Falko Baustian

SPOC-AI: How large language models change the future of document processing

We encounter documents, whether physical or digital, on a daily basis in the form of emails, letters, articles, applications, invoices and many others. Documents contain a lot of information conveyed not only through text but also through tables, diagrams, and images. Hence, document understanding and processing are essential tasks that occur regularly and offer significant potential for automation. But since documents exist in various formats and are often not standardized, automatic document processing is rather challenging. The objective of the SPOC-AI project is the development of AI-powered assistance systems for automatic processing of complex, document-based workflows. The involved project partners are the Computational Intelligence Lab (CITlab) at the University of Rostock, the research-driven company Planet AI, and the technology agency Altow, all based in Rostock. The CITlab team studies fundamental concepts and methods connected to Large Language Models (LLMs) within the context of document processing, which also hold relevance and applicability across other domains. The research focuses on three key topics: hallucination detection, parameter-efficient fine-tuning, and the effective utilization of external resources.

Hallucination detection

Not every answer generated by a language model is reliable. Sometimes, the models produce responses that sound plausible and convincing but are factually inaccurate. Such generations are referred to as AI hallucinations. The main causes of hallucinations are incomplete, inconsistent, or even incorrect information in the training data, leading to insufficient representations within the model parameters, and a training paradigm that requires the model to always generate an answer, even when uncertain. Hallucinations pose a significant challenge in LLM applications, particularly in critical domains such as medicine, where factual accuracy is essential. In agent-driven assistance systems, errors caused by hallucinations can propagate and amplify, potentially leading to flawed decisions, compromised safety, and reduced user trust.

Dr. Falko Baustian
Institut für Mathematik
Universität Rostock
Haus 3, Raum 226
Ulmenstraße 69, 18057 Rostock
falko.baustian@uni-rostock.de
https://www.mathematik.uni-rostock.de/unser-institut/mitarbeitende/baustian/

DOI-Nr.: 10.26125/ngzz-as15

A promising approach for hallucination detection in LLMs is uncertainty quantification, which leverages additional information present in the model's output. To understand these methods, it is helpful to first grasp how the models process input prompts and generate their output. Textual input is first transformed into a sequence of tokens, each representing a word, subword, or punctuation mark, and these tokens are then embedded into a vector space, where semantic and syntactic relationships are captured through numerical representations. For images, similar methods exist that decompose visual data into tokens and embed them into a vector space. An LLM consists of multiple layers that progressively transform embedded vectors, resulting in a vector representing the probability distribution of the next token over all tokens in the vocabulary. Output generation is determined by selecting the most probable token at each step. For uncertainty quantification, the probabilities are used to derive scores that assess the model's confidence in its outputs and to detect potential hallucinations. The methods use either the probability of the generated tokens, known as log-probabilities (logprobs), or the entropy of the full probability distribution, as defined in information theory. More sophisticated scores additionally consider the relevance of each token in relation to the output sentence.



Fig. 1: Possible relevance weights for tokens in a sequence. The tokens <capital>, <France> and <Paris> clearly define the meaning of the sentence shown by the comparatively high relevance. Tokens like <The>, <of>, <is> and <.> do carry less semantic weight.

Parameter-efficient fine-tuning

The rise of new-generation AI models began with the introduction of the attention mechanism in the seminal paper "Attention is all you need" in 2017 [1], reaching widespread impact with the 2022 launch of ChatGPT. In the attention layers of a model, the input is not processed sequentially but rather in parallel, enabling the model to efficiently capture global contextual relationships. Another key factor behind the leap in model capabilities is the growing depth of architectures and the resulting substantial increase in numbers of parameters, supported by rapid advancements in computing hardware. LLMs with parameter counts in the several billions require an extensive corpus of data for effective training. The cost of training new frontier models from scratch is substantial, often reaching millions of dollars due to the need for massive computational resources, extensive datasets, and significant

energy consumption. As a result, such training efforts are financially infeasible for many organizations, particularly in academia. Fine-tuning pre-trained foundation models to optimize them for specific tasks and domains offer a cost-effective and more sustainable alternative. In assistance systems, agents can perform diverse tasks by leveraging specialized versions of the same LLM, each fine-tuned for a specific task.

