USING SPAN QUERIES TO OPTIMIZE FOR CACHE AND ATTENTION LOCALITY

Paul Castro [1] **Nick Mitchell** [1] **Nathan Ordonez** [1] **Thomas Parnell** [1] **Mudhakar Srivatsa** [1] **Antoni Viros i Martin** [1]

ABSTRACT

Clients are evolving beyond chat completion, and now include a variety of innovative inference-time scaling and deep reasoning techniques. At the same time, inference servers remain heavily optimized for chat completion and thus largely use linear KV cache strategies. Prior work has shown that large improvements to KV cache hit rate are possible if inference servers evolve towards these more non-linear use cases. However, they offer solutions that are also optimized for a single use case, RAG. In this paper, we demonstrate that chat, RAG, inference-time scaling, and agentic workloads can be expressed as special cases of a more general structure, the *span query*. The critical distinction that had been assumed by prior work lies in whether the order of the inputs matter — do they *commute*? In chat, they do not. In RAG, they often do. A span query is an expression tree of inference calls, linked together with commutativity constraints. We describe span query syntax and semantics. We show how they can be automatically optimized to improve KV cache locality. We show how a small change to vLLM (affecting only 492 lines) can enable high-performance execution of span queries. Using this stack, we demonstrate that span queries can achieve 10-20x reductions in TTFT for two distinct non-chat use cases. Finally, we show that span queries can also be optimized to improve *attention locality*, so as to avoid the so-called lost-in-the-middle problem. We demonstrate that an attention-optimized span query on a 2b parameter model vastly outperforms the accuracy of a stock inference server using an 8b model.

1 INTRODUCTION

Cache locality is core to the viability of transformer-based large language models (Ayoub et al., 2025) (Li et al., 2025). The Key-Value (KV) cache amortizes the quadratic complexity of self-attention (Vaswani et al., 2017), reducing "prefill" GPU load both within and across requests. In this paper we focus on improving cross-request KV cache locality for emerging agentic workloads (PwC, 2025) and for the ever-evolving and innovative suite of inference time scaling (ITS) patterns (Davis et al., 2024; Zhou et al., 2025).

Inference servers such as vLLM (Kwon et al., 2023) are optimized for *prefix-based* memory reuse. For example, in a chatbot the chat history accumulates linearly and thus, with each turn of a chat, the prefix remains the same. Model servers have grown to embody this reuse pattern. In contrast, agentic and ITS workloads combine the output of multiple separate calls, and make extensive use of a corpus of post-training knowledge — the so-called retrieval augmented generation or RAG pattern (Liang et al., 2024; Sarthi et al., 2024). Neither one these has a prefix-based pattern of reuse.

Rather, with these emerging workloads, the order in which input is reused changes from request to request. For example, with RAG the first request may retrieve fragments F_1 and F_2 and present them in that order. The second request may present F_2 and F_3, thus the order of F_2 has changed. The tricky part is that whether or not order matters, i.e. whether or not it is valid to permute the order of input, depends on the application. There is no blanket rule of thumb, as observed by (Yao et al., 2025) in their CacheBlend work. Implicit in their approach is that, since this constraint is not known, it is better to optimize for the case that order does matter. We take a complementary approach.

The central premise of the paper is this: chat, RAG, ITS, and agentic workloads are special cases of a more general structure, and this more general structure hinges on the expression of whether order matters. The algebraic way to express this is via operator *commutativity*: is A, B the same as B, A? In section 4, we introduce the *Span Query*, a declarative intermediate representation that allows one to express how model calls can be arranged into a hierarchical expression tree that is linked together with commutativity constraints. For example, a span query for RAG might express that the retrieved fragments can be sequenced in any order. A span query for a judge/generator can express that the judge can inspect the generated candidates in any order.

If the span query expresses that order matters, techniques such as CacheBlend can be used. If order does not matter,

[1]IBM Research. Correspondence to: Nick Mitchell <nickm@us.ibm.com>.

Proceedings of the 8^{th} MLSys Conference, Santa Clara, CA, USA, 2025. Copyright 2025 by the author(s).

we shall show in section 5 how to automatically optimize span queries in order to improve KV cache locality and thus reduce time-to-first-token (TTFT) and prefill GPU load by as much as 20x — far exceeding the 3–4x reported by CacheBlend. Prior work on Block Attention (Ma et al., 2025) achieves similar gains, but in a way that is hard-coded for a single use case (RAG). Even on cache miss, we shall show that commutative reuse reduces prefill load by 3x due to attention sparsity. We show how we did this with the help of only 492 lines of code changed in vLLM.

