

Trustworthy and Efficient LLMs Meet Databases

Kyoungmin Kim
kyoung-min.kim@epfl.ch
EPFL
Switzerland

Anastasia Ailamaki
anastasia.ailamaki@epfl.ch
EPFL
Switzerland

Abstract

In the rapidly evolving AI era with large language models (LLMs) at the core, making LLMs more trustworthy and efficient, especially in output generation (inference), has gained significant attention. This is to reduce plausible but faulty LLM outputs (a.k.a hallucinations) and meet the highly increased inference demands. This tutorial explores such efforts and makes them transparent to the database community. Understanding these efforts is essential in harnessing LLMs in database tasks and adapting database techniques to LLMs. Furthermore, we delve into the synergy between LLMs and databases, highlighting new opportunities and challenges in their intersection. This tutorial aims to share with database researchers and practitioners essential concepts and strategies around LLMs, reduce the unfamiliarity of LLMs, and inspire joining in the intersection between LLMs and databases.

1 Introduction

Large language models (LLMs) have recently transformed various fields with their ability to understand and generate human-like text. In the database domain, researchers are leveraging LLMs to tackle complex data management tasks [55, 194]. LLMs can function not only as assistants for database administrators (DBAs) [271, 340] but also as internal components of database systems, optimizing query plans [8, 168] and translating natural languages to SQLs [224].

Beyond these applications, key concepts and advancements from the LLM community remain underexplored by database researchers. This tutorial aims to bridge that gap by focusing on enhancing the trustworthiness and efficiency of LLMs. Improving trustworthiness involves reducing hallucinations [124] to ensure LLMs generate accurate, factual responses, thereby increasing their reliability in database tasks requiring precise answers and reasoning. Enhancing efficiency focuses on decreasing inference latency and boosting throughput.

Inference efficiency is particularly important because, while training LLMs demands substantial resources and expertise, inference occurs daily across numerous users, leading to significant operational costs. For instance, OpenAI handles millions of requests, incurring substantial monthly expenses to run ChatGPT [73, 215]. Integrating LLMs with external data sources, such as vector databases and document retrieval systems in retrieval-augmented generation

(RAG) [154], increases the number and complexity of LLM calls, especially with longer inputs. Recent trends in chain-of-thought and multi-path reasoning, exemplified by models like OpenAI's o1 [205], further amplify inference demands, as generating final answers may require multiple LLM calls to enhance trustworthiness.

From a systems perspective, improving LLM inference efficiency parallels database management system (DBMS) development, presenting opportunities for database researchers to contribute to creating more efficient LLMs, promoting economic and environmental sustainability by reducing the CO2 footprint associated with extensive GPU usage.

After introducing the essential ideas in making LLMs more trustworthy and efficient, we will explain the intersection of LLMs and databases with new challenges and opportunities.

1.1 Target Audience and Prerequisites

Our tutorial is designed for conference attendees, focusing on three key areas to maximize engagement:

Trustworthy LLMs (Section 2.1): Aimed at individuals seeking to effectively utilize large language models (LLMs) in database tasks with minimal errors. Prerequisites include experience with LLMs like ChatGPT and the distinction between training and inference in machine learning. No in-depth knowledge of LLM internals is required.

Efficient LLMs (Section 2.2): Targeted at those interested in enhancing LLM inference efficiency or contributing to the development of fast LLM inference systems by applying database techniques. Prerequisites include basic database knowledge and an understanding of GPUs. Familiarity with Transformer architecture, attention mechanisms, and key-value (KV) caching is advantageous.

LLMs Meet Databases (Section 2.3): Intended for participants exploring new research opportunities at the intersection of databases and LLMs. A background in databases, including OLAP, relational algebra, cost-based query optimization, and approximate/adaptive query processing, will be helpful.

Our goal is to bridge the gap between essential LLM knowledge and the database community, enabling researchers already utilizing LLMs to uncover and develop unexplored ideas. Rather than merely listing state-of-the-art papers, we employ consistent visuals and focus on core concepts and insights, facilitating a deeper understanding and navigation of the evolving LLM landscape.

1.2 Tutorial Length

The intended length of this tutorial is 1.5 hour, with 40, 30, and 20 minutes each for Sections 2.1, 2.2, and 2.3, respectively.

2 Tutorial Outline

The tutorial is structured into three main sections, addressing critical aspects of LLMs and their interplay with database systems.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

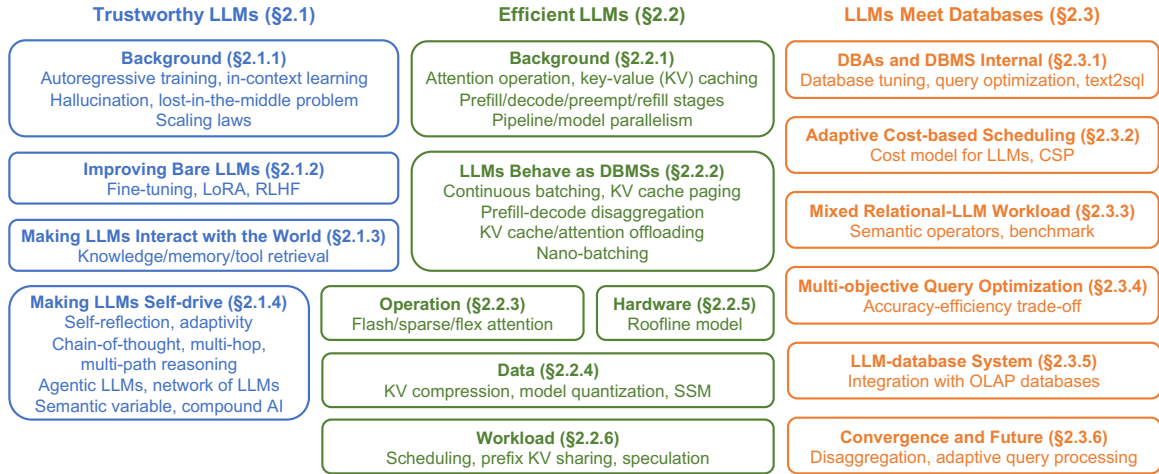


Figure 1: Tutorial outline (each subsection with keywords).

Figure 1 visualizes the outline with keywords for each subsection. Section 2.1 focuses on improving the trustworthiness of LLMs, exploring challenges such as hallucination and context limitations while presenting state-of-the-art solutions to improve the accuracy and reliability of generated outputs. Section 2.2 emphasizes efficiency, covering optimization strategies for inference, data management, and hardware utilization. Finally, Section 2.3 highlights the convergence of LLMs and databases, exploring opportunities for integration, new workloads, and emerging system designs. Since the field is changing fast, we will regularly reflect new information until the tutorial date.

2.1 Trustworthy LLMs

The first part of the tutorial explains the efforts to reduce hallucinations and make LLMs more trustworthy, using an analogy that LLMs resemble humans. We explain background (Section 2.1.1), how LLMs can solely improve (Section 2.1.2), how LLMs can improve by interacting with the external world (Section 2.1.3), and how LLMs can automatically make such decisions and interact with other LLMs (Section 2.1.4).

2.1.1 Background. Large Language Models (LLMs) function as text-in, text-out systems, generating texts based on their training. Training an LLM is akin to nurturing a child: by exposing it to extensive text data, the model acquires world knowledge and reasoning abilities. This process involves predicting the most probable next token in a sequence, a type of self-supervised learning. For a sequence of tokens, the model learns to predict the latter tokens based on the preceding ones, enabling it to generate coherent text continuations.

Fine-tuning refines this process for specific tasks or domains, similar to how individuals specialize in particular professions. In contrast, in-context learning provides additional information or examples within the input without altering the model’s parameters, akin to consulting external references during an open-book exam. Many prompting techniques [19, 34, 112, 240, 247, 258, 275, 319] including chain-of-thought prompting [144, 289] and its variants [18, 312] may leverage in-context learning to enhance performance.

During inference, LLMs generate texts autoregressively, producing one token at a time. This process may involve deterministic methods like greedy or beam search, or probabilistic approaches such as nucleus sampling [80, 114], which helps avoid selecting low-probability tokens.

However, LLMs experience hallucinations [13, 124, 270, 305], generating plausible-sounding but incorrect or fabricated information. This is an unavoidable aspect of LLMs [13, 305] which arises from limitations in capturing real-world knowledge, inherent approximations in training and inference, input noise, etc. Even slight input perturbations can significantly influence hallucinations [65, 101], and the detection of hallucinations has been a major problem [47, 54, 77, 195, 204, 233, 264].

Additionally, the lost-in-the-middle problem [117, 181] indicates that LLMs may struggle to utilize information located in the middle of long contexts, often performing better when relevant information is at the beginning or end of the input, exhibiting a U-shaped performance curve. This phenomenon has been attributed to inherent attention biases within LLMs, where tokens at the start and end of the input receive higher attention, regardless of their relevance [117]. This tendency can lead to increased hallucinations as context lengthens [230].

Scaling laws [113, 130, 221] explain that error rates decrease as model size and training data increase, with optimal scaling requiring proportional growth in both [113]. However, this may not hold for smaller models [221]. Laws can also relate to temporal loss in the training curve [297], downstream tasks [121], model quantization [306], transfer learning [14, 111], number of generated samples [27], and inference time [205] with the advance of using long, complex reasoning paths. Due to automatic prompting techniques [43, 136, 255, 302, 330] and that larger models tend to be less sensitive to prompt variations [75], we focus less on prompting techniques.

Target. The audience will distinguish pre-training, fine-tuning, and in-context learning phases of LLMs, and understand the inherent challenges in making LLMs trustworthy.

2.1.2 Improving Bare LLMs. We briefly explain the approaches to improve the LLM itself to make it more trustworthy. Since LLM

is a specific class of machine learning (ML) models, general ML approaches to enhance accuracy may work for LLMs. However, as such approaches have been extensively studied from the classic ML era, we target more LLM-specific approaches.

As it is infeasible to increase the model size indefinitely, and the models typically follow the Transformer architecture [274], efforts have been put to increase or augment training data (where LLMs themselves can be used to generate data) [36, 64, 82, 88, 142, 157, 163, 235, 241, 279, 341], improve data quality (again, LLMs can be used to clean data) [23, 66, 72, 103, 143, 328], make inferences more robust [91, 276], and apply better training and fine-tuning methods.

Specifically, fine-tuning covers a broad spectrum of work for example, parameter-efficient fine-tuning (PEFT) [116, 119, 152, 162, 222], instruction tuning [51, 198, 199, 243, 285, 288], reinforcement learning from human feedback (RLHF) [30, 52, 57, 81, 102, 134, 167, 208, 250, 262], and direct preference optimization (DPO) [83, 138, 234, 338]. RLHF leverages human feedback to train a reward model in reinforcement learning (RL), guiding the LLM through RL to produce desired outputs. DPO simplifies the alignment process by directly optimizing the policy model without a separate reward model. While these approaches rely on RL that continuously interacts with human or the world external to LLMs, such interactions are often limited to training and do not occur in inferences, which we explain in the following sections.

Other than the training methods, a new model architecture of differential Transformer [315] reduces the distractions of the models to focus on unnecessary information in the long context, which works similarly to robust decoding strategies [91, 276].

Target. The audience will learn about tuning LLMs to make them more trustworthy and aligned with user intentions.

2.1.3 Making LLMs Interact with the World: Adding Eyes and Hands. LLMs alone can encounter knowledge, memory, and capability limitations [326]. Their knowledge is confined to the static information encoded during training, leading to potential inaccuracies over time. Memory constraints arise from limited context windows, hindering the handling of extended conversations. Additionally, their text-based nature restricts interactions with the physical world. To address these challenges, LLMs can retrieve knowledge, memory, and tools.

This section focuses on *what* and *how* to retrieve. *When* to retrieve is the key to autonomy and will be detailed in the next section.

Knowledge retrieval is represented by well-known retrieval-augmented generation (RAG) [76, 90, 126, 154, 166]. Based on the data type, it can fetch knowledge from knowledge graphs [38, 109, 169, 211, 225, 245, 266, 301, 309], tables [9, 22, 40, 45, 96, 98, 115, 128, 150, 158, 190, 265, 320], images [41, 42, 314], not just documents. The data may be chunked/vectorized, stored in vector databases, then similar chunks are searched online. While vector similarities are typically used, more advanced similarity scores are possible, e.g., using dual or cross encoders [203, 239].

Memory retrieval attempts to overcome the limited context size of LLMs by storing previously seen tokens as key-value pairs [25, 183, 193, 281, 294] and fetching relevant pairs in upcoming requests, managing memory stores in hierarchical or partitioned way [145, 287] or even as a database [118]. Fetching information from long input can also be done without maintaining a separate

memory store, but by sparsifying the model layers [15, 186]. One can relate low-rank adapters and mixture-of-experts [29, 71, 78, 110, 119, 155, 228, 292] with memory retrieval since lightweight model parameters are fine-tuned per specific task and domain, and dynamically fetched at online inferences.

Tool retrieval searches for the APIs to interact with external environments [188, 197, 218, 229, 232, 246, 286, 300]. One can connect LLMs with databases to call SQLs that can help answering user questions [22]. Constrained decoding [20, 69, 92, 93] allows output to follow specific structure which can increase correctness and efficiency.

The challenges in retrieval include the followings. 1) Heterogeneity: LLMs are text-based, but knowledge can be of any type. Even for text retrieval, heterogeneous lengths and intents between queries and documents can lead to suboptimal retrieval accuracy [74, 89], and the vector-similarity search may be too simple to retrieve necessary information [87, 210]. 2) Scalability: Not only that LLMs have limited context or data they can utilize per inference, but maintaining a large set of retrieval entities and retrieving a subset may incur overheads [269, 303]. While approximation can mitigate the search overhead and make the search negligible to LLM inference costs, it is limited to vector-similarity search, and generalization to more complex searches [131, 135, 137, 239] remains challenging. 3) Sparsity: This is also relevant to data sparsity and noise [56], where relevant data is sparse compared to large information pools. 4) Reliability: Retrieved knowledge may be imperfect [277].

Target. The audience will understand how LLMs can interact with the world and exploit external knowledge to overcome the limits in using LLMs alone.