The core idea of parameter-efficient fine-tuning (PEFT) is to preserve the original parameters of the pre-trained LLM while training only a small set of additional parameters. Most computations in AI models are performed through matrix multiplication involving the model's learned weight matrices. To these matrices, additional parameters are added for example in the form of a new matrix that is parameterized via low-rank factorization - specifically, as the product of two smaller matrices - thereby significantly reducing the overall parameter count [2]. During fine-tuning, only the newly added parameters are updated using the task- or domain-specific dataset, while the original model weights remain fixed. Empirical studies and the common application of PEFT methods in practice show that task- or domain-specific performance can be notably increased by fine-tuning with a fraction of the parameters, often achieving performance comparable to full fine-tuning while reducing memory consumption, training time, and computational overhead by orders of magnitude. The additional weights can be saved as so-called adapters and activated on demand, enabling a single model to support multiple specializations through corresponding adapters. PEFT methods also reduce the loss of previously learned knowledge known as catastrophic forgetting during the fine-tuning process.

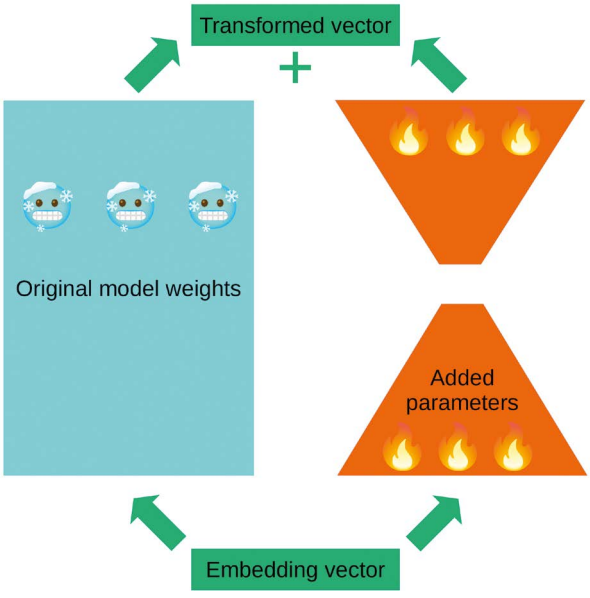


Fig. 2: Visualization of the fine-tuning approach. Original model weights remain fixed while only a small amount of added parameters are trained.

Effective utilization of external resources

When using your own documents or data sources with LLMs, long documents cannot be included directly in the prompt since each models input is restricted by a fixed context window limit. One could fine-tune the model on specific documents, but this is not always feasible since it still requires significant computational resources, data preparation, and specialized expertise. A common alternative is Retrieval-Augmented Generation (RAG), which addresses the limitations of fixed context windows by retrieving relevant information from external knowledge sources at runtime. The provided documents are split into chunks and embedded into a semantic vector space. When a user submits a query, the prompt is similarly embedded and compared against the document vectors using cosine similarity to identify the most relevant chunks. These chunks are then appended to the original prompt, enriching it with relevant information before being processed by the LLM to provide an answer. RAG is a widely used method when integrating documents and databases into assistance systems. The approach excels at handling straightforward retrieval-based queries but also has its limitations. The model's fixed context window restricts the number of chunks used per request making it difficult to address questions that require complex or distributed information across large documents.

Graph-based RAG variants enhance text chunk embeddings by simultaneously extracting entities and their semantic relationships by using LLMs, enabling richer, more structured knowledge representation. A subsequent entity resolution step merges matching entities across different chunks. The resulting graph can be stored in a graph database, allowing information retrieval through specialized query languages or agent-driven frameworks. In cooperation with the Department of Systems Biology and Bioinformatics at the University of Rostock, the CITlab team employs the graph-enhanced RAG approach to partially automate the generation of disease maps and their corresponding submaps. Disease maps visually represent complex relationships among genes, proteins, metabolites, and phenotypes, typically constructed manually by systems biologists based on information extracted from scientific literature. The automated pipeline takes PubMed articles as input, performs text chunking, and then employs an LLM to extract genes, proteins, metabolites, phenotypes, and their associated relationships. A normalization step maps the identified entities to standardized databases, among them NCBI Gene, UniProt and MeSH, resolving synonymous notations and simplifying the following entity resolution. The resulting knowledge graph is adapted to systems biology graphical notation (SBGN) and can be directly utilized to generate disease maps and submaps using standard tools. In addition, the pipeline is integrated into an assistance system that uses LLM agents to retrieve information from the graph. The agents can utilize predefined tools that execute fixed queries in a graph query language, or they can employ LLMs to generate sophisticated, dynamic queries as needed.