Finally, we share a surprising emergent property of span queries: we can also optimize them for *attention locality*. Smaller models struggle to attend to important information located in the middle of long inputs — termed the "lost in the middle" phenomenon by (Liu et al., 2024). In section 6, we show how span queries can be optimized to improve attention locality, by treating the expression tree like a divide-and-conquer tree — c.f. map/reduce (Dean & Ghemawat, 2008). We show that, as with KV cache locality, attention locality optimization also hinges on commutativity. We demonstrate that an optimized span query on a 2b model outperforms the accuracy of an 8b model. This was not our original design point, and so it was satisfying to find that commutativity constraints have more general applicability.

2 A Brief Tour of vLLM

This section provides an overview of the interaction of vLLM and its key-value (KV) cache. We concentrate on the chat completion API, which is shared with other inference servers. The chat completion API takes as input a sequence of typed *messages*. The most common message types are `System`, `User`, and `Assistant`; in what follows you may see these designated by their first letter (**S**, **U**, **A**). A system message instructs to the model as to how best to respond, a user message represents a question being posed to the model, and an assistant message represents the output of some prior model interaction. Models are trained on a language that has been encoded into a lexicon of *tokens*. We thus consider the input to be sequences of typed tokens.

Second, vLLM breaks token sequences into fixed (but configurable) size *blocks*.[1] All KV cache operations (insertion, lookup, eviction, etc.) are performed at the block granularity. vLLM does not cache partial blocks across requests. For example, if vLLM is configured to have a block size of 16 tokens (the default setting as of this writing) and a message tokenizes to 17 tokens, the last token might not be cached.

Third, cached output blocks (i.e. those *generated* by the chat completion API call) implicitly depend on the input that generated them. This dependence is a product of the Transformers' architecture. The hash code for a block B_i, the ith in a sequence, depends on the hash code of block B_{i-1}. Thus, when provided input blocks $[I_1, \ldots, I_n]$ and observing output blocks $[O_1, \ldots, O_m]$, the KV cache can be considered to be keyed by some hash of $[I_1, \ldots, I_n, O_1, \ldots, O_m]$. Furthermore, when performing a cache lookup for a given sequence $[B_1, \ldots, B_n]$, vLLM stops scanning for cache hits after the first miss; i.e. it only searches for a *prefix* of cached blocks $[B_1, \ldots, B_{h \leq n}]$.

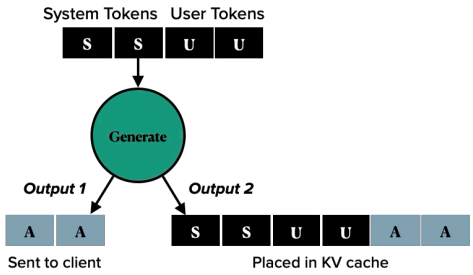


Figure 1. The "dual output paradox": the model server emits one thing to the client and something different to KV cache.

Finally, vLLM's chat completion API can be considered as having two distinct outputs. We term this the *dual output paradox*. One is sent to the client $[O_1, \ldots, O_M]$, and the other is inserted into KV cache (as illustrated in Figure 1). In some use cases (subsection 2.1), this paradox can be resolved by the client, as it can prefix each request with the history of all prior interactions. For other use cases (section 3), resolving the paradox requires changes to vLLM such as those introduced in this paper.

2.1 Use Case: Chat Completion

With that background in mind, we illustrate how well-suited vLLM's architecture is to the chat completion use case. Consider the example in Figure 2. For illustration purposes, the hypothetical block holds 2 tokens. If a user submits "Hello" (say 1 token) and the model generates "How are you?" (say 4 tokens), then vLLM will cache 2 blocks: `[Hello, how]` on the first block and `[are, you]` on the second block. The final generated token `?` is not cached for future requests as it lies on a partially filled block.

The chat session continues as the user responds "I am fine". The OpenAI chat completion API is stateless. Therefore, the chat client prepends the 5 tokens of chat history (referred to as "context" tokens) to the second request, which becomes "Hello How are you? I am fine". vLLM can avoid recomputing 4 out of 5 of the context tokens (the fifth, as per above, was not cached due to a partially filled block). In our chat completion example cache hit rate approaches 100% as the chat session progresses. This demonstrates that

[1] The vLLM literature uses "page" while the source and we use "block" to mean a fixed-size token sequence. (Ma et al., 2025) uses "block" to mean a variably-sized sequence — to us, a "span".