2.1.4 Making LLMs Self-drive: Adding Brain. Now we have more powerful models and interactions with the world. The last part is how we can make LLMs smart enough to maximize these capacities, adding autonomy. Self-consistency and major voting enables a simple yet effective solution for increasing consistency [283], however, it fails to generate accurate and diverse answers [33, 44, 46] and is yet passive. More active approaches include self-reflection and adaptive retrieval [11, 32, 123, 125, 159, 185, 331], which adaptively retrieves information multiple times based on the generated output, model confidence, query complexity, or fine-tuned policies. This is particularly helpful for chain-of-thought/multi-hop reasoning and question answering [175, 278, 282, 284, 329].

The next step is to use multiple reasoning paths instead of a single path. This multi-path reasoning has been an effective approach for driving LLMs [224, 226, 324, 332]. While the exact mechanism remains closed, OpenAI's o1 model is assumed to plan subtasks, conduct these, and revise the results to decide whether to extend the current plan or generate different plans, forming a tree-like reasoning structure. They suggest a new scaling law that LLM accuracy increases with inference time, not only with training time and data [205].

Agentic LLM indicates that LLMs can act as agents, selecting actions based on observations [49, 179, 256, 313]. Multiple agents exploit collaborative reasoning, parallel processing, diversity, and specialization akin to humans [35, 104, 172, 212, 214, 227, 299]. Semantic variables [173] regard LLM input and output tokens as dependent variables to explicitly model control flows.

A broader view includes compound AI [323] where AI and non-AI components interact with each other, including retrievals, control flows, agentic LLMs, and more. An interesting example is automated research process [187, 259].

Target. The audience will learn about approaches to make LLMs self-driving and build systems around LLMs for complex tasks.

2.2 Efficient LLMs

The second part of the tutorial demystifies the internals of LLM inference process and explains the efforts to make it more efficient, using an analogy that LLMs behave as DBMSs. We explain background (Section 2.2.1) and how LLM inference systems resemble DBMSs in improving their efficiency (Section 2.2.2). We then explain further work for each dimension of operation (Section 2.2.3), data (Section 2.2.4), hardware (Section 2.2.5), and workload (Section 2.2.6).

2.2.1 Background. The dominant Transformer architecture employs an attention mechanism [274] that calculates similarity scores between a token and its preceding tokens, effectively capturing inter-token relationships and managing extended contexts. This process has quadratic complexity, but key-value (KV) caching [58] optimizes it by storing and reusing these computations, reducing the complexity to linear during inference. Non-attention operations mostly consist of matrix multiplications and activations.

Inference in Large Language Models (LLMs) involves two primary phases: prefill and decode. During the prefill, the model processes input tokens to generate the initial output token. The attention operates with quadratic complexity due to the absence of precomputed KVs, making it compute-intensive. During the decode, the model generates subsequent tokens sequentially, each time using the last generated token as input. Here, the attention leverages KV caching, resulting in linear complexity relative to the number of processed tokens and reading their KVs, which makes this phase more memory-intensive.

In case of multiple requests, they face a race condition as in multi-tenant systems. If the GPU memory is insufficient to keep all requests' KVs, some running requests are preempted (evicted), releasing their KVs from the memory, and restarted (refilled) later [146]. Due to the low PCIe bandwidth, the released KVs are often recomputed when restarted, instead of offloading to other storage devices and loading back. Multiple requests in either prefill or decode steps can be batched to amortize the cost of loading model weights from GPU memory.

Note that the model weights also occupy the GPU memory. When the model size exceeds a single GPU capacity, techniques like model and pipeline parallelism [105, 120, 257] distribute model weights across multiple GPUs. This partitioning introduces data transfer overhead between GPUs.

Target. The audience will understand the KV caching and different phases of LLM inference requests, and how they compete for the same GPU resource.

2.2.2 LLM Inference Systems: LLMs Behave as DBMSs. LLM inference systems (e.g., vLLM [146]) behave similarly to (in-memory) DBMS. KVs and model weights correspond to the data, which are

maintained in GPU memory. Operators include matrix multiplications, activations, attentions, and data transfers. Compared to OLAP in databases, the operations are simpler yet much more time-sensitive, where the requests should be served in real-time.

Significant efforts have been made to increase the efficiency of LLM inference [342], largely based on operating and database systems. ORCA [318] forms a new batch of requests after each iteration (of prefills and decodes) whenever resources are available. Thus, a new request does not have to wait for all current running requests to finish, just like the pipelining in OS. vLLM [146] adopts paging and virtual memory to manage KVs, reducing memory fragmentation and enlarging the batch size. Since prefills are typically more costly than decodes, making stalls for decodes when batched together, SARATHI [4, 5] chunks prefills to reduce pipeline bubbles. Some other work [217, 263, 339] rather disaggregates the prefills and decodes into different GPUs, so the workload for each GPU is homogeneous. vTENSOR [298] decouples the KV cache management and attention computation of vLLM for better flexibility and performance. NANOFlow [343] splits each batch into nano-batches for finer-grained pipelining, increasing the overlap of computation, memory operation, and data transfer between GPUs. It also hides CPU scheduling latency by asynchronous scheduling. INFINIGEN [149] offloads the KVs to CPU memory to extend the KV cache and reloads the KVs from CPU layer-wise, but fetches a subset of KVs for efficiency, similarly to sparse attentions (Section 2.2.3). INSTINFERR [213] offloads KVs and attention computations to flash drives, just like the storage-disaggregation and computation pushdown in databases [310]. NEO [127] selectively offloads attention computations and KVs from GPU to CPU, in order to maximize both GPU and CPU utilization.

Target. The audience will understand why LLMs behave similarly to DBMSs and how database techniques can improve LLM inference efficiency. In subsequent sections, the audience will learn about efforts and challenges in further improving efficiency in four dimensions: operation data, hardware, and workload.

2.2.3 Operation: Attention. While matrix multiplications take the major portion in LLM latency in general [5], attentions can dominate the runtime for large inputs due to their quadratic complexity. FlashAttention [59, 60, 249] has become a de facto standard as an efficient attention implementation, utilizing recent GPU technologies to boost the inference speed. The ideas include kernel operator fusion and GPU cache-aware KV transfer. As in approximate query processing (AQP) in databases, sparse attentions [15, 178, 186] do not compute the full attention scores for all preceding tokens but a subset as an approximation. Some attentions rather optimize for long contexts [2, 62]. FlexAttention [107] offers flexible and performant implementation of such attentions.

2.2.4 Data: KV and Model Weights. Reading KVs from GPU memory in decode-attentions is similar to sequential table scan. As KVs are maintained per each attention layer, reading KVs for a layer can overlap with other layers' operators [149, 263]. While offloading KVs and attention computation have been popular recently [127, 149, 213], we need to be careful as it is challenging to predict the output lengths of LLM requests and thus their utilization patterns, and a KV for a single token may consume a few MBs. KVs of long documents can be precomputed, compressed, and fetched

for later retrievals [184]. To reduce memory latency, one can opt for KV sharing across different attention heads [7, 26, 50], KV compression [63, 129, 176, 184, 236], model quantization [94, 97, 106, 147, 164, 200, 251, 295, 306], or different model architectures than Transformer, such as State Space Models (SSMs) [61, 100] that do not rely on attentions, thereby not generating KVs. While hybrid architectures [10, 67, 99, 108, 171, 223, 237] can balance between the efficiency of SSMs and memorization capacity of Transformers, SSMs remain niche in the market [17]. A recent work even shows that tokenizers can be removed from the models [209].

2.2.5 Hardware: Theory and Practice. We briefly explain 1) the roofline model [322] and 2) some efforts to overcome the hardware limits [161, 238, 253, 261, 317, 333] or leverage advanced hardware for LLM inference [170, 304]. The roofline model is based on the computation speed (e.g., GPU FLOPS) and memory bandwidth, which acts as a theoretical hardware bound and determines whether an operator is compute-intensive or memory-intensive across different inputs.

2.2.6 Workload: Scheduling, Prefix Sharing, Speculation. To handle multiple LLM requests, LLM inference systems implement request scheduler to send LLM requests to appropriate machines or GPUs to maximize throughput or minimize latency. Assuming independent requests, early schedulers either prioritize prefills [146] or decodes [4], which tend to optimize latency or throughput, respectively. More complex schedulers consider fairness [252, 291] while compromising performance, or predict the output lengths of requests (not known in advance) and schedule shorter requests first [85, 231, 337].

If different LLM requests can share a prefix in their inputs, the KVs of the prefix can be stored just once and reused for multiple requests [334, 335]. This forms a trie structure with shared prefixes. However, a single-token difference in inputs may invalidate the sharing of KVs of all subsequent tokens. To increase the sharing opportunity, [311] uses the KVs of multiple token sequences to approximate the KVs of the concatenated sequence. The mechanism is similar to the speculation in OLAP [260] and healing protocol in transactions [293] in databases.

This speculation and healing patterns also appear in speculative decoding [153] and model cascades [39, 177, 316, 325, 336], accelerating the generation of tokens by leveraging smaller, faster models then validate the tokens using larger models, since the validation costs less than the generation.

2.3 LLMs Meet Databases

The last part of the tutorial discusses the intersection between LLMs and databases, opportunities and challenges in how we can exploit LLMs for databases, how the development of databases can help LLMs, and how we can exploit new types of workloads and integrations of LLMs and databases. We explain from more well-known to more untapped, deeper integrations in Sections 2.3.1-2.3.5 and provide more proactive visions in Section 2.3.6.

2.3.1 LLMs for DBs: DBAs and DBMS Internal. We briefly explain how LLMs are utilized for well-known tasks of DBAs and DBMS internals such as database tuning [271, 340], text2sql [151, 224] and query optimization [8, 168]. As we mentioned in Section 1, we will

not cover every detail, as many of these efforts are covered in a previous tutorial [194] and its additional list of papers [55].

2.3.2 DBs for LLMs: Adaptive Cost-based Scheduling. Unlike the sophisticated query optimizers in DBMSs, LLMs lack cost models and cost-based scheduling of LLM requests. [3] measures the batch times across various inputs (number of tokens to process and KV size to read). [322] computes batch times based on the roofline models. These can be used to model batch times and formulate the problem of finding optimal schedules as a constrained satisfaction problem (CSP) [139]. While schedulers try to avoid preemptions to maximize performance, [139] shows that harnessing preemptions can rather reduce overall latency compared to zero-preemptions. As the exact hardware utilization of each request is not known in advance, the scheduling should be adaptive based on the observations, and it has not been explored much to schedule dependent requests connected via semantic variables or shared prefixes [139].

2.3.3 DBs with LLMs: Mixed Relational-LLM Workload. Not only solving existing tasks with LLMs, LLMs offer new functionalities when integrated into DBMSs. Semantic operators [216] extend relational operators to batch-process the tabular data with LLMs (e.g., filters and joins using LLMs), which can be regarded as an AQP. Workloads with LLMs provide a justification to use LLMs inside DBMSs (heavy LLM calls in plan optimization can be negligible compared to query execution with LLMs). However, different pipelines (with semantic operators) lead to different accuracy and efficiency, thus defining the equivalence between two pipelines is non-trivial. Furthermore, more complex pipelines or LLM calls do not always guarantee higher accuracy [37, 74], and searching similar entities with LLMs can be replaced with efficient vector-similarity searches [177, 242] as a type of model cascade.

2.3.4 DBs with LLMs: Multi-objective Query Optimization and Benchmarks. The challenge is therefore how we can automatically find good pipelines for mixed relational-LLM workloads under the multi-objective of accuracy and efficiency [22, 273, 321] as in compound AI systems [95, 244, 267]. This calls for development of accurate cost models and accuracy-prediction models for LLMs and mixed relational-LLM workloads, in order to enable the holistic optimization of query plans consisting of both relational and non-relational operators. The cost model itself can be learned via LLMs (or any ML models), possibly using RLHF or feedback from query execution without human intervention, where such an automatic training data generation is one of the advantages of solving database tasks compared to conventional ML tasks (e.g., natural language processing with human-labeled translation data) [140]. Another model for predicting the output accuracy or detecting hallucination may be chosen from the scaling laws (using the general fact that larger models are more trustworthy) or separately trained.

To balance efficiency and accuracy, during the physical query optimization we should select proper models (ones used for execution) to avoid calling heavy LLMs unnecessarily. Depending on the complexity of the task, simple ML models with a small set of supervised data [123], or larger deep generative models such as in tabular foundation models tailored to domain-specific data [141, 160, 308], or LLMs with world knowledge and reasoning capacity [296] can fit the task with different accuracy-efficiency trade-offs.

Small language models (SLMs) [1, 16, 67, 189, 196, 202, 226] are also a good choice. Automatically finding the best prompt configuration [136, 280] tailored to the mixed workloads and more (e.g., previously mentioned fine-tuning or multi-hop/multi-path reasoning with adaptivity during inference) might be desired.

Furthermore, unlike the TPC benchmarks for databases, another problem is that there is no comprehensive benchmark for relational-LLM workloads yet. [22, 182] provide exploratory benchmarks without focusing on semantic joins.

2.3.5 DBs with LLMs: Integrated System. Other than the cost models, we also need DBMSs with native LLM support to increase the optimization opportunities, alike systems for relational-vector workloads [242, 327]. Current prototypes for relational-LLM workloads [177, 182, 216] separate table processing (e.g., pandas [192]) and LLM inference engine (e.g., vLLM [146]).

To maximize efficiency and scalability, we should focus on hardware utilization, data movement [122], caching hot data, locating computations close to data (e.g., computation pushdown in storage-aggregation setting) [84, 191], asynchronous API calls [95], balancing loads, and multi-tenancy just like in DBMSs [248, 310]. One also has to decide whether to maintain a separate vector database for faster online vector retrievals, or use just-in-time vectorization for reducing storage overhead. This also applies to precomputing KVs of data tokens [184] for faster LLM inferences or not, but with a higher caution as KVs are typically larger than vectors.

2.3.6 Convergence and Future. We envision LLMs and databases to converge (e.g., neuro-symbolic systems [48, 79, 268, 321]), more than just applying the techniques from one domain to another. A new LLM inference system optimized for DBMSs might be developed from an open-source cloud DBMS, utilizing recent implementations and optimizations for processing relational operators, such as storage-disaggregation and computation pushdown for scalable data and model management [310], GPU-based OLAP processing [86, 219, 220] for the full use of GPUs for both relational operators and LLMs, hybrid operators with heterogeneous data transfer paths [53, 310], adaptive query execution [307] and more. A unified query optimizer and data model for both relational data, KVs, and model weights, could offer opportunities for better data management and hardware utilization. Finally, if we look into the near future, we could also harness the emerging CXL technology for memory disaggregation [6, 148, 180] to manage model weights and KVs, and increased interest in pruning unnecessary data in OLAP [21, 24, 174, 206] could lead to higher trustworthiness (due to reduced noise) and efficiency (due to less data to process) in the relational-LLM workloads, with connections to online aggregation [254] and incremental view maintenance [28].