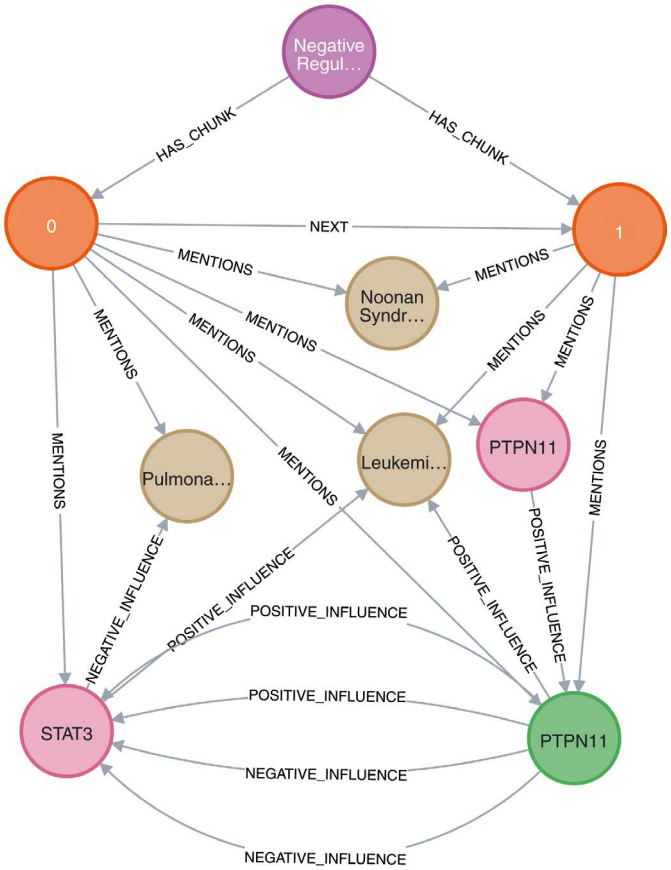


Fig. 3: Knowledge graph with relations between genes (pink), proteins (green), and metabolites (brown) extracted from chunks (orange) of a PubMed article (purple).

A wide range of use cases for the graph-based RAG framework are currently being explored at CITlab, including industrial applications with Planet AI, industry foundation classes for buildings with the Chair of Marine Data Science, and species interactions with the Chair of Aquaculture and Sea-Ranching, both at the University of Rostock, and medical studies with the Medical Informatics Laboratory at the University Medicine Greifswald.

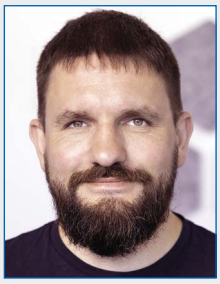
The project SPOC-AI is partially funded by the European Union through the Technologie-Beratungs-Institut Mecklenburg-Vorpommern.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin (2017): "Attention is all you need". *Advances in Neural Information Processing Systems*, **30**.

[2] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen (2021): "LoRA: Lowrank Adaptation of Large Language Models". *International Conference on Learning Representations*.

Dr. Falko Baustian



Falko Baustian studied mathematics at the University of Rostock. In his doctoral research and as a postdoc, he studied regularity properties of solutions of partial differential equations and their applications in mathematical finance. Later, he joined the company Planet AI as an AI researcher. In 2025, he returned to the University of Rostock as junior research group lead for the SPOC-AI project for AI-driven automation of document processing at the Computational Intelligence Lab.

Join yPC and shape the future of the Physical Chemistry Community!

Do you like to **work in a team**?
Are you **creative** or like to **connect to others**?
Then join the **young Physical Chemists**!

Our goals:
Connect early career scientists!
Represent their interests in and outside DBG!
Support them in their career development!

Interested? Get in contact!
ypc@go.bunsen.de

young PHYSICAL CHEMISTS
yPC