can then extract from the depths of the model and from the enrichment data precisely those connections that are most likely to generate a correct answer," explains Jonas Hübotter from the Learning & Adaptive Systems Group, who developed the new method as part of his Ph.D. studies.

Enriching AI with specific data

"The method is particularly suitable for companies, scientists or other users who want to use general AI in a specialized field that is only covered partially or not at all by the AI training data," adds Andreas Krause, head of the research group and Director of the ETH AI Center.

For example, users can feed their locally stored data into a large language model (LLM), such as Llama. The so-called SIFT algorithm (Selecting Informative data for Fine-Tuning), developed by ETH computer scientists, can then use the additional data provided to select specific information that is most closely related to the question.

Relationship vectors in multidimensional space

The algorithm uses the structure according to which the language information is organized in the AI's large language model (LLM) to find related information. The models divide the language information in their training data into word parts.

The semantic and syntactic relationships between the word parts are then arranged as connecting arrows—known in the field as vectors—in a multidimensional space. The dimensions of space, which can number in the thousands, arise from the relationship parameters that the LLM independently identifies during training using the general data.

Angle between arrows as measure of correlation

Relational arrows pointing in the same direction in this vector space indicate a strong correlation. The larger the angle between two vectors, the less two units of information relate to one another.

The SIFT algorithm developed by ETH researchers now uses the direction of the relationship vector of the input query (prompt) to identify those information relationships that are closely related to the question but at the same time complement each other in terms of content.

"The angle between the vectors corresponds to the relevance of the content, and we can use the angles to select specific data that reduces uncertainty," explains Hübotter.

Less overlap from redundant information

By contrast, the most common method used to date for selecting the information suitable for the answer, known as the nearest neighbor method, tends to accumulate redundant information that is widely available. The difference between the two methods becomes clear when looking at an example of a query prompt that is composed of several pieces of information.

To answer the two-part question "How old is Roger Federer and how many children does he have?" the nearest neighbor method considers similar information such as "Roger Federer is 43 years old" and "Roger Federer's birthday is 8 August 1981" to be equally relevant.

Information about his children, which is relevant for the second part of the question, is sometimes missing. It is overlaid by birth date information, which occurs much more frequently in the AI training data.

The SIFT algorithm, however, takes into account the extent to which the pieces of information included complement each other, i.e. whether the information vectors point in different directions. This allows relevant information to be identified for both aspects of the question.

More reliable answers with much smaller models

However, targeted information selection not only improves the quality of responses. It can also be used to reduce the ever-increasing computing power required by AI applications.

By indirectly measuring uncertainty, the model can decide for itself how much more data is needed to provide a sufficiently reliable answer. Consequently, the computational overhead required by an LLM can be systematically adapted to the complexity of the question and the availability of relevant information.

Since SIFT continuously adapts the weighting of the arrow directions to its calculations during data retrieval, the enriched model becomes increasingly reliable the more it is used. This is known as test-time training and can be used to achieve the same output performance with smaller models.

"In tests with standard data sets, we used SIFT tuning to outperform even the best current AI models with models up to 40 times smaller," emphasizes Hübotter.

Identifying added value of relevant data

Additional applications for the SIFT algorithm are opening up in terms of data evaluation. As Krause explains, "We can track which enrichment data SIFT selects. They are closely related to the question and therefore particularly relevant to this subject area. This could be used in medicine, for example, to investigate which laboratory analyses or measurement values are significant for a specific diagnosis and which are less so."

Hübotter is presenting his approach at the International Conference on Learning Representations (ICLR) in Singapore. In December, the ETH researchers won the prize for the Best Scientific Article for their method at the NeurIPS Annual Conference on Neural Information Processing Systems (NeurIPS) in the "Finetuning in Modern Machine Learning" workshop.

More information: Jonas Hübotter et al, Efficiently Learning at Test-Time: Active Fine-Tuning of LLMs, *arXiv* (2024). DOI: 10.48550/arxiv.2410.08020

Provided by ETH Zurich

Citation: Algorithm can make AI responses increasingly reliable with less computational overhead (2025, April 24) retrieved 14 August 2025 from https://techxplore.com/news/2025-04-algorithm-ai-responses-reliable-overhead.html