Target. The audience will understand the different depths of LLM-database integrations and be able to find interesting research topics from each of the integration, which are closely related to the current and near-future trends of databases.

3 Related Tutorial

Xupeng et al. [194] presented a tutorial at SIGMOD 2024 about the role of data management in the development (training, fine-tuning) and deployment (inference) of LLMs. It focused on how the

knowledge is encoded into model parameters and extracted during the inference, and explained the concept of KV caching but not LLM inference systems. Trummer [272] presented at VLDB 2023 about Transformer architecture, pre-training/fine-tuning/prompting in LLMs, and LLM applications in data management. As pointed out by [194], most of other tutorials presented at SIGMOD and VLDB about AI and databases focused on traditional machine learning and deep learning tasks not tailored to LLMs [31, 70, 156, 165, 201, 207, 290], or specific LLM-related applications such as tabular data understanding [12] and queries with natural languages [132, 133]. Dong et al. [68] presented at SIGKDD 2023 about the role of LLMs in building intelligent AR/VR assistants.

In this tutorial, we will focus on more recent, general approaches to enhance the trustworthiness and efficiency of LLMs, which have not been addressed in previous tutorials. For trustworthiness, we will start with enhancing LLMs alone, LLMs with tools, and agentic LLMs and collaboration. For efficiency, we will explain how LLM inference systems resemble DBMSs. Then we will discuss how we can integrate LLMs and databases in depth. We expect that these are what researchers, who aim to use LLMs in their applications or optimize LLMs using database techniques, need to know about. Instead of a common analogy that LLMs are knowledge bases as they generate plausible facts, we will use analogies that LLMs behave as DBMSs and improve as how humans solve challenging problems.

References

- [1] Marah I Abidin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Björck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *CoRR* abs/2404.14219 (2024). <https://doi.org/10.48550/ARXIV.2404.14219> arXiv:2404.14219
- [2] Amey Agrawal, Junda Chen, Íñigo Goiri, Ramachandran Ramjee, Chaojie Zhang, Alexey Tumanov, and Esha Choukse. 2024. Mnemosyne: Parallelization Strategies for Efficiently Serving Multi-Million Context Length LLM Inference Requests Without Approximations. *CoRR* abs/2409.17264 (2024). <https://doi.org/10.48550/ARXIV.2409.17264> arXiv:2409.17264
- [3] Amey Agrawal, Nitin Kedia, Jayashree Mohan, Ashish Panwar, Nipun Kwatra, Bhargav S. Gulavani, Ramachandran Ramjee, and Alexey Tumanov. 2024. VIDUR: A Large-Scale Simulation Framework for LLM Inference. In *Proceedings of the Seventh Annual Conference on Machine Learning and Systems, MLSys 2024, Santa Clara, CA, USA, May 13-16, 2024*, Phillip B. Gibbons, Gennady Pekhimenko, and Christopher De Sa (Eds.). mlsys.org. https://proceedings.mlsys.org/paper_files/paper/2024/hash/b74a8de47d2b3c928360e0a011f48351-Abstract-Conference.html
- [4] Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Alexey Tumanov, and Ramachandran Ramjee. 2024. Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve. In *18th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2024, Santa Clara, CA, USA, July 10-12, 2024*, Ada Gavrilovska and Douglas B. Terry (Eds.). USENIX Association, 117–134. <https://www.usenix.org/conference/osdi24/presentation/agrawal>
- [5] Amey Agrawal, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, and Ramachandran Ramjee. 2023. SARATHI: Efficient LLM Inference

- by Piggybacking Decodes with Chunked Prefixes. *CoRR* abs/2308.16369 (2023). <https://doi.org/10.48550/ARXIV.2308.16369> arXiv:2308.16369
- [6] Minseon Ahn, Thomas Willhalm, Norman May, Donghun Lee, Suprasad Mutalik Desai, Daniel Booss, Jungmin Kim, Navneet Singh, Daniel Ritter, and Oliver Reibholz. 2024. An Examination of CXL Memory Use Cases for In-Memory Database Management Systems using SAP HANA. *Proc. VLDB Endow.* 17, 12 (2024), 3827–3840. <https://www.vldb.org/pvldb/vol17/p3827-ahn.pdf>
 - [7] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 4895–4901. <https://doi.org/10.18653/V1/2023.EMNLP-MAIN.298>
 - [8] Peter Akioyamen, Zixuan Yi, and Ryan Marcus. 2024. The Unreasonable Effectiveness of LLMs for Query Optimization. *arXiv preprint arXiv:2411.02862* (2024).
 - [9] Uday Allu, Biddwan Ahmed, and Vishesh Tripathi. 2024. Beyond Extraction: Contextualising Tabular Data for Efficient Summarisation by Language Models. *CoRR* abs/2401.02333 (2024). <https://doi.org/10.48550/ARXIV.2401.02333> arXiv:2401.02333
 - [10] Quentin Anthony, Yury Tokpanov, Paolo Glorioso, and Beren Millidge. 2024. BlackMamba: Mixture of Experts for State-Space Models. *CoRR* abs/2402.01771 (2024). <https://doi.org/10.48550/ARXIV.2402.01771> arXiv:2402.01771
 - [11] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=hSyW5go0v8>
 - [12] Gilbert Badaro and Paolo Papotti. 2022. Transformers for Tabular Data Representation: A Tutorial on Models and Applications. *Proc. VLDB Endow.* 15, 12 (2022), 3746–3749. <https://doi.org/10.14778/3554821.3554890>
 - [13] Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. LLMs Will Always Hallucinate, and We Need to Live With This. *CoRR* abs/2409.05746 (2024). <https://doi.org/10.48550/ARXIV.2409.05746> arXiv:2409.05746
 - [14] Matthew Barnett. 2024. An Empirical Study of Scaling Laws for Transfer. *CoRR* abs/2408.16947 (2024). <https://doi.org/10.48550/ARXIV.2408.16947> arXiv:2408.16947
 - [15] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *CoRR* abs/2004.05150 (2020). arXiv:2004.05150 <https://arxiv.org/abs/2004.05150>
 - [16] Loubna Ben Allal, Anton Lozhkov, and Elie Bakouch. 2024. SmolLM - blazingly fast and remarkably powerful. <https://huggingface.co/blog/smolLM>. Accessed: December 15, 2024.
 - [17] Nathan Benaich and Ian Hogarth. 2024. State of AI Report 2024. <https://www.stateof.ai>.
 - [18] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 17682–17690. <https://doi.org/10.1609/AAAILV38I16.29720>
 - [19] Maciej Besta, Florim Memedi, Zhenyu Zhang, Robert Gerstenberger, Nils Blach, Piotr Nyczyk, Marcin Copik, Grzegorz Kwasniewski, Jürgen Müller, Lukas Gianinazzi, Ales Kubicek, Hubert Niewiadomski, Onur Mutlu, and Torsten Hoefler. 2024. Topologies of Reasoning: Demystifying Chains, Trees, and Graphs of Thoughts. *CoRR* abs/2401.14295 (2024). <https://doi.org/10.48550/ARXIV.2401.14295> arXiv:2401.14295
 - [20] Luca Beurer-Kellner, Marc Fischer, and Martin T. Vechev. 2024. Guiding LLMs The Right Way: Fast, Non-Invasive Constrained Generation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. <https://openreview.net/forum?id=pXaEYzrFae>
 - [21] Altan Birlir, Alfons Kemper, and Thomas Neumann. 2024. Robust Join Processing with Diamond Hardened Joins. *Proc. VLDB Endow.* 17, 11 (2024), 3215–3228. <https://www.vldb.org/pvldb/vol17/p3215-birlir.pdf>
 - [22] Asim Biswal, Liana Patel, Siddharth Jha, Amog Kamsetty, Shu Liu, Joseph E. Gonzalez, Carlos Guestrin, and Matei Zaharia. 2024. Text2SQL is Not Enough: Unifying AI and Databases with TAG. *CoRR* abs/2408.14717 (2024). <https://doi.org/10.48550/ARXIV.2408.14717> arXiv:2408.14717
 - [23] Quinten Bolding, Baohao Liao, Brandon James Denis, Jun Luo, and Christof Monz. 2023. Ask Language Model to Clean Your Noisy Translation Data. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 3215–3236. <https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.212>
 - [24] Angela Bonifati, Stefania Dumbrava, George Fletcher, Jan Hidders, Matthias Hofer, Wim Martens, Filip Murlak, Joshua Shinavier, Slawek Staworko, and Dominik Tomaszuk. 2023. Threshold Queries. *SIGMOD Rec.* 52, 1 (2023), 64–73. <https://doi.org/10.1145/3604437.3604452>
 - [25] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Casirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 2206–2240. <https://proceedings.mlr.press/v162/borgeaud22a.html>
 - [26] William Brandon, Mayank Mishra, Aniruddha Nrusimha, Rameswar Panda, and Jonathan Ragan-Kelley. 2024. Reducing Transformer Key-Value Cache Size with Cross-Layer Attention. *CoRR* abs/2405.12981 (2024). <https://doi.org/10.48550/ARXIV.2405.12981> arXiv:2405.12981
 - [27] Bradley C. A. Brown, Jordan Juravsky, Ryan Saul Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. *CoRR* abs/2407.21787 (2024). <https://doi.org/10.48550/ARXIV.2407.21787> arXiv:2407.21787
 - [28] Mihai Budiu, Tej Chajed, Frank McSherry, Leonid Ryzhyk, and Val Tannen. 2023. DBSP: Automatic Incremental View Maintenance for Rich Query Languages. *Proc. VLDB Endow.* 16, 7 (2023), 1601–1614. <https://doi.org/10.14778/3587136.3587137>
 - [29] Shiyi Cao, Shu Liu, Tyler Griggs, Peter Schafhalter, Xiaoxuan Liu, Ying Sheng, Joseph E. Gonzalez, Matei Zaharia, and Ion Stoica. 2024. MoE-Lightning: High-Throughput MoE Inference on Memory-constrained GPUs. *arXiv preprint arXiv:2411.11217* (2024).
 - [30] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphaël Ségerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krashennnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca D. Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Trans. Mach. Learn. Res.* 2023 (2023). <https://openreview.net/forum?id=bx24KpJ4Eb>
 - [31] Chengliang Chai, Nan Tang, Ju Fan, and Yuyu Luo. 2023. Demystifying Artificial Intelligence for Data Preparation. In *Companion of the 2023 International Conference on Management of Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023*, Sudipto Das, Ippokratis Pandis, K. Selçuk Candan, and Sihem Amer-Yahia (Eds.). ACM, 13–20. <https://doi.org/10.1145/3555041.3589406>
 - [32] Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. DUAL-REFLECT: Enhancing Large Language Models for Reflective Translation through Dual Learning Feedback Mechanisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 693–704. <https://aclanthology.org/2024.acl-short.64>
 - [33] Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R. Bowman, and Kyunghyun Cho. 2024. Two Failures of Self-Consistency in the Multi-Step Reasoning of LLMs. *Trans. Mach. Learn. Res.* 2024 (2024). <https://openreview.net/forum?id=5nBqY1y96B>
 - [34] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *CoRR* abs/2310.14735 (2023). <https://doi.org/10.48550/ARXIV.2310.14735> arXiv:2310.14735
 - [35] Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. 2023. Multi-Agent Consensus Seeking via Large Language Models. *CoRR* abs/2310.20151 (2023). <https://doi.org/10.48550/ARXIV.2310.20151> arXiv:2310.20151
 - [36] Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Trans. Assoc. Comput. Linguistics* 11 (2023), 191–211. https://doi.org/10.1162/TACL_A_00542
 - [37] Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. 2024. Are more llm calls all you need? towards scaling laws of compound inference systems. *arXiv preprint arXiv:2403.02419* (2024).
 - [38] Liyi Chen, Panrong Tong, Zhongming Jin, Ying Sun, Jieping Ye, and Hui Xiong. 2024. Plan-on-Graph: Self-Correcting Adaptive Planning of Large Language Model on Knowledge Graphs. *CoRR* abs/2410.23875 (2024). <https://doi.org/10.48550/ARXIV.2410.23875> arXiv:2410.23875
 - [39] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *CoRR* abs/2305.05176 (2023). <https://doi.org/10.48550/ARXIV.2305.05176>