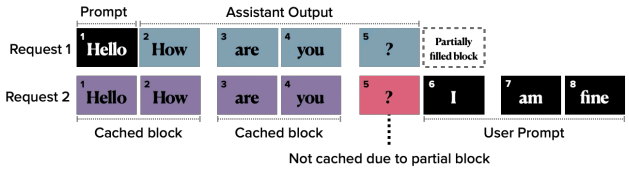


Figure 2. Chat completion use case. Each rectangle is a token, a cache block fits 2 tokens, and a token's sequence position is shown in upper left corner. 80% hit rate on Request 2 (4 of 5 input tokens are cached), which asymptotes to 100% as the chat progresses.

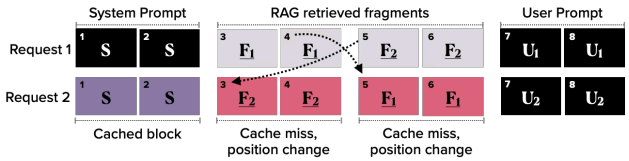


Figure 3. RAG use case. 33% hit rate on Request 2 (2 of 6 input tokens cached), which asymptotes to 0% as F_1, F_2 grow.

vLLM is well optimized for the chat completion use case.

With chat completion, blocks maintain their position within the accumulating chat context. For example, for the duration of the chat session from Figure 2 "Hello" will always appear first, "How" always second, and so on. Chat completion has an append-only pattern of updates. In contrast, while non-chat use cases also have reuse across requests, the position of reused blocks changes from request to request.

3 Non-Chat Use Cases

Next, we introduce the two non-chat use cases: retrieval-augmented generation (RAG) and "nested generation".

3.1 Use Case: Retrieval-augmented Generation

Retrieval-augmented generation (RAG) is a technique for supplementing a chat request with knowledge from an online corpus (Liang et al., 2024; Sarthi et al., 2024). Chat clients that support RAG operate in two phases: indexing and retrieval. Figure 3 illustrates a RAG use case.

First the corpus is *indexed*. A simple indexing strategy fragments the corpus into contiguous segments and then *embeds* each fragment in a vector space. In Figure 3, the corpus consists of embedded fragments $F_1, F_2, \ldots$.

Second, in response to a user message, the client embeds it in the vector space and then *retrieves* a list of nearby vectors. The closest vectors are de-embedded back to text and sent, along with the user's message, to the chat completion server. In Figure 3, the first query retrieves fragments in the order $[F_1, F_2]$ and the second query retrieves fragments in the reverse order. Even though there is reuse of the fragments across the requests, vLLM will not consider them to be cache hits because the prefix of F_2 in the second request differs from that in the first request. For this reason, as the length of the retrieved fragments grows, RAG use cases will asymptote to a 0% cache hit rate.

Recent work introduced *block attention* (Ma et al., 2025) with the goal of improving KV cache locality for RAG. We generalize that work in this paper, implementing it within the constraints of the production-grade vLLM model server, and to provide a generalized API that can be automatically optimized to support new use cases.

3.2 Use Case: Nested Generation

RAG is a proven way to improve output quality without expensive re-training (so-called test-time or inference-time scaling). A complementary inference scaling approach is to wire together model calls into workflows. This allows each model call in a flow to focus on a smaller part of the larger problem. The emerging deep research systems and agentic workflows are examples of this strategy. Whereas chat completion has an additive and append-only pattern of token accumulation, these use cases have a *nested* structure.

Definition 1 (Nested Generation) *A set of model calls G is a* nested generation *if, for any calls $g_1, g_2 \in G$, only the output of g_2 (and none of the input to g_2) is fed as input to g_1. That is, if they nest like conventional function calls.*

Consider Figure 4. This diagram illustrates a common example of nested generation: the *judge-generator* inference-time scaling technique (Davis et al., 2024; Zhou et al., 2025; Snell et al., 2025). Say we want to generate a high quality job application email. Using this technique one first generates a set of candidate emails, instructing the model (via system and user messages) what its job is, and on the specifics of the email (your name, qualifications, etc.). Next, the generated email candidates (the output of the "inner" model calls) is passed to the "outer" model call. The instructions on what constitutes a good email given to the outer generation will differ from those given to the inner generates. Thus, while there is reuse, from inner to outer model calls, there is no common prefix as we had in the case of chat completion.