arXiv:2305.05176

- [40] Wenhui Chen. 2023. Large Language Models are few(1)-shot Table Reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, 1090–1100. <https://doi.org/10.18653/V1/2023.FINDINGS-EACL.83>
- [41] Wenhui Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 5558–5570. <https://doi.org/10.18653/V1/2022.EMNLP-MAIN.375>
- [42] Wenhui Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. 2023. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/forum?id=XSEBx0iSjFQ>
- [43] Weizhe Chen, Sven Koenig, and Bistra Dilkina. 2024. RePrompt: Planning by Automatic Prompt Engineering for Large Language Models Agents. *CoRR abs/2406.11132* (2024). <https://doi.org/10.48550/ARXIV.2406.11132> arXiv:2406.11132
- [44] Wenqing Chen, Weicheng Wang, Zhixuan Chu, Kui Ren, Zibin Zheng, and Zhichao Lu. 2024. Self-Para-Consistency: Improving Reasoning Tasks at Low Cost for Large Language Models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 14162–14167. <https://doi.org/10.18653/V1/2024.FINDINGS-ACL.842>
- [45] Wenhui Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1026–1036. <https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.91>
- [46] Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal Self-Consistency for Large Language Model Generation. *CoRR abs/2311.17311* (2023). <https://doi.org/10.48550/ARXIV.2311.17311> arXiv:2311.17311
- [47] Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. Unified Hallucination Detection for Multimodal Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 3235–3252. <https://doi.org/10.18653/V1/2024.ACL-LONG.178>
- [48] Zui Chen, Zihui Gu, Lei Cao, Ju Fan, Samuel Madden, and Nan Tang. 2023. Symphony: Towards Natural Language Query Answering over Multi-modal Data Lakes. In *13th Conference on Innovative Data Systems Research, CIDR 2023, Amsterdam, The Netherlands, January 8-11, 2023*. www.cidrdb.org. <https://www.cidrdb.org/cidr2023/papers/p51-chen.pdf>
- [49] Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, and Xiuqiang He. 2024. Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects. *CoRR abs/2401.03428* (2024). <https://doi.org/10.48550/ARXIV.2401.03428> arXiv:2401.03428
- [50] Sai Sena Chinnakonduru and Astarag Mohapatra. 2024. Weighted Grouped Query Attention in Transformers. *CoRR abs/2407.10855* (2024). <https://doi.org/10.48550/ARXIV.2407.10855> arXiv:2407.10855
- [51] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 1931–1942. <http://proceedings.mlr.press/v139/cho21a.html>
- [52] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4299–4307. <https://proceedings.neurips.cc/paper/2017/hash/d5e2c0dad503c91f91df240d0cd4e49-Abstract.html>
- [53] Periklis Chrysogelos, Manos Karpasiotakis, Raja Appuswamy, and Anastasia Ailamaki. 2019. HetExchange: Encapsulating heterogeneous CPU-GPU parallelism in JIT compiled engines. *Proc. VLDB Endow.* 12, 5 (2019), 544–556. <https://doi.org/10.14778/3303753.3303760>
- [54] Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 1419–1436. <https://aclanthology.org/2024.emnlp-main.84>
- [55] code4DB. 2024. LLM4DB: A Curated List of Resources on Large Language Models for Databases. <https://github.com/code4DB/LLM4DB>. Accessed: 2024-11-30.
- [56] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonello, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 719–729. <https://doi.org/10.1145/3626772.3657834>
- [57] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe Reinforcement Learning from Human Feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=TyFrPOKYXw>
- [58] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2978–2988. <https://doi.org/10.18653/V1/P19-1285>
- [59] Tri Dao. 2024. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=mZn2Xyh9Ec>
- [60] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html
- [61] Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. <https://openreview.net/forum?id=ztn8FCR1td>
- [62] Tri Dao, Daniel Haziza, Francisco Massa, and Grigory Sizov. 2023. Flash-Decoding for long-context inference. <https://pytorch.org/blog/flash-decoding/>. Accessed: December 15, 2024.
- [63] Alessio Devoto, Yu Zhao, Simone Scardapane, and Pasquale Minervini. 2024. A Simple and Effective L₂ Norm-Based Strategy for KV Cache Compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 18476–18499. <https://aclanthology.org/2024.emnlp-main.1027>
- [64] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhao Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 1679–1705. <https://doi.org/10.18653/V1/2024.FINDINGS-ACL.97>
- [65] Peng Ding, Jingyu Wu, Jun Kuang, Dan Ma, Xuezhi Cao, Xunliang Cai, Shi Chen, Jiajun Chen, and Shujian Huang. 2024. Hallu-PI: Evaluating Hallucination in Multi-modal Large Language Models within Perturbed Inputs. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, Jianfei Cai, Mohan S. Kankanalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (Eds.). ACM, 10707–10715. <https://doi.org/10.1145/3664467.3681251>
- [66] Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 1286–1305. <https://doi.org/10.18653/V1/2021.EMNLP-MAIN.98>
- [67] Xin Dong, Yonggan Fu, Shizhe Diao, Womni Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen,

- Yoshi Suhara, et al. 2024. Hymba: A Hybrid-head Architecture for Small Language Models. *arXiv preprint arXiv:2411.13676* (2024).
- [68] Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. 2023. Towards Next-Generation Intelligent Assistants Leveraging LLM Techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (Eds.). ACM, 5792–5793. <https://doi.org/10.1145/3580305.3599572>
- [69] Yixin Dong, Charlie F Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2024. XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models. *arXiv preprint arXiv:2411.15100* (2024).
- [70] Alexey Drutsa, Valentina Fedorova, Dmitry Ustalov, Olga Megorskaya, Evfrosiniya Zermirnova, and Daria Baidakova. 2020. Crowdsourcing Practice for Efficient Data Labeling: Aggregation, Incremental Relabeling, and Pricing. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, David Maier, Rachel Pottinger, AnHai Doan, Wang-Chiew Tan, Abdussalam Alawini, and Hung Q. Ngo (Eds.). ACM, 2623–2627. <https://doi.org/10.1145/3318464.3383127>
- [71] Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen S. Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhiheng Chen, and Claire Cui. 2022. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 5547–5569. <https://proceedings.mlr.press/v162/du22c.html>
- [72] Yaxin Du, Rui Ye, Yuchi Fengting, Wanru Zhao, Jingjing Qu, Yanfeng Wang, and Siheng Chen. 2024. Data Quality Control in Federated Instruction-tuning of Large Language Models. *CoRR abs/2410.11540* (2024). <https://doi.org/10.48550/ARXIV.2410.11540> arXiv:2410.11540
- [73] Fabio Duarte. 2024. Number of ChatGPT Users (Dec 2024). <https://explodingtopics.com/blog/chatgpt-users>. Accessed: 2024-12-15.
- [74] Matous Eibich, Shivay Nagpal, and Alexander Fred-Ojala. 2024. ARAGOG: Advanced RAG Output Grading. *CoRR abs/2404.01037* (2024). <https://doi.org/10.48550/ARXIV.2404.01037> arXiv:2404.01037
- [75] Oluwale Fagbohun, Rachel M. Harrison, and Anton Dereventsov. 2024. An Empirical Categorization of Prompting Techniques for Large Language Models: A Practitioner’s Guide. *CoRR abs/2402.14837* (2024). <https://doi.org/10.48550/ARXIV.2402.14837> arXiv:2402.14837
- [76] Wenqi Fan, Yujian Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, Ricardo Baeza-Yates and Francesco Bonchi (Eds.). ACM, 6491–6501. <https://doi.org/10.1145/3637528.3671470>
- [77] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nat.* 630, 8017 (2024), 625–630. <https://doi.org/10.1038/S41586-024-07421-0>
- [78] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *J. Mach. Learn. Res.* 23 (2022), 120:1–120:39. <https://jmlr.org/papers/v23/21-0998.html>
- [79] Jonathan Feldstein, Paulius Dilkas, Vaishak Belle, and Efthymia Tsamoura. 2024. Mapping the Neuro-Symbolic AI Landscape by Architectures: A Handbook on Augmenting Deep Learning Through Symbolic Reasoning. *CoRR abs/2410.22077* (2024). <https://doi.org/10.48550/ARXIV.2410.22077> arXiv:2410.22077
- [80] Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. 2024. Closing the Curious Case of Neural Text Degeneration. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=dONpC9GL1o>
- [81] Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. How to Evaluate Reward Models for RLHF. *CoRR abs/2410.14872* (2024). <https://doi.org/10.48550/ARXIV.2410.14872> arXiv:2410.14872
- [82] Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Scott Yih, Luke Zettlemoyer, and Mike Lewis. 2023. InCoder: A Generative Model for Code Infilling and Synthesis. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/forum?id=hQwb-lbM6EL>
- [83] Yuhua Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. 2024. Mitigating Hallucination in Multimodal Large Language Model via Hallucination-targeted Direct Preference Optimization. *arXiv preprint arXiv:2411.10436* (2024).
- [84] Yao Fu, Leyang Xue, Yeqi Huang, Andrei-Octavian Brabete, Dmitrii Ustiugov, Yuvraj Patel, and Luo Mai. 2024. ServerlessLLM: Locality-Enhanced Serverless Inference for Large Language Models. *CoRR abs/2401.14351* (2024). <https://doi.org/10.48550/ARXIV.2401.14351> arXiv:2401.14351
- [85] Yichao Fu, Siqi Zhu, Runlong Su, Aurick Qiao, Ion Stoica, and Hao Zhang. 2024. Efficient LLM Scheduling by Learning to Rank. *CoRR abs/2408.15792* (2024). <https://doi.org/10.48550/ARXIV.2408.15792> arXiv:2408.15792
- [86] Henning Funke and Jens Teubner. 2020. Data-parallel query processing on non-uniform data. *Proceedings of the VLDB Endowment* 13, 6 (2020), 884–897.
- [87] Hang Gao and Yongfeng Zhang. 2024. VRSD: Rethinking Similarity and Diversity for Retrieval in Large Language Models. *CoRR abs/2407.04573* (2024). <https://doi.org/10.48550/ARXIV.2407.04573> arXiv:2407.04573
- [88] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *CoRR abs/2101.00027* (2021). arXiv:2101.00027 <https://arxiv.org/abs/2101.00027>
- [89] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 1762–1777. <https://doi.org/10.18653/V1/2023.ACL-LONG.99>
- [90] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *CoRR abs/2312.10997* (2023). <https://doi.org/10.48550/ARXIV.2312.10997> arXiv:2312.10997
- [91] Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2024. DeCoRe: Decoding by Contrasting Retrieval Heads to Mitigate Hallucinations. *arXiv preprint arXiv:2410.18860* (2024).
- [92] Saibo Geng, Berkay Döner, Chris Wendler, Martin Josifoski, and Robert West. 2024. Sketch-Guided Constrained Decoding for Boosting Blackbox Large Language Models without Logit Access. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 234–245. <https://aclanthology.org/2024.acl-short.23>
- [93] Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 10932–10952. <https://doi.org/10.18653/V1/2023.EMNLP-MAIN.674>
- [94] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2021. A Survey of Quantization Methods for Efficient Neural Network Inference. *CoRR abs/2103.13630* (2021). arXiv:2103.13630 <https://arxiv.org/abs/2103.13630>
- [95] In Gim, Seung-seob Lee, and Lin Zhong. 2024. Asynchronous LLM Function Calling. *arXiv preprint arXiv:2412.07017* (2024).
- [96] Michael R. Glass, Xueqing Wu, Ankita Rajaram Naik, Gaetano Rossiello, and Alfio Gliozzo. 2023. Retrieval-Based Transformer for Table Augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 5635–5648. <https://doi.org/10.18653/V1/2023.FINDINGS-ACL.348>
- [97] Ruihao Gong, Yang Yong, Shiqiao Gu, Yushi Huang, Chengtao Lv, Yunchen Zhang, Dacheng Tao, and Xianglong Liu. 2024. LLMC: Benchmarking Large Language Model Quantization with a Versatile Compression Toolkit. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024*, Franck Dernoncourt, Daniel Preoticiu-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, 132–152. <https://aclanthology.org/2024.emnlp-industry.12>
- [98] Yury Gorishniy, Ivan Rubachev, Nikolay Kartashev, Daniil Shlenskii, Akim Kotelnikov, and Artem Babenko. 2023. TabR: Unlocking the Power of Retrieval-Augmented Tabular Deep Learning. *CoRR abs/2307.14338* (2023). <https://doi.org/10.48550/ARXIV.2307.14338> arXiv:2307.14338
- [99] Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *CoRR abs/2312.00752* (2023). <https://doi.org/10.48550/ARXIV.2312.00752> arXiv:2312.00752
- [100] Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently Modeling Long Sequences with Structured State Spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=uYLFoz1vIAC>
- [101] Nuno Miguel Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. *Trans. Assoc. Comput. Linguistics* 11

- (2023), 1500–1517. https://doi.org/10.1162/TACL_A_00615
- [102] Çağlar Gülçehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced Self-Training (ReST) for Language Modeling. *CoRR abs/2308.08998* (2023). <https://doi.org/10.48550/ARXIV.2308.08998> arXiv:2308.08998
- [103] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks Are All You Need. *CoRR abs/2306.11644* (2023). <https://doi.org/10.48550/ARXIV.2306.11644> arXiv:2306.11644
- [104] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, 8048–8057. <https://www.ijcai.org/proceedings/2024/890>
- [105] Aaron Harlap, Deepak Narayanan, Amar Phanishayee, Vivek Seshadri, Nikhil R. Devanur, Gregory R. Ganger, and Phillip B. Gibbons. 2018. PipeDream: Fast and Efficient Pipeline Parallel DNN Training. *CoRR abs/1806.03377* (2018). arXiv:1806.03377 <http://arxiv.org/abs/1806.03377>
- [106] Jahid Hasan. 2024. Optimizing Large Language Models through Quantization: A Comparative Analysis of PTQ and QAT Techniques. *arXiv preprint arXiv:2411.06084* (2024).
- [107] Horace He, Driss Guessous, Yanbo Liang, and Joy Dong. 2024. FlexAttention: The Flexibility of PyTorch with the Performance of FlashAttention. <https://pytorch.org/blog/flexattention/>.
- [108] Wei He, Kai Han, Yehui Tang, Chengcheng Wang, Yujie Yang, Tianyu Guo, and Yunhe Wang. 2024. DenseMamba: State Space Models with Dense Hidden Connection for Efficient Large Language Models. *CoRR abs/2403.00818* (2024). <https://doi.org/10.48550/ARXIV.2403.00818> arXiv:2403.00818
- [109] Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. *CoRR abs/2402.07630* (2024). <https://doi.org/10.48550/ARXIV.2402.07630> arXiv:2402.07630
- [110] Xu Owen He. 2024. Mixture of A Million Experts. *CoRR abs/2407.04153* (2024). <https://doi.org/10.48550/ARXIV.2407.04153> arXiv:2407.04153
- [111] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling Laws for Transfer. *CoRR abs/2102.01293* (2021). arXiv:2102.01293 <https://arxiv.org/abs/2102.01293>
- [112] Michael Hewing and Vincent Leinhos. 2024. The Prompt Canvas: A Literature-Based Practitioner Guide for Creating Effective Prompts in Large Language Models. *arXiv preprint arXiv:2412.05127* (2024).
- [113] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. *CoRR abs/2203.15556* (2022). <https://doi.org/10.48550/ARXIV.2203.15556> arXiv:2203.15556
- [114] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. <https://openreview.net/forum?id=rygGQyrFvH>
- [115] Zijin Hong, Zheng Yuan, Hao Chen, Qinggang Zhang, Feiran Huang, and Xiao Huang. 2024. Knowledge-to-SQL: Enhancing SQL Generation with Data Expert LLM. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 10997–11008. <https://doi.org/10.18653/V1/2024.FINDINGS-ACL.653>
- [116] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2790–2799. <http://proceedings.mlr.press/v97/houlsby19a.html>
- [117] Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T. Le, Abhishek Kumar, James R. Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating Positional Attention Bias Improves Long Context Utilization. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 14982–14995. <https://doi.org/10.18653/V1/2024.FINDINGS-ACL.890>
- [118] Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. 2023. ChatDB: Augmenting LLMs with Databases as Their Symbolic Memory. *CoRR abs/2306.03901* (2023). <https://doi.org/10.48550/ARXIV.2306.03901> arXiv:2306.03901
- [119] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=nZvKeeFYf9>
- [120] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Xu Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. 2019. GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.), 103–112. <https://proceedings.neurips.cc/paper/2019/hash/093f65e080a295f8076b1c5722a46aa2-Abstract.html>
- [121] Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. 2024. Scaling Laws for Downstream Task Performance of Large Language Models. *CoRR abs/2402.04177* (2024). <https://doi.org/10.48550/ARXIV.2402.04177> arXiv:2402.04177
- [122] Andrei Ivanov, Nikoli Dryden, Tal Ben-Nun, Shigang Li, and Torsten Hoefler. 2021. Data Movement Is All You Need: A Case Study on Optimizing Transformers. In *Proceedings of the Fourth Conference on Machine Learning and Systems, MLSys 2021, virtual, April 5-9, 2021*, Alex Smola, Alex Dimakis, and Ion Stoica (Eds.). mlsys.org. https://proceedings.mlsys.org/paper_files/paper/2021/hash/bc86e95606a6392f51f95a8de106728d-Abstract.html
- [123] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (Eds.). Association for Computational Linguistics, 7036–7050. <https://doi.org/10.18653/V1/2024.NAACL-LONG.389>
- [124] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (2023), 248:1–248:38. <https://doi.org/10.1145/3571730>
- [125] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards Mitigating Hallucination in Large Language Models via Self-Reflection. *CoRR abs/2310.06271* (2023). <https://doi.org/10.48550/ARXIV.2310.06271> arXiv:2310.06271
- [126] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 9237–9251. <https://doi.org/10.18653/V1/2023.EMNLP-MAIN.574>
- [127] Xuanlin Jiang, Yang Zhou, Shiyi Cao, Ion Stoica, and Minlan Yu. 2024. NEO: Saving GPU Memory Crisis with CPU Offloading for Online LLM Inference. *arXiv preprint arXiv:2411.01142* (2024).
- [128] Deokhyung Kang, Baekjin Jung, Yunsu Kim, and Gary Geunbae Lee. 2024. Denoising Table-Text Retrieval for Open-Domain Question Answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakrani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, 4634–4640. <https://aclanthology.org/2024.lrec-main.414>
- [129] Hao Kang, Qingru Zhang, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, and Tuo Zhao. 2024. Gear: An efficient kv cache compression recipe for near-lossless generative inference of llm. *arXiv preprint arXiv:2403.05527* (2024).
- [130] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *CoRR abs/2001.08361* (2020). arXiv:2001.08361 <https://arxiv.org/abs/2001.08361>
- [131] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 6769–6781. <https://doi.org/10.18653/V1/2020.EMNLP-MAIN.550>
- [132] George Katsogiannis-Meimarakis and Georgia Koutrika. 2021. A Deep Dive into Deep Learning Approaches for Text-to-SQL Systems. In *SIGMOD ’21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*,

- Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava (Eds.). ACM, 2846–2851. <https://doi.org/10.1145/3448016.3457543>
- [133] George Katsogiannis-Meimarakis, Mike Xydias, and Georgia Koutrika. 2023. Natural Language Interfaces for Databases with Deep Learning. *Proc. VLDB Endow.* 16, 12 (2023), 3878–3881. <https://doi.org/10.14778/3611540.3611575>
- [134] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A Survey of Reinforcement Learning from Human Feedback. *CoRR* abs/2312.14925 (2023). <https://doi.org/10.48550/ARXIV.2312.14925> arXiv:2312.14925
- [135] Zixuan Ke, Weize Kong, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Bridging the Preference Gap between Retrievers and LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 10438–10451. <https://doi.org/10.18653/V1/2024.ACL-LONG.562>
- [136] Omar Khattab, Arnab Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=sY5N0zY5Od>
- [137] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 39–48. <https://doi.org/10.1145/3397271.3401075>
- [138] Dahyun Kim, Yungi Kim, Wonho Song, Hyeonwoo Kim, Yunsu Kim, Sanghoon Kim, and Chanjun Park. 2024. sDPO: Don't Use Your Data All at Once. *CoRR* abs/2403.19270 (2024). <https://doi.org/10.48550/ARXIV.2403.19270> arXiv:2403.19270
- [139] Kyoungmin Kim, Kijae Hong, Caglar Gulcehre, and Anastasia Ailamaki. 2024. The Effect of Scheduling and Preemption on the Efficiency of LLM Inference Serving. *arXiv preprint arXiv:2411.07447* (2024).
- [140] Kyoungmin Kim, Jisung Jung, In Seo, Wook-Shin Han, Kangwoo Choi, and Jaehyok Chong. 2022. Learned Cardinality Estimation: An In-depth Study. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary G. Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 1214–1227. <https://doi.org/10.1145/3514221.3526154>
- [141] Kyoungmin Kim, Sangoh Lee, Injung Kim, and Wook-Shin Han. 2024. ASM: Harmonizing Autoregressive Model, Sampling, and Multi-dimensional Statistics Merging for Cardinality Estimation. *Proc. ACM Manag. Data* 2, 1 (2024), 45:1–45:27. <https://doi.org/10.1145/3639300>
- [142] Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashevski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2024. Evaluating Language Models as Synthetic Data Generators. *arXiv preprint arXiv:2412.03679* (2024).
- [143] Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2023. The Stack: 3 TB of permissively licensed source code. *Trans. Mach. Learn. Res.* 2023 (2023). <https://openreview.net/forum?id=pxpbTtUEpD>
- [144] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html
- [145] Juri Kong, Hong Liang, Yuan Zhang, Hongxiang Li, Pengcheng Shen, and Fang Lu. 2024. Dynamic semantic memory retention in large language models: An exploration of spontaneous retrieval mechanisms. (2024).
- [146] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (Eds.). ACM, 611–626. <https://doi.org/10.1145/3600006.3613165>
- [147] Jiedong Lang, Zhehao Guo, and Shuyu Huang. 2024. A Comprehensive Study on Quantization Techniques for Large Language Models. *arXiv preprint arXiv:2411.02530* (2024).
- [148] Sangjin Lee, Alberto Lerner, Philippe Bonnet, and Philippe Cudré-Mauroux. 2024. Database Kernels: Seamless Integration of Database Systems and Fast Storage via CXL. In *14th Conference on Innovative Data Systems Research, CIDR 2024, Chaminade, HI, USA, January 14-17, 2024*. www.cidrdb.org. <https://www.cidrdb.org/cidr2024/papers/p43-lee.pdf>
- [149] Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim. 2024. InfiniGen: Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management. In *18th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2024, Santa Clara, CA, USA, July 10-12, 2024*, Ada Gavrilovska and Douglas B. Terry (Eds.). USENIX Association, 155–172. <https://www.usenix.org/conference/osdi24/presentation/lee>
- [150] Younghun Lee, Sungchul Kim, Tong Yu, Ryan A. Rossi, and Xiang Chen. 2024. Learning to Reduce: Optimal Representations of Structured Data in Prompting Large Language Models. *CoRR* abs/2402.14195 (2024). <https://doi.org/10.48550/ARXIV.2402.14195> arXiv:2402.14195
- [151] Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin Su, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, et al. 2024. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. *arXiv preprint arXiv:2411.07763* (2024).
- [152] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 3045–3059. <https://doi.org/10.18653/V1/2021.EMNLP-MAIN.243>
- [153] Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast Inference from Transformers via Speculative Decoding. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 19274–19286. <https://proceedings.mlr.press/v202/leviathan23a.html>
- [154] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [155] Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024. MixLoRA: Enhancing Large Language Models Fine-Tuning with LoRA based Mixture of Experts. *CoRR* abs/2404.15159 (2024). <https://doi.org/10.48550/ARXIV.2404.15159> arXiv:2404.15159
- [156] Guoliang Li, Xuanhe Zhou, and Lei Cao. 2021. AI Meets Database: AI4DB and DB4AI. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava (Eds.). ACM, 2859–2866. <https://doi.org/10.1145/3448016.3457542>
- [157] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruva Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah M. Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Raghavi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujoy Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. 2024. DataComp-LM: In search of the next generation of training sets for language models. *CoRR* abs/2406.11794 (2024). <https://doi.org/10.48550/ARXIV.2406.11794> arXiv:2406.11794
- [158] Jinyang Li, Binyuan Hui, Ge Qu, Jiayi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin Chen-Chuan Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. Can LLM Already Serve as A Database Interface? A Blg Bench for Large-Scale Database Grounded Text-to-SQLs. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/83fc8fab1710363050bbd1d4b8cc0021-Abstract-Datasets_and_Benchmarks.html
- [159] Lingyu Li, Yixu Wang, Haiquan Zhao, Shuqi Kong, Yan Teng, Chunbo Li, and Yingchun Wang. 2024. Reflection-Bench: probing AI intelligence with reflection. *arXiv preprint arXiv:2410.16270* (2024).
- [160] Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2024. TableGPT: Table Fine-tuned GPT for Diverse Table Tasks. *Proc. ACM Manag. Data* 2, 3 (2024), 176. <https://doi.org/10.1145/3654979>

- [161] Xiangyu Li, Yuanchun Li, Yuanzhe Li, Ting Cao, and Yunxin Liu. 2024. FlexNN: Efficient and Adaptive DNN Inference on Memory-Constrained Edge Devices. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, ACM MobiCom 2024, Washington D.C., DC, USA, November 18-22, 2024*, Weisong Shi, Deepak Ganesan, and Nicholas D. Lane (Eds.). ACM, 709–723. <https://doi.org/10.1145/3636534.3649391>
- [162] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 4582–4597. <https://doi.org/10.18653/V1/2021.ACL-LONG.353>
- [163] Yichuan Li, Kaize Ding, Jianling Wang, and Kyumin Lee. 2024. Empowering Large Language Models for Textual Data Augmentation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 12734–12751. <https://doi.org/10.18653/V1/2024.FINDINGS-ACL.756>
- [164] Yuhang Li, Mingzhu Shen, Jian Ma, Yan Ren, Mingxin Zhao, Qi Zhang, Ruihao Gong, Fengwei Yu, and Junjie Yan. 2021. MQBench: Towards Reproducible and Deployable Model Quantization Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c20ad4d76fe97759aa27a0c99b6f710-Abstract-round1.html>
- [165] Yuliang Li, Xiaolan Wang, Zhengjie Miao, and Wang-Chiew Tan. 2021. Data Augmentation for ML-driven Data Preparation and Integration. *Proc. VLDB Endow.* 14, 12 (2021), 3182–3185. <https://doi.org/10.14778/3476311.3476403>
- [166] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024*, Franck Dernoncourt, Daniel Preotiu-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, 881–893. <https://aclanthology.org/2024.emnlp-industry.66>
- [167] Zihao Li, Zhuoran Yang, and Mengdi Wang. 2023. Reinforcement Learning with Human Feedback: Learning Dynamic Choices via Pessimism. *CoRR abs/2305.18438* (2023). <https://doi.org/10.48550/ARXIV.2305.18438> arXiv:2305.18438
- [168] Zhaodonghui Li, Haitao Yuan, Huiming Wang, Gao Cong, and Lidong Bing. 2024. LLM-R2: A Large Language Model Enhanced Rule-based Rewrite System for Boosting Query Efficiency. *CoRR abs/2404.12872* (2024). <https://doi.org/10.48550/ARXIV.2404.12872> arXiv:2404.12872
- [169] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, Fuchun Sun, and Kunlun He. 2024. A Survey of Knowledge Graph Reasoning on Graph Types: Static, Dynamic, and Multi-Modal. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 12 (2024), 9456–9478. <https://doi.org/10.1109/TPAMI.2024.3417451>
- [170] Sean Lie. 2023. Cerebras Architecture Deep Dive: First Look Inside the Hardware/Software Co-Design for Deep Learning. *IEEE Micro* 43, 3 (2023), 18–30. <https://doi.org/10.1109/MM.2023.3256384>
- [171] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedico, Erez Safahi, Shaked Meiron, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zuzman, and Yoav Shoham. 2024. Jamba: A Hybrid Transformer-Mamba Language Model. *CoRR abs/2403.19887* (2024). <https://doi.org/10.48550/ARXIV.2403.19887> arXiv:2403.19887
- [172] Jonathan Light, Min Cai, Weiqin Chen, Guanzhi Wang, Xiusi Chen, Wei Cheng, Yisong Yue, and Ziniu Hu. 2024. Strategist: Learning Strategic Skills by LLMs via Bi-Level Tree Search. *CoRR abs/2408.10635* (2024). <https://doi.org/10.48550/ARXIV.2408.10635> arXiv:2408.10635
- [173] Chaofan Lin, Zhenhua Han, Chengruidong Zhang, Yuqing Yang, Fan Yang, Chen Chen, and Lili Qiu. 2024. Parrot: Efficient Serving of LLM-based Applications with Semantic Variable. In *18th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2024, Santa Clara, CA, USA, July 10-12, 2024*, Ada Gavrilovska and Douglas B. Terry (Eds.). USENIX Association, 929–945. <https://www.usenix.org/conference/osdi24/presentation/lin-chaofan>
- [174] Yiming Lin and Sharad Mehrotra. 2024. PLAQUE: Automated Predicate Learning at Query Time. *Proc. ACM Manag. Data* 2, 1 (2024), 46:1–46:25. <https://doi.org/10.1145/3639301>
- [175] Zicheng Lin, Tian Liang, Jiahao Xu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. 2024. Critical Tokens Matter: Token-Level Contrastive Estimation Enhance LLM’s Reasoning Capability. *arXiv preprint arXiv:2411.19943* (2024).
- [176] Akide Liu, Jing Liu, Zizheng Pan, Yefei He, Gholamreza Haffari, and Bohan Zhuang. 2024. MiniCache: KV Cache Compression in Depth Dimension for Large Language Models. *CoRR abs/2405.14366* (2024). <https://doi.org/10.48550/ARXIV.2405.14366> arXiv:2405.14366
- [177] Chunwei Liu, Matthew Russo, Michael J. Cafarella, Lei Cao, Peter Baile Chen, Zui Chen, Michael J. Franklin, Tim Kraska, Samuel Madden, and Gerardo Vitagliano. 2024. A Declarative System for Optimizing AI Workloads. *CoRR abs/2405.14696* (2024). <https://doi.org/10.48550/ARXIV.2405.14696> arXiv:2405.14696
- [178] Di Liu, Meng Chen, Baotong Lu, Huiqiang Jiang, Zhenhua Han, Qianxi Zhang, Qi Chen, Chengruidong Zhang, Bailu Ding, Kai Zhang, Chen Chen, Fan Yang, Yuqing Yang, and Lili Qiu. 2024. RetrievalAttention: Accelerating Long-Context LLM Inference via Vector Retrieval. *CoRR abs/2409.10516* (2024). <https://doi.org/10.48550/ARXIV.2409.10516> arXiv:2409.10516
- [179] Hao Liu and Pieter Abbeel. 2023. Emergent Agentic Transformer from Chain of Hindsight Experience. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 21362–21374. <https://proceedings.mlr.press/v202/liu23a.html>
- [180] Jinsu Liu, Hamid Hadian, Hanchen Xu, Daniel S. Berger, and Huaicheng Li. 2024. Dissecting CXL Memory Performance at Scale: Analysis, Modeling, and Optimization. *CoRR abs/2409.14317* (2024). <https://doi.org/10.48550/ARXIV.2409.14317> arXiv:2409.14317
- [181] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguistics* 12 (2024), 157–173. https://doi.org/10.1162/TACL_A_00638
- [182] Shu Liu, Asim Biswal, Audrey Cheng, Xiangxi Mo, Shiyi Cao, Joseph E. Gonzalez, Ion Stoica, and Matei Zaharia. 2024. Optimizing LLM Queries in Relational Workloads. *CoRR abs/2403.05821* (2024). <https://doi.org/10.48550/ARXIV.2403.05821> arXiv:2403.05821
- [183] Weiwei Liu, Zecheng Tang, Juntao Li, Kehai Chen, and Min Zhang. 2024. MemLong: Memory-Augmented Retrieval for Long Text Modeling. *CoRR abs/2408.16967* (2024). <https://doi.org/10.48550/ARXIV.2408.16967> arXiv:2408.16967
- [184] Yuhang Liu, Hanchen Li, Yihua Cheng, Siddhant Ray, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, Michael Maire, Henry Hoffmann, Ari Holtzman, and Junchen Jiang. 2024. CacheGen: KV Cache Compression and Streaming for Fast Large Language Model Serving. In *Proceedings of the ACM SIGCOMM 2024 Conference, ACM SIGCOMM 2024, Sydney, NSW, Australia, August 4-8, 2024*, ACM, 38–56. <https://doi.org/10.1145/3651890.3672274>
- [185] Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024. RA-ISF: Learning to Answer and Understand from Retrieval Augmentation via Iterative Self-Feedback. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 4730–4749. <https://doi.org/10.18653/V1/2024.FINDINGS-ACL.281>
- [186] Chao Lou, Zixia Jia, Zilong Zheng, and Kewei Tu. 2024. Sparser is Faster and Less is More: Efficient Sparse Attention for Long-Range Transformers. *CoRR abs/2406.16747* (2024). <https://doi.org/10.48550/ARXIV.2406.16747> arXiv:2406.16747
- [187] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *CoRR abs/2408.06292* (2024). <https://doi.org/10.48550/ARXIV.2408.06292> arXiv:2408.06292
- [188] Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. 2024. ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities. *CoRR abs/2408.04682* (2024). <https://doi.org/10.48550/ARXIV.2408.04682> arXiv:2408.04682
- [189] Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D. Lane, and Mengwei Xu. 2024. Small Language Models: Survey, Measurements, and Insights. *CoRR abs/2409.15790* (2024). <https://doi.org/10.48550/ARXIV.2409.15790> arXiv:2409.15790
- [190] Kaixin Ma, Hao Cheng, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. 2022. Open-domain Question Answering via Chain of Reasoning over Heterogeneous Knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 5360–5374. <https://doi.org/10.18653/V1/2022.FINDINGS-EMNLP.392>
- [191] Ziming Mao, Tian Xia, Zhanghao Wu, Wei-Lin Chiang, Tyler Griggs, Romil Bhardwaj, Zongheng Yang, Scott Shenker, and Ion Stoica. 2024. SkyServe: Serving AI Models across Regions and Clouds with Spot Instances. *CoRR abs/2411.01438* (2024). <https://doi.org/10.48550/ARXIV.2411.01438> arXiv:2411.01438
- [192] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman (Eds.). 56–61. <https://doi.org/10.25080/Majora-92bf1922-00a>

- [193] Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-Editing Memory in a Transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/forum?id=MkbcAHlYgyS>
- [194] Xupeng Miao, Zhihao Jia, and Bin Cui. 2024. Demystifying Data Management for Large Language Models. In *Companion of the 2024 International Conference on Management of Data, SIGMOD/PODS 2024, Santiago AA, Chile, June 9-15, 2024*, Pablo Barceló, Nayat Sánchez-Pi, Alexandra Meliou, and S. Sudarshan (Eds.). ACM, 547–555. <https://doi.org/10.1145/3626246.3654683>
- [195] Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained Hallucination Detection and Editing for Language Models. *CoRR* abs/2401.06855 (2024). <https://doi.org/10.48550/ARXIV.2401.06855> arXiv:2401.06855
- [196] M. Mehdi Mojarradi, Lingyi Yang, Robert McCraith, and Adam Mahdi. 2024. Improving In-Context Learning with Small Language Model Ensembles. *CoRR* abs/2410.21868 (2024). <https://doi.org/10.48550/ARXIV.2410.21868> arXiv:2410.21868
- [197] Laurent Mombaerts, Terry Ding, Adi Banerjee, Florian Felice, Jonathan Taws, and Tarik Borogovac. 2024. Meta Knowledge for Retrieval Augmented Large Language Models. *CoRR* abs/2408.09017 (2024). <https://doi.org/10.48550/ARXIV.2408.09017> arXiv:2408.09017
- [198] Niklas Muennighoff, Qian Liu, Armel Randy Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and Shayne Longpre. 2024. OctoPack: Instruction Tuning Code Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=mw1PWNSWZP>
- [199] Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative Representational Instruction Tuning. *CoRR* abs/2402.09906 (2024). <https://doi.org/10.48550/ARXIV.2402.09906> arXiv:2402.09906
- [200] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. 2021. A White Paper on Neural Network Quantization. *CoRR* abs/2106.08295 (2021). arXiv:2106.08295 <https://arxiv.org/abs/2106.08295>
- [201] Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish. 2022. Responsible Data Integration: Next-generation Challenges. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary G. Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 2458–2464. <https://doi.org/10.1145/3514221.3522567>
- [202] Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, Junda Wu, Ashish Singh, Yu Wang, Jiuxiang Gu, Franck Dernoncourt, Nesreen K. Ahmed, Nedim Lipka, Ruiyi Zhang, Xiang Chen, Tong Yu, Sungchul Kim, Hanieh Deilamsalehy, Namyong Park, Mike Rimer, Zhehao Zhang, Huanrui Yang, Ryan A. Rossi, and Thien Huu Nguyen. 2024. A Survey of Small Language Models. *CoRR* abs/2410.20011 (2024). <https://doi.org/10.48550/ARXIV.2410.20011> arXiv:2410.20011
- [203] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large Dual Encoders Are Generalizable Retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 9844–9855. <https://doi.org/10.18653/V1/2022.EMNLP-MAIN.669>
- [204] Noa Nonkes, Sergei Agaronian, Evangelos Kanoulas, and Roxana Petcu. 2024. Leveraging Graph Structures to Detect Hallucinations in Large Language Models. *CoRR* abs/2407.04485 (2024). <https://doi.org/10.48550/ARXIV.2407.04485> arXiv:2407.04485
- [205] OpenAI. 2024. Introducing OpenAI o1. <https://openai.com/o1>. Accessed: 2024-12-15.
- [206] Laurel J. Orr, Srikanth Kandula, and Surajit Chaudhuri. 2019. Pushing Data-Induced Predicates Through Joins in Big-Data Clusters. *Proc. VLDB Endow.* 13, 3 (2019), 252–265. <https://doi.org/10.14778/3368289.3368292>
- [207] Laurel J. Orr, Atindriyo Sanyal, Xiao Ling, Karan Goel, and Megan Leszczynski. 2021. Managing ML Pipelines: Feature Stores and the Coming Wave of Embedding Ecosystems. *Proc. VLDB Endow.* 14, 12 (2021), 3178–3181. <https://doi.org/10.14778/3476311.3476402>
- [208] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html
- [209] Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srini Iyer. 2024. Byte Latent Transformer: Patches Scale Better Than Tokens. (2024). <https://ai.meta.com/research/publications/byte-latent-transformer-patches-scale-better-than-tokens/>
- [210] James Jie Pan, Jianguo Wang, and Guoliang Li. 2024. Survey of vector database management systems. *VLDB J.* 33, 5 (2024), 1591–1615. <https://doi.org/10.1007/S00778-024-00864-X>
- [211] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering* (2024).
- [212] Xuchen Pan, Dawei Gao, Yuexiang Xie, Zhewei Wei, Yaliang Li, Bolin Ding, Ji-Rong Wen, and Jingren Zhou. 2024. Very Large-Scale Multi-Agent Simulation in AgentScope. *CoRR* abs/2407.17789 (2024). <https://doi.org/10.48550/ARXIV.2407.17789> arXiv:2407.17789
- [213] Xiurui Pan, Endian Li, Qiao Li, Shengwen Liang, Yizhou Shan, Ke Zhou, Yingwei Luo, Xiaolin Wang, and Jie Zhang. 2024. InstInfer: In-Storage Attention Offloading for Cost-Effective Long-Context LLM Inference. *CoRR* abs/2409.04992 (2024). <https://doi.org/10.48550/ARXIV.2409.04992> arXiv:2409.04992
- [214] Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (Eds.). ACM, 2:1–2:22. <https://doi.org/10.1145/3586183.3606763>
- [215] Dylan Patel and Afzal Ahmad. 2023. The Inference Cost of Search Disruption – Large Language Model Cost Analysis. <https://www.semianalysis.com/p/the-inference-cost-of-search-disruption>. Accessed: 2024-12-15.
- [216] Liana Patel, Siddharth Jha, Carlos Guestrin, and Matei Zaharia. 2024. LOTUS: Enabling Semantic Queries with LLMs Over Tables of Unstructured and Structured Data. *CoRR* abs/2407.11418 (2024). <https://doi.org/10.48550/ARXIV.2407.11418> arXiv:2407.11418
- [217] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Riccardo Bianchini. 2024. Splitwise: Efficient Generative LLM Inference Using Phase Splitting. In *51st ACM/IEEE Annual International Symposium on Computer Architecture, ISCA 2024, Buenos Aires, Argentina, June 29 - July 3, 2024*. IEEE, 118–132. <https://doi.org/10.1109/ISCA59077.2024.00019>
- [218] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large Language Model Connected with Massive APIs. *CoRR* abs/2305.15334 (2023). <https://doi.org/10.48550/ARXIV.2305.15334> arXiv:2305.15334
- [219] Johns Paul, Bingsheng He, Shengliang Lu, and Chiew Tong Lau. 2020. Improving execution efficiency of just-in-time compilation based query processing on GPUs. *Proceedings of the VLDB Endowment* 14, 2 (2020), 202–214.
- [220] Johns Paul, Shengliang Lu, and Bingsheng He. 2021. Database Systems on GPUs. *Found. Trends Databases* 11, 1 (2021), 1–108. <https://doi.org/10.1561/19000000076>
- [221] Tim Pearce and Jinyeop Song. 2024. Reconciling Kaplan and Chinchilla Scaling Laws. *CoRR* abs/2406.12907 (2024). <https://doi.org/10.48550/ARXIV.2406.12907> arXiv:2406.12907
- [222] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, Paola Merlo, Jörg Tiedemann, and Reut Tsarfay (Eds.). Association for Computational Linguistics, 487–503. <https://doi.org/10.18653/V1/2021.EACL-MAIN.39>
- [223] Jonathan Pilault, Mahan Fathi, Orhan Firat, Chris Pal, Pierre-Luc Bacon, and Ross Goroshin. 2023. Block-State Transformers. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/16ccd203e9e3696a7ab0dcf568316379-Abstract-Conference.html
- [224] Mohammadreza Pourreza, Hailong Li, Ruoxi Sun, Yeounoh Chung, Shayan Talaei, Gaurav Tarlok Kakkar, Yu Gan, Amin Saberi, Fatma Ozcan, and Sercan Ö. Arik. 2024. CHASE-SQL: Multi-Path Reasoning and Preference Optimized Candidate Selection in Text-to-SQL. *CoRR* abs/2410.01943 (2024). <https://doi.org/10.48550/ARXIV.2410.01943> arXiv:2410.01943
- [225] Tavva Prudhivithi, Chakrabarty Swattik, and Selvakumar Prakash. 2024. Enhancing Retrieval Augmented Generation Systems with Knowledge Graphs. In *2024 International Conference on Electrical, Computer and Energy Technologies (ICEET)*. IEEE, 1–8.
- [226] Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solvers. *CoRR* abs/2408.06195 (2024). <https://doi.org/10.48550/ARXIV.2408.06195> arXiv:2408.06195

- [227] Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024. Scaling Large-Language-Model-based Multi-Agent Collaboration. *CoRR* abs/2406.07155 (2024). <https://doi.org/10.48550/ARXIV.2406.07155> arXiv:2406.07155
- [228] Yulei Qian, Fengcun Li, Xiangyang Ji, Xiaoyu Zhao, Jianchao Tan, Kefeng Zhang, and Xunliang Cai. 2024. EPS-MoE: Expert Pipeline Scheduler for Cost-Efficient MoE Inference. *CoRR* abs/2410.12247 (2024). <https://doi.org/10.48550/ARXIV.2410.12247>
- [229] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=dHng2O0Jjr>
- [230] Han Qiu, Jiaxing Huang, Peng Gao, Qin Qi, Xiaoqin Zhang, Ling Shao, and Shijian Lu. 2024. LongHalQA: Long-Context Hallucination Evaluation for MultiModal Large Language Models. *CoRR* abs/2404.09962 (2024). <https://doi.org/10.48550/ARXIV.2410.09962> arXiv:2410.09962
- [231] Haoran Qiu, Weichao Mao, Archit Patke, Shengkun Cui, Saurabh Jha, Chen Wang, Hubertus Franke, Zbigniew T. Kalbarczyk, Tamer Basar, and Ravishankar K. Iyer. 2024. Efficient Interactive LLM Serving with Proxy Model-based Sequence Length Prediction. *CoRR* abs/2404.08509 (2024). <https://doi.org/10.48550/ARXIV.2404.08509> arXiv:2404.08509
- [232] Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2024. Towards Completeness-Oriented Tool Retrieval for Large Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, Edorado Serra and Francesca Spezzano (Eds.). ACM, 1930–1940. <https://doi.org/10.1145/3627673.3679847>
- [233] Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomás Cerný. 2024. Detecting Hallucinations in Large Language Model Generation: A Token Probability Approach. *CoRR* abs/2405.19648 (2024). <https://doi.org/10.48550/ARXIV.2405.19648> arXiv:2405.19648
- [234] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html
- [235] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. <https://jmlr.org/papers/v21/20-074.html>
- [236] Isaac Rehg. 2024. KV-Compress: Paged KV-Cache Compression with Variable Compression Rates per Attention Head. *CoRR* abs/2410.00161 (2024). <https://doi.org/10.48550/ARXIV.2410.00161> arXiv:2410.00161
- [237] Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. 2024. Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling. *arXiv preprint arXiv:2406.07522* (2024).
- [238] Minsoo Rhu, Natalia Gimelshein, Jason Clemons, Arslan Zulfiqar, and Stephen W. Keckler. 2016. vDNN: Virtualized deep neural networks for scalable, memory-efficient neural network design. In *49th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 2016, Taipei, Taiwan, October 15-19, 2016*. IEEE Computer Society, 18:1–18:13. <https://doi.org/10.1109/MICRO.2016.7783721>
- [239] Guilherme Rosa, Luiz Bonifacio, Vitor Jeronimo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. In defense of cross-encoders for zero-shot retrieval. *arXiv preprint arXiv:2212.06121* (2022).
- [240] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *CoRR* abs/2402.07927 (2024). <https://doi.org/10.48550/ARXIV.2402.07927> arXiv:2402.07927
- [241] Gaurav Sahu, Pau Rodriguez, Issam H. Laradji, Parmida Atighehchian, David Vázquez, and Dzmitry Bahdanau. 2022. Data Augmentation for Intent Classification with Off-the-shelf Large Language Models. In *Proceedings of the 4th Workshop on NLP for Conversational AI, ConvAI@ACL 2022, Dublin, Ireland, May 27, 2022*, Bing Liu, Alexandros Papangelis, Stefan Ultes, Abhinav Rastogi, Yun-Nung Chen, Georgios Spithourakis, Elnaz Nouri, and Weiyan Shi (Eds.). Association for Computational Linguistics, 47–57. <https://doi.org/10.18653/V1/2022.NLP4CONVAI-1.5>
- [242] Viktor Sanca and Anastasia Ailamaki. 2024. Efficient Data Access Paths for Mixed Vector-Relational Search. In *Proceedings of the 20th International Workshop on Data Management on New Hardware, DaMoN 2024, Santiago, Chile, 10 June 2024*, Carsten Binnig and Nesime Tatbul (Eds.). ACM, 6:1–6:9. <https://doi.org/10.1145/3662010.3663448>
- [243] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesh Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=9Vrb9D0Wt4>
- [244] Keshav Santhanam, Deepti Raghavan, Muhammad Shahir Rahman, Thejas Venkatesh, Neha Kunjal, Pratiksha Thaker, Philip Alexander Levis, and Matei Zaharia. 2024. ALTO: An Efficient Network Orchestrator for Compound AI Systems. In *Proceedings of the 4th Workshop on Machine Learning and Systems, EuroMLSys 2024, Athens, Greece, 22 April 2024*. ACM, 117–125. <https://doi.org/10.1145/3642970.3655844>
- [245] Bhaskarjit Sarmah, Dhagash Mehta, Benika Hall, Rohan Rao, Sunil Patel, and Stefano Pasquali. 2024. HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 608–616.
- [246] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html
- [247] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoungJun Han, Seviun Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson C. Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Miserlis Hoyle, and Philip Resnik. 2024. The Prompt Report: A Systematic Survey of Prompting Techniques. *CoRR* abs/2406.06608 (2024). <https://doi.org/10.48550/ARXIV.2406.06608> arXiv:2406.06608
- [248] Robert Schulze, Tom Schreiber, Ilya Yatsishin, Ryadh Dahimene, and Alexey Mirodov. 2024. ClickHouse-Lightning Fast Analytics for Everyone. *Proceedings of the VLDB Endowment* 17, 12 (2024), 3731–3744.
- [249] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision. *CoRR* abs/2407.08608 (2024). <https://doi.org/10.48550/ARXIV.2407.08608> arXiv:2407.08608
- [250] Lior Shani, Aviv Rosenberg, Asaf B. Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szepes, Avinatan Hassidim, Yossi Matias, and Rémi Munos. 2024. Multi-turn Reinforcement Learning from Preference Human Feedback. *CoRR* abs/2405.14655 (2024). <https://doi.org/10.48550/ARXIV.2405.14655> arXiv:2405.14655
- [251] Wenqi Shao, Mengzhao Chen, Zhaoqiang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. 2024. OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=8Wuvhh0LYW>
- [252] Ying Sheng, Shiyi Cao, Dacheng Li, Banghua Zhu, Zhuohan Li, Danyang Zhuo, Joseph E. Gonzalez, and Ion Stoica. 2024. Fairness in Serving Large Language Models. In *18th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2024, Santa Clara, CA, USA, July 10-12, 2024*, Ada Gavrilovska and Douglas B. Terry (Eds.). USENIX Association, 965–988. <https://www.usenix.org/conference/osdi24/presentation/sheng>
- [253] Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. 2023. FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 31094–31116. <https://proceedings.mlr.press/v202/sheng23a.html>
- [254] Nikhil Sheoran, Supawit Chockchowwat, Arav Chheda, Suwen Wang, Riya Verma, and Yongjoo Park. 2023. A Step Toward Deep Online Aggregation. *Proc. ACM Manag. Data* 1, 2 (2023), 124:1–124:28. <https://doi.org/10.1145/3589269>
- [255] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 4222–4235. <https://doi.org/10.18653/>

- [256] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/1b44b87bb782e6954cd888628510e90-Abstract-Conference.html
- [257] Mohammad Shoeibi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *CoRR* abs/1909.08053 (2019). [arXiv:1909.08053](https://arxiv.org/abs/1909.08053) <http://arxiv.org/abs/1909.08053>
- [258] Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 To Be Reliable. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/forum?id=98p5x5L5af>
- [259] Tomer Simon. 2024. The scientist of the scientist. *AI Soc.* 39, 2 (2024), 803–804. <https://doi.org/10.1007/S00146-022-01544-6>
- [260] Panagiotis Sioulas, Viktor Sanca, Ioannis Mytilinis, and Anastasia Ailamaki. 2021. Accelerating Complex Analytics using Speculation. In *11th Conference on Innovative Data Systems Research, CIDR 2021, Virtual Event, January 11-15, 2021, Online Proceedings*. www.cidrdb.org. http://cidrdb.org/cidr2021/papers/cidr2021_paper03.pdf
- [261] Yixin Song, Zeyu Mi, Haotong Xie, and Haibo Chen. 2024. PowerInfer: Fast Large Language Model Serving with a Consumer-grade GPU. In *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles, SOSP 2024, Austin, TX, USA, November 4-6, 2024*, Emmett Witchel, Christopher J. Rossbach, Andrea C. Arpaci-Dusseau, and Kimberly Keeton (Eds.). ACM, 590–606. <https://doi.org/10.1145/3694715.3695964>
- [262] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). <https://proceedings.neurips.cc/paper/2020/hash/1f8985d556929e98d3ef9b86448f951-Abstract.html>
- [263] Foteini Strati, Sara McAllister, Amar Phanishayee, Jakub Tarnawski, and Ana Klimovic. 2024. DéjàVu: KV-cache Streaming for Fast, Fault-tolerant Generative LLM Serving. In *Forty-first International Conference on Machine Learning, ICLR 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. <https://openreview.net/forum?id=AbGbGZFYOd>
- [264] Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 14379–14391. <https://doi.org/10.18653/V1/2024.FINDINGS-ACL.854>
- [265] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 645–654.
- [266] Jiashuo Sun, Chengjin Xu, Luminyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model with Knowledge Graph. *CoRR* abs/2307.07697 (2023). <https://doi.org/10.48550/ARXIV.2307.07697>
- [267] Xin Tan, Yimin Jiang, Yitao Yang, and Hong Xu. 2024. Teola: Towards End-to-End Optimization of LLM-based Applications. *CoRR* abs/2407.00326 (2024). <https://doi.org/10.48550/ARXIV.2407.00326>
- [268] James Thorne, Majid Yazdani, Marzieh Saeidi, Fabrizio Silvestri, Sebastian Riedel, and Alon Y. Levy. 2021. From Natural Language Processing to Neural Databases. *Proc. VLDB Endow.* 14, 6 (2021), 1033–1039. <https://doi.org/10.14778/3447689.3447706>
- [269] Bing Tian, Haikun Liu, Yuhang Tang, Shihai Xiao, Zhuohui Duan, Xiaofei Liao, Xuechang Zhang, Junhua Zhu, and Yu Zhang. 2024. FusionANNS: An Efficient CPU/GPU Cooperative Processing Architecture for Billion-scale Approximate Nearest Neighbor Search. *CoRR* abs/2409.16576 (2024). <https://doi.org/10.48550/ARXIV.2409.16576>
- [270] S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *CoRR* abs/2401.01313 (2024). <https://doi.org/10.48550/ARXIV.2401.01313>
- [271] Immanuel Trummer. 2022. DB-BERT: A Database Tuning Tool that "Reads the Manual". In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary G. Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 190–203. <https://doi.org/10.1145/3514221.3517843>
- [272] Immanuel Trummer. 2023. From BERT to GPT-3 Codex: Harnessing the Potential of Very Large Language Models for Data Management. *CoRR* abs/2306.09339 (2023). <https://doi.org/10.48550/ARXIV.2306.09339>
- [273] Matthias Urban and Carsten Binnig. 2024. CAESURA: Language Models as Multi-Modal Query Planners. In *14th Conference on Innovative Data Systems Research, CIDR 2024, Chaminade, HI, USA, January 14-17, 2024*. www.cidrdb.org. <https://www.cidrdb.org/cidr2024/papers/p14-urban.pdf>
- [274] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [275] Shubham Vatsal and Harsh Dubey. 2024. A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks. *CoRR* abs/2407.12994 (2024). <https://doi.org/10.48550/ARXIV.2407.12994>
- [276] Jonas Waldendorf, Barry Haddow, and Alexandra Birch. 2024. Contrastive Decoding Reduces Hallucinations in Large Multilingual Machine Translation Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, 2526–2539. <https://aclanthology.org/2024.eacl-long.155>
- [277] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arik. 2024. Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models. *CoRR* abs/2410.07176 (2024). <https://doi.org/10.48550/ARXIV.2410.07176>
- [278] Jinyuan Wang, Junlong Li, and Hai Zhao. 2023. Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning. *arXiv preprint arXiv:2310.13552* (2023).
- [279] Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and Yunhong Wang. 2024. A Survey on Data Synthesis and Augmentation for Large Language Models. *CoRR* abs/2410.12896 (2024). <https://doi.org/10.48550/ARXIV.2410.12896>
- [280] Wenyi Wang, Hisham A Alyahya, Dylan R Ashley, Oleg Serikov, Dmitrii Khizbullin, Francesco Faccio, and Jürgen Schmidhuber. 2024. How to Correctly do Semantic Backpropagation on Language-based Agentic Systems. *arXiv preprint arXiv:2412.03624* (2024).
- [281] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting Language Models with Long-Term Memory. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/ebd82705f44793b6f9ade5a669d0f0bf-Abstract-Conference.html
- [282] Xiaochen Wang, Junqing He, Liang Chen, Reza Haf Zhe Yang, Yiru Wang, Xiangdi Meng, Kunhao Pan, and Zhifang Sui. 2024. SG-FSM: A Self-Guiding Zero-Shot Prompting Paradigm for Multi-Hop Question Answering Based on Finite State Machine. *arXiv preprint arXiv:2410.17021* (2024).
- [283] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/forum?id=1PL1NIMMrw>
- [284] Xuezhi Wang and Denny Zhou. 2024. Chain-of-Thought Reasoning Without Prompting. *CoRR* abs/2402.10200 (2024). <https://doi.org/10.48550/ARXIV.2402.10200>
- [285] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoki Okazaki (Eds.). Association for Computational Linguistics, 13484–13508. <https://doi.org/10.18653/V1/2023.ACL-LONG.754>
- [286] Zhiruo Wang, Zhoujun Cheng, Hao Zhu, Daniel Fried, and Graham Neubig. 2024. What Are Tools Anyway? A Survey from the Language Model Perspective. *CoRR* abs/2403.15452 (2024). <https://doi.org/10.48550/ARXIV.2403.15452>
- [287] Zheng Wang, Shu Xian Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024. M-RAG: Reinforcing Large Language Model Performance through Retrieval-Augmented Generation with Multiple Partitions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