Figure 5 illustrates this situation of reuse without a common prefix. In Request 1, a candidate email is generated using system prompt S_1 and instructions U_1. This produces the candidate email A_1 (A for Assistant output). In Request 2, a second email candidate is generated. Here, we get cache hits on S_1 and U_1, because there is a common prefix. The same is not true for Request 3 which acts as the judge. Even though the tokens from A_1 and A_2 reside, in some form, in KV cache, they are useless when provided as input to Request 3. The judge has the instructions $S_2 \neq S_1$ as a prefix. As a result, cache hit rate approaches 0% as the

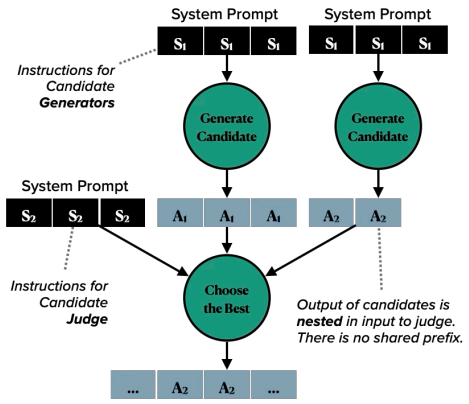


Figure 4. An example of nested generation: the *judge-generator* inference-time scaling strategy.

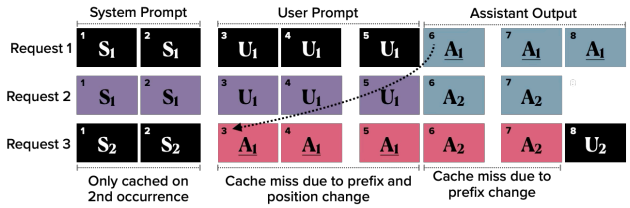


Figure 5. Nested generation use case, following on from Figure 3. 29% hit rate on Request 3 (2 of 7 input tokens cached), which asymptotes to 0% as assistant output grows.

lengths of the generated candidates A_1, A_2 grow.

4 The Span Query

We introduce a declarative intermediate representation, the *span query*. In this section, we define the syntax and semantics of span queries, provide a graphical representation that helps us to reason about queries, and show how this purely declarative form can be transformed and optimized for high-throughput execution.

4.1 Query Syntax

Our goal is to identify a declarative form that expresses our use cases as special cases of a more general structure. To that end, we start with inspiration from the observation made in Definition 1: non-chat use cases seem to have the rather conventional structure of a function call graph. We thus define the span query as a declarative *expression tree* over a set of operators. An expression tree is simply the parse tree of some client-facing library or language. We defer the topic of language design to future work.

Definition 2 (Span Query) *A* span query *is an expression tree over the operators* $\mathbf{C}, \mathbf{R}, \mathbf{F}, +, \bowtie, \mathbf{S}, \mathbf{A}, \mathbf{U}, \mathbf{G}$. *Consult Table 1 for definitions of the operators.*

Table 1. A span query is an expression tree over these operators. Desugaring of operators is discussed in subsection 5.1.

(a) Syntactic sugar

C	Chat completion
R	corpus Retrieval
F	corpus Fragment

(b) Message joins

$+$	commutative
$\bowtie$	non-commutative

(c) Core operators

S,**A**,**U**	System, User, Assistant messages
G	Generate new tokens

We show in subsection 5.1 that $\mathbf{C}$, $\mathbf{R}$, $\mathbf{F}$, and $\mathbf{G^1}$ can be desugared and expressed in terms of the core operators. Before we get there, it is helpful to walk through our use cases with a visualization. The visuals demonstrate how each use case is an instance of this general span query structure.

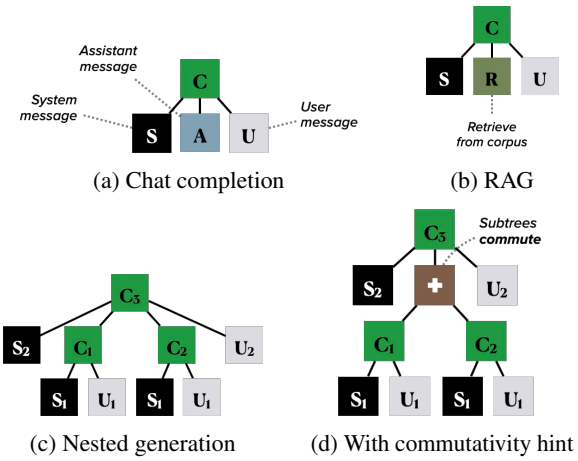
4.2 Query Visualization

We introduce a visualization of span queries. The visuals presented in this paper focus on the relations between the nodes (the edges) rather than node attributes. For example, an $\mathbf{S}$ node represents a system message, but the visualization does not show the content of this message. An $\mathbf{R}$ node represents retrieval from a corpus, but the visualization does not identify the location of the corpus. The attributes are nonetheless important, and will come into play when we serialize span queries in subsection 5.2.

Consider the visualization of a chat completion span query shown in Figure 6a. At the root of this span query is the $\mathbf{C}$ operator. This operator represents chat completion with input provided by *joining* its children together. Conventionally speaking, the input to $\mathbf{C}$ is the *prompt* to the chat completion. In this case, the $\mathbf{C}$ operator is fed three typed messages: a system message $\mathbf{S}$, the output of the prior turn of chat $\mathbf{A}$, and the user's question $\mathbf{U}$.

The chat query as shown does not explicitly govern how those three children should be joined together when presented to the model server. For chat completion it may seem obvious, i.e. that the three children should be concatenated in the order given. For chat completion, order matters. For example, one should expect lower quality output if the system message $\mathbf{S}$ were placed at the end, or if $\mathbf{U}$ came before $\mathbf{A}$. The order of the messages matters for causal attention (Vaswani et al., 2017), because later messages can attend to earlier messages but not vice versa. There is an asymmetry, and a concomitant lack of commutativity.

Figure 6b visualizes a RAG query. The structure is nearly identical to the chat completion query from Figure 6a, with $\mathbf{R}$ replacing $\mathbf{A}$. The $\mathbf{R}$ represents retrieval from a corpus. Implicit in retrieval $\mathbf{R}$ is a second ordering constraint: how



(a) Chat completion (b) RAG

(c) Nested generation (d) With commutativity hint

Figure 6. Example span queries.

should the retrieved fragments be joined together when presented as input to the **C**? It seems reasonable to infer a default join semantics from the **R** node, that order does not matter. We can automatically desugar them in a way that makes ordering constraint explicit.

Figure 6c visualizes a nested generation query. Here again, are several layers of ordering semantics that need to be teased out. For example, if the output of the inner generates can be safely permuted, then it is a fruitful exercise to reuse their KV cache entries. Some instances of nested generation may permit permuting the order of the inner generates, others may not. We need more semantic information from the user before optimization is allowed or effective.

4.3 Query Semantics

A span query is declarative, which means query execution is free of side effects. From this, it is clear cut that every sub-tree of a node can be safely dispatched in parallel. The output of every execution is another span query, thus the set of valid span queries is closed under query execution.

There is more room for interpretation regarding how a query should *join* child sub-trees in order to construct the input to the parent. This structure, of parallel fork but semantically constrained join, has strong analogies to the map-reduce paradigm (Dean & Ghemawat, 2008).

For our purposes, the semantics of joining messages depends largely on *commutativity* required by the query. Does order matter, (M_1, M_2) versus (M_2, M_1) for that given pair of messages? The answer will vary by use case, hence a span query must allow that choice to be expressed.

Definition 3 *Span Query Join Operators* A span query can express a message ordering constraint with two operators. $\bowtie$ indicates a non-commutative join of the inputs. $+$ indicates the opposite, a commutative join.

Table 2. Lines of Code (LoC) changed or added to vLLM.

vLLM Source File	LoC
`core/block_pool.py`	68
`core/kv_cache_manager.py`	89
`core/kv_cache_utils.py`	73
`core/sched/output.py`	4
`core/sched/scheduler.py`	19
`core/single_type_kv_cache_manager.py`	4
`worker/gpu_model_runner.py`	232
Total	492

For example, Figure 6d shows how a user might amend the nested generation span query from Figure 6c to introduce a commutativity hint $+$. This hint serves to relax the ordering constraints, thus providing more opportunities for the optimizations covered in the remainder of this paper.

5 Improving KV Cache Locality

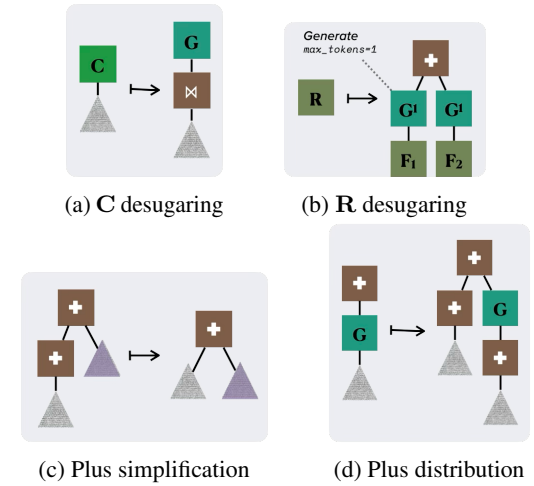
We now describe how span queries, with help from a small change to vLLM, can be optimized for KV cache locality. This involves four parts: a) a high-level optimizer that rewrites span queries; b) a span query tokenizer; c) a low-level optimizer that rewrites token sequences; d) the aforementioned small updates to vLLM. All told, the changes described in this section have a remarkably small impact on vLLM. We added or modified a total of 492 lines of Python source code out of 260,000 across all of vLLM. Table 2 breaks down our changes by source file.