- Papers), *ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 1966–1978. <https://doi.org/10.18653/V1/2024.ACL-LONG.108>
- [288] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=gEZrGCozdqR>
- [289] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
- [290] Steven Whang and Jae-Gil Lee. 2020. Data Collection and Quality Challenges for Deep Learning. *Proc. VLDB Endow.* 13, 12 (2020), 3429–3432. <https://doi.org/10.14778/3415478.3415562>
- [291] Bingyang Wu, Yinmin Zhong, Zili Zhang, Gang Huang, Xuanzhe Liu, and Xin Jin. 2023. Fast Distributed Inference Serving for Large Language Models. *CoRR abs/2305.05920* (2023). <https://doi.org/10.48550/ARXIV.2305.05920>
- [292] Xun Wu, Shaohan Huang, and Furu Wei. 2024. Mixture of LoRA Experts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=uWvKBCYh4S>
- [293] Yingjun Wu, Chee Yong Chan, and Kian-Lee Tan. 2016. Transaction Healing: Scaling Optimistic Concurrency Control on Multicores. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, Fatma Özcan, Georgia Koutrika, and Sam Madden (Eds.). ACM, 1689–1704. <https://doi.org/10.1145/2882903.2915202>
- [294] Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing Transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=TrjbxxzRcnf->
- [295] Guangxuan Xiao, Ji Lin, Micaël Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. In *International Conference on Machine Learning, ICLR 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 38087–38099. <https://proceedings.mlr.press/v202/xiao23c.html>
- [296] Junjie Xing, Yeye He, Mengyu Zhou, Haoyu Dong, Shi Han, Dongmei Zhang, and Surajit Chaudhuri. 2024. Table-LLM-Specialist: Language Model Specialists for Tables using Iterative Generator-Validator Fine-tuning. *CoRR abs/2410.12164* (2024). <https://doi.org/10.48550/ARXIV.2410.12164>
- [297] Yizhe Xiong, Xiansheng Chen, Xin Ye, Hui Chen, Zijia Lin, Haoran Lian, Jianwei Niu, and Guiguang Ding. 2024. Temporal Scaling Law for Large Language Models. *CoRR abs/2404.17785* (2024). <https://doi.org/10.48550/ARXIV.2404.17785>
- [298] Jiale Xu, Rui Zhang, Cong Guo, Weiming Hu, Zihan Liu, Feiyang Wu, Yu Feng, Shixuan Sun, Changxu Shao, Yuhong Guo, Junpeng Zhao, Ke Zhang, Minyi Guo, and Jingwen Leng. 2024. vTensor: Flexible Virtual Tensor Management for Efficient LLM Serving. *CoRR abs/2407.15309* (2024). <https://doi.org/10.48550/ARXIV.2407.15309>
- [299] Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See-Kiong Ng, and Jiashi Feng. 2024. MAGIC: Investigation of Large Language Model Powered Multi-Agent in Cognition, Adaptability, Rationality and Collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 7315–7332. <https://aclanthology.org/2024.emnlp-main.416>
- [300] Qiancheng Xu, Yongqi Li, Heming Xia, and Wenjie Li. 2024. Enhancing Tool Retrieval with Iterative Feedback from Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 9609–9619. <https://aclanthology.org/2024.findings-emnlp.561>
- [301] Sheng Xu, Mike Chen, and Shuwen Chen. 2024. Enhancing Retrieval-Augmented Generation Models with Knowledge Graphs: Innovative Practices Through a Dual-Pathway Approach. In *International Conference on Intelligent Computing*. Springer, 398–409.
- [302] Weijia Xu, Andrzej Banburski, and Nebojsa Jojic. 2024. Reprompting: Automated Chain-of-Thought Prompt Inference Through Gibbs Sampling. In *Forty-first International Conference on Machine Learning, ICLR 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. <https://openreview.net/forum?id=D8zn1DnTuj>
- [303] Yuming Xu, Hengyu Liang, Jin Li, Shuotao Xu, Qi Chen, Qianxi Zhang, Cheng Li, Ziyue Yang, Fan Yang, Yuqing Yang, Peng Cheng, and Mao Yang. 2023. SPFresh: Incremental In-Place Update for Billion-Scale Vector Search. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (Eds.). ACM, 545–561. <https://doi.org/10.1145/3600006.3613166>
- [304] Yi Xu, Ziming Mao, Xiangxi Mo, Shu Liu, and Ion Stoica. 2024. Pie: Pooling CPU Memory for LLM Inference. *arXiv preprint arXiv:2411.09317* (2024).
- [305] Ziwei Xu, Sanjay Jain, and Mohan S. Kankanalli. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *CoRR abs/2401.11817* (2024). <https://doi.org/10.48550/ARXIV.2401.11817>
- [306] Zifei Xu, Alexander Lan, Wanzin Yazar, Tristan Webb, Sayeh Sharify, and Xin Wang. 2024. Scaling laws for post-training quantized large language models. *CoRR abs/2410.12119* (2024). <https://doi.org/10.48550/ARXIV.2410.12119>
- [307] Maryann Xue, Yingyi Bu, Abhishek Somani, Wenchen Fan, Ziqi Liu, Steven Chen, Herman Van Hovell, Bart Samwel, Mostafa Mokhtar, Rk Korlapati, Andy Lam, Yunxiao Ma, Vuk Ercegovic, Jiexing Li, Alexander Behm, Yuanjian Li, Xiao Li, Sriram Krishnamurthy, Amit Shukla, Michalis Petropoulos, Sameer Paranjpye, Reynold Xin, and Matei Zaharia. 2024. Adaptive and Robust Query Execution for Lakehouses At Scale. *Proc. VLDB Endow.* 17, 12 (2024), 3947–3959. <https://www.vldb.org/pvldb/vol17/p3947-bu.pdf>
- [308] Scott Yak, Yihe Dong, Javier Gonzalvo, and Sercan Arik. 2023. IngesTables: Scalable and Efficient Training of LLM-Enabled Tabular Foundation Models. In *NeurIPS 2023 Second Table Representation Learning Workshop*.
- [309] Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Etting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. 2024. CRAG - Comprehensive RAG Benchmark. *CoRR abs/2406.04744* (2024). <https://doi.org/10.48550/ARXIV.2406.04744>
- [310] Yifei Yang, Xiangyao Yu, Marco Serafini, Ashraf Aboulnaga, and Michael Stonebraker. 2024. FlexpushdownDB: rethinking computation pushdown for cloud OLAP DBMSs. *VLDB J.* 33, 5 (2024), 1643–1670. <https://doi.org/10.1007/S00778-024-00867-8>
- [311] Jiayi Yao, Hanchen Li, Yuhang Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. 2024. CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion. *CoRR abs/2405.16444* (2024). <https://doi.org/10.48550/ARXIV.2405.16444>
- [312] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafra, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html
- [313] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafra, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=WE_vluYUL-X
- [314] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-Augmented Multimodal Language Modeling. In *International Conference on Machine Learning, ICLR 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 39755–39769. <https://proceedings.mlr.press/v202/yasunaga23a.html>
- [315] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. 2024. Differential transformer. *arXiv preprint arXiv:2410.05258* (2024).
- [316] Zihao Ye, Ruihang Lai, Bo-Ru Lu, Chien-Yu Lin, Size Zheng, Lequn Chen, Tianqi Chen, and Luis Ceze. 2024. Cascade inference: Memory bandwidth efficient shared prefix batch decoding.
- [317] Chengye Yu, Tianyu Wang, Zili Shao, Linjie Zhu, Xu Zhou, and Song Jiang. 2024. Twinpilots: A new computing paradigm for gpu-cpu parallel llm inference. In *Proceedings of the 17th ACM International Systems and Storage Conference*. 91–103.
- [318] Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. 2022. Orca: A Distributed Serving System for Transformer-Based Generative Models. In *16th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2022, Carlsbad, CA, USA, July 11-13, 2022*, Marcos K. Aguilera and Hakim Weatherspoon (Eds.). USENIX Association, 521–538. <https://www.usenix.org/conference/osdi22/presentation/yyu>
- [319] Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards Better Chain-of-Thought Prompting Strategies: A Survey. *CoRR abs/2310.04959*

- (2023). <https://doi.org/10.48550/ARXIV.2310.04959> arXiv:2310.04959
- [320] Ruize Yuan, Xiang Ao, Li Zeng, and Qing He. 2024. DRAMA: Dynamic Multi-Granularity Graph Estimate Retrieval over Tabular and Textual Question Answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, 5365–5375. <https://aclanthology.org/2024.lrec-main.477>
- [321] Ye Yuan, Bo Tang, Tianfei Zhou, Zhiwei Zhang, and Jianbin Qin. 2024. nsDB: Architecting the Next Generation Database by Integrating Neural and Symbolic Systems (Vision). *Proc. VLDB Endow.* 17, 11 (2024), 3283–3289. <https://www.vldb.org/pvldb/vol17/p3283-tang.pdf>
- [322] Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Zhe Zhou, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, Yan Yan, Beidi Chen, Guangyu Sun, and Kurt Keutzer. 2024. LLM Inference Unveiled: Survey and Roofline Model Insights. *CoRR abs/2402.16363* (2024). <https://doi.org/10.48550/ARXIV.2402.16363> arXiv:2402.16363
- [323] Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, and Ali Ghodsi. 2024. The Shift from Models to Compound AI Systems. <https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/>.
- [324] Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, Wanli Ouyang, and Dongzhan Zhou. 2024. LLaMA-Berry: Pairwise Optimization for O1-like Olympiad-Level Mathematical Reasoning. *CoRR abs/2410.02884* (2024). <https://doi.org/10.48550/ARXIV.2410.02884> arXiv:2410.02884
- [325] Kai Zhang, Liqian Peng, Congchao Wang, Alec Go, and Xiaozhong Liu. 2024. LLM Cascade with Multi-Objective Optimal Consideration. *CoRR abs/2410.08014* (2024). <https://doi.org/10.48550/ARXIV.2410.08014> arXiv:2410.08014
- [326] Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve Anything To Augment Large Language Models. *CoRR abs/2310.07554* (2023). <https://doi.org/10.48550/ARXIV.2310.07554> arXiv:2310.07554
- [327] Qianxi Zhang, Shuotao Xu, Qi Chen, Guoxin Sui, Jiaotong Xie, Zhizhen Cai, Yaoqi Chen, Yinxuan He, Yuqing Yang, Fan Yang, Mao Yang, and Lidong Zhou. 2023. VBASE: Unifying Online Vector Similarity Search and Relational Queries via Relaxed Monotonicity. In *17th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2023, Boston, MA, USA, July 10-12, 2023*, Roxana Geambasu and Ed Nightingale (Eds.). USENIX Association, 377–395. <https://www.usenix.org/conference/osdi23/presentation/zhang-qianxi>
- [328] Shuo Zhang, Zezhou Huang, and Eugene Wu. 2024. Data Cleaning Using Large Language Models. *arXiv preprint arXiv:2410.15547* (2024).
- [329] Xiaoming Zhang, Ming Wang, Xiaocui Yang, Daling Wang, Shi Feng, and Yifei Zhang. 2024. Hierarchical Retrieval-Augmented Generation Model with Rethink for Multi-hop Question Answering. *arXiv preprint arXiv:2408.11875* (2024).
- [330] Yue Zhang, Hongliang Fei, Dingcheng Li, and Ping Li. 2022. PromptGen: Automatically Generate Prompts using Generative Models. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, 30–37. <https://doi.org/10.18653/v1/2022.FINDINGS-NAACL.3>
- [331] Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-form Open-Domain Question Answering. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 6963–6975. <https://doi.org/10.18653/v1/2024.FINDINGS-ACL.415>
- [332] Ziqi Zhang, Cunxiang Wang, Xiao Xiong, Yue Zhang, and Donglin Wang. 2024. Nash CoT: Multi-Path Inference with Preference Equilibrium. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 14572–14587. <https://aclanthology.org/2024.emnlp-main.807>
- [333] Xuanlei Zhao, Bin Jia, Haotian Zhou, Ziming Liu, Shenggan Cheng, and Yang You. 2024. HeteGen: Efficient Heterogeneous Parallel Inference for Large Language Models on Resource-Constrained Devices. In *Proceedings of the Seventh Annual Conference on Machine Learning and Systems, MLSys 2024, Santa Clara, CA, USA, May 13-16, 2024*, Phillip B. Gibbons, Gennady Pekhimenko, and Christopher De Sa (Eds.). msys.org. https://proceedings.msys.org/paper_files/paper/2024/hash/5431dca75a8d2abc1fb51e89e8324f10-Abstract-Conference.html
- [334] Yilong Zhao, Shuo Yang, Kan Zhu, Lianmin Zheng, Baris Kasikci, Yang Zhou, Jiarong Xing, and Ion Stoica. 2024. BlendServe: Optimizing Offline Inference for Auto-regressive Large Models with Resource-aware Batching. *arXiv preprint arXiv:2411.16102* (2024).
- [335] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Jeff Huang, Chuyue Sun, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark W. Barrett, and Ying Sheng. 2023. Efficiently Programming Large Language Models using SGLang. *CoRR abs/2312.07104* (2023). <https://doi.org/10.48550/ARXIV.2312.07104> arXiv:2312.07104
- [336] Wenhao Zheng, Yixiao Chen, Weitong Zhang, Souvik Kundu, Yun Li, Zhengzhong Liu, Eric P Xing, Hongyi Wang, and Huaxiu Yao. 2024. CITER: Collaborative Inference for Efficient Large Language Model Decoding with Token-Level Routing. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.
- [337] Zangwei Zheng, Xiaozhe Ren, Fuzhao Xue, Yang Luo, Xin Jiang, and Yang You. 2023. Response Length Perception and Sequence Scheduling: An LLM-Empowered LLM Inference Pipeline. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/ce7f3405c782f761fac7f849b41ae9a-Abstract-Conference.html
- [338] Han Zhong, Guhao Feng, Wei Xiong, Li Zhao, Di He, Jiang Bian, and Liwei Wang. 2024. DPO Meets PPO: Reinforced Token Optimization for RLHF. *CoRR abs/2404.18922* (2024). <https://doi.org/10.48550/ARXIV.2404.18922> arXiv:2404.18922
- [339] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. 2024. DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving. In *18th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2024, Santa Clara, CA, USA, July 10-12, 2024*, Ada Gavrilovska and Douglas B. Terry (Eds.). USENIX Association, 193–210. <https://www.usenix.org/conference/osdi24/presentation/zhong-yinmin>
- [340] Xuanhe Zhou, Guoliang Li, and Zhiyuan Liu. 2023. LLM As DBA. *CoRR abs/2308.05481* (2023). <https://doi.org/10.48550/ARXIV.2308.05481> arXiv:2308.05481
- [341] Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. 2024. A Survey on Data Augmentation in Large Model Era. *CoRR abs/2401.15422* (2024). <https://doi.org/10.48550/ARXIV.2401.15422> arXiv:2401.15422
- [342] Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhang Dong, and Yu Wang. 2024. A Survey on Efficient Inference for Large Language Models. *CoRR abs/2404.14294* (2024). <https://doi.org/10.48550/ARXIV.2404.14294> arXiv:2404.14294
- [343] Kan Zhu, Yilong Zhao, Liangyu Zhao, Gefei Zuo, Yile Gu, Dedong Xie, Yufei Gao, Qinyu Xu, Tian Tang, Zihao Ye, et al. 2024. NanoFlow: Towards Optimal Large Language Model Serving Throughput. *arXiv preprint arXiv:2408.12757* (2024).