5.1 High-level Cache Locality Optimizations

We break query optimization for KV cache locality into two layers, the first of which is covered here. This high-level optimizer operates on expression trees. It transforms them to be more explicit in their semantics, to resolve dual output paradoxes, and to simplify them to remove redundancies.

The high-level optimizer executes a suite of tree transformation in a fixed point iteration. Four such rules are illustrated in Figure 7, using the visualization from subsection 4.2. Each illustration shows the left-hand side and right-hand side of a query-to-query transformation. The optimizer looks for matches in a given tree, applies the first match, splices in the replacement sub-tree, and then iterates this process until convergence.

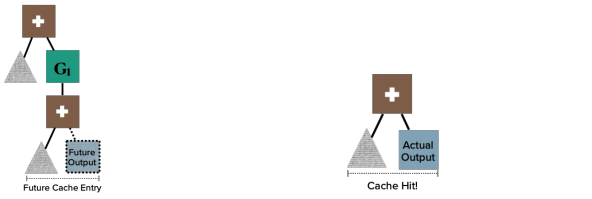
Figure 7a and Figure 7b show two of a family of desugaring rules. Figure 7a replaces a **C** with **G** (generate new tokens) linked to an explicit $\bowtie$ (non-commutative join of children messages). Figure 7b replaces a retrieval node **R** with an explicit commutative join $+$ over the retrieved fragments.



(a) **C** desugaring (b) **R** desugaring

(c) Plus simplification (d) Plus distribution

Figure 7. Some tree rewriting rules from the high-level optimizer.



(a) Optimized (b) Eventually presented to outer generate

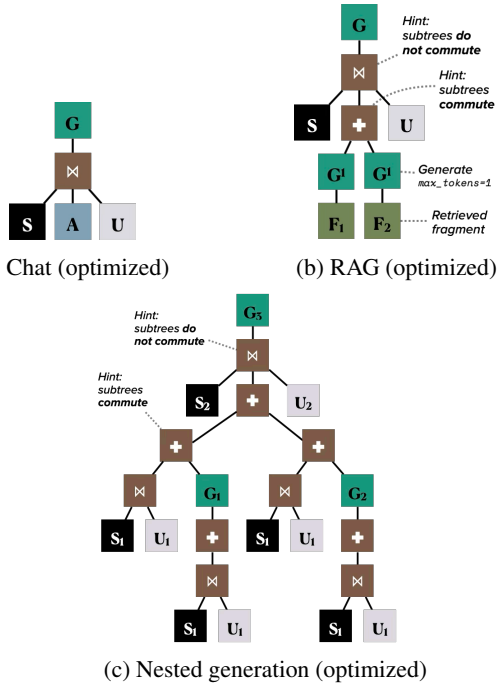
Figure 8. Plus distribution resolves the dual output paradox, because it aligns query structure with the behavior of model servers.

The desugaring rule for **R** relies on $\mathbf{G^1}$, which represents token generate **G** but with `max_tokens` parameter set to 1. This allows vLLM to *prepare* the fragments prior to their use in the full query.[2] Figure 9 illustrates the optimized forms of the chat completion and RAG queries from earlier.

Figure 7c shows one of a family of tree simplification rules. This rule leverages the commutativity of **+** to avoid an unnecessary chain of commutativity hints.

Figure 7d illustrates the "plus distribution" rule, which is used to resolve the dual output paradox. This rule is valid, because it is valid to distribute the commutative **+** across token generation **G**. The rule is also effective. Once we have distributed the **+** inside of the **G**, we ensure that future use of the *output* of the token generation **G** will be prefixed by the input; this is precisely the constraint underlying the dual output paradox. The effectiveness of this "ahead of time" planning is illustrated in Figure 8. Figure 9c illustrates the optimized form of the example query from Figure 6d.

[2]There is an asymmetry between cache lookup and insertion. It is straightforward to modify vLLM to avoid biasing at insertion time. It would be a monumental effort to change vLLM's behavior w.r.t. prefilling. Instead, we opt to "prepare" the fragments in a way that is independent of context. Span queries can express this.



(a) Chat (optimized) (b) RAG (optimized)

(c) Nested generation (optimized)

Figure 9. Optimized span queries. Notice that RAG, after optimization, becomes a special case of nested generation.

5.2 Query Tokenization

First, we must encode span queries into a form that vLLM can consume. This encoding step allows clients to operate on queries of arbitrary complexity, while allowing vLLM to do its job with minimal changes. Since the core of vLLM operates on token sequences, we must "parenthesize" a query. The parentheses will denote the boundaries of each sub-tree in the serialized form — c.f. (Merth et al., 2024).

We considered but ultimately decided against several approaches. Prior work (Ma et al., 2025) assumes that a list of spans is given as input. To support span queries of arbitrary nesting, we need something more general than that. We could introduce a separate reverse index data structure that maps sequence index to metadata about the sub-tree in which that token resides. It also may be possible to dig more deeply into the vLLM logic so that it accepts trees rather than sequences. We defer these approaches to future work, as they will necessitate large changes to vLLM.

Our chosen span query encoding produces minimal impact on vLLM. The general approach should not be surprising: employ special tokens to represent the parenthesization. Table 3 defines the special tokens. In practice, we can combine these so as to reduce the number of special tokens to two; e.g. `(` can be reused in place of `()`, and whitespace is usually a safe pad token.

Definition 4 (Span) *In a token sequence* $T = [T_1, \ldots, T_n]$,

Tokens	Meaning
□	Padding to achieve block alignment
(	Start of independent sub-tree
)(	Boundary between independent siblings
)n	End independent sub-tree of length n tokens

Table 3. Special tokens used by query tokenization.

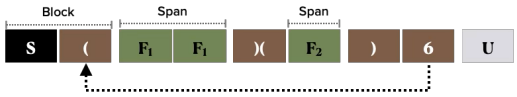


Figure 10. Tokenized RAG span query from Figure 9b. Block size of 2 tokens. S_1, U_1, F_2 are each messages of 1 token and F_1 occupies 2 tokens. The **6** points to the start of the `plus` sub-tree.

a span *is a subsequence of* $S \subseteq T$ *such that* $S = [T_i, \ldots, T_j]$ *and* T_i *and* T_j *are parentheses tokens.*

To further minimize changes to vLLM, we use a minor variant of this special token approach. We append to the end token) a number that points to the sequence position at which that serialized sub-tree begins. This design complicates the span query tokenization logic, but allows us to avoid some complexity in vLLM proper — code to maintain the parsing stack necessary to support nested parenthesization. This strategy is illustrated in Figure 10.

Some caution is needed when using special tokens. There is no free lunch. Novel special tokens require model fine tuning. Overloading the meaning of existing tokens may reduce accuracy. We take the latter approach.

5.3 Query Low-level Optimization

In order to maximize KV cache locality, reduce overheads, and simplify our vLLM footprint, a serialized span query may need to be rewritten. In this paper, we address two concerns: block alignment and coping with vLLM's propensity to avoid caching trailing partial blocks.

5.3.1 *Block Alignment*

We rewrite a serialized span query to ensure that the special tokens in a sequence are aligned to block boundaries. We do this by employing the padding special token from Table 3.

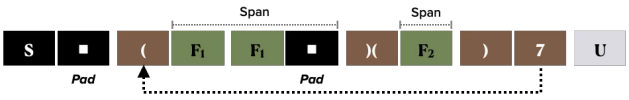
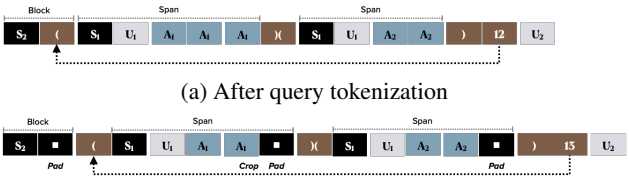


Figure 11. RAG span query after low-level optimizations are applied to the tokenized span query in Figure 10. The **7** points to the start of the `plus` sub-tree. Table 3 defines the special tokens.



(a) After query tokenization

(b) After low-level optimization pass

Figure 12. Tokenization and low-level optimization of the nested generation span query from Figure 9c. Block size of 2 tokens. Inner generates produce A_1 (3 tokens) and A_2 (2 tokens).

Figure 11 illustrates how padding the serialized query from Figure 10 can be used to ensure block boundary alignment.

By enforcing block alignment, the upcoming logic in subsection 5.4 can operate under that assumption. Rather than scanning every token in a block for special tokens, we can limit this search to the first token in every block. This also ensures that all special tokens in cached spans remain in KV cache, as they will only be part of partially filled blocks if the length of a span is smaller than the length of a block.

5.3.2 *Trailing Partial Blocks*

We may need to rewrite a serialized span query for the nested generation use case. If the output of an inner generate does not fully fill the last block, then there will be a disparity between what is sent back to the client and what is stored in KV cache — recall the dual output paradox from Figure 1.

This situation is visualized in Figure 12, which shows a tokenized span query for the nested generation example from Figure 9c. The output of the first inner generate is denoted A_1, which consists of three tokens in this example. Thus, the last token of the first inner generate will not be cached. Without deeper changes to vLLM (or to model fine tuning), we are forced to accept either cache misses (where, because of the dual output paradox, we send as input something different than was output to KV cache) or reduced accuracy (due to the need to crop, in order to prevent the paradox). For now, we accept that the third token of the first inner generate output A_1 must be cropped in order to get good cache locality.

To understand better why cache hit rate suffers and why padding alone cannot help, recall from section 2 that vLLM scans only for the *prefix* of the blocks of a token sequence that are cache hits. Upon reaching a block that is not in cache, it scans no further for cache hits. If we do not crop the third token of A_1, vLLM's prefix scan will find a cache miss on the block that starts with the third token of A_1.

5.4 Changes to the vLLM Scheduler Layer

Our first set of changes to vLLM is done to the "scheduler" layer of vLLM. This level of vLLM performs operations that are specific to a single client request. In particular, this logic includes KV cache management. In contrast, the layer below (subsection 5.5) deals with optimizing GPU utilization across client requests.

When scanning a block sequence for cache hits, vLLM normally accumulates a hash code. This is how it computes a hash for each block in a way that depends on the hash of preceding blocks. We update this logic to allow disabling the incorporation of the accumulated hash into a block's hash. If the first token in a block is `(`, i.e. the start of a span sub-tree, we suspend the accumulation logic. If the first token in a block is `)`, i.e. the end of a span sub-tree, we resume the accumulation logic.

5.5 Changes to the vLLM GPU Runner Layer

On a cache hit, the fetched block's KV vectors likely require updates before they can be reused. As visualized in Figure 2, Figure 3, and Figure 5, each cached KV vector has a positional encoding. If the position of a cached block (i.e. the position of its prior use) differs from its new position, adjustments are necessary. This adjustment is termed *repositioning* — a.k.a. a ReRoPE, which reverses and then re-applies a Rotational Positional Encoding (Su et al., 2024).

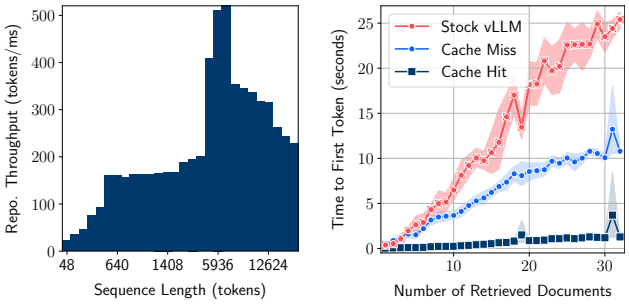
In order to implement a high-throughput ReROPE we must tap into the lower "GPU runner" layer of vLLM. This will allow us to perform cross-request batching and to avoid race conditions between requests.

5.5.1 CIDRA Repositioning Algorithm

We implemented a ReROPE algorithm in the GPU runner layer. Prior work (Ma et al., 2025) does not handle the scenario where multiple concurrent requests reuse overlapping sets of KV vectors. To address this, we introduce the Concurrent In-place Duplicating ReROPE Algorithm (CIDRA).

CIDRA first formulates the dependency graph of block repositionings — e.g. block A moved to block B, B to C, and so on. This allows CIDRA to operate *in-place*, minimizing scratch memory needed. If a node has out degree greater than 1 (signaling that concurrent requests need to relocate a block but to different positions), CIDRA will duplicate that block. In other words, CIDRA duplicates nodes until it is left with a strict permutation graph.

The algorithm then performs a strongly connected components analysis of the repositioning graph. This analysis identifies the cycles and exposes the independent subgraphs that can be batched (via bin packing) for parallel GPU execution. For large cycles exceeding the batch size, CIDRA falls back on CPU-based processing. This is rare in our experiments, where cycles are small (often size 2).



(a) Repositioning throughput (b) TTFT on RAG

Figure 13. CIDRA repositioning throughput and TTFT of span query execution (using CIDRA) in a RAG microbenchmark. The right-hand figure compares stock vLLM execution to one using span queries. Even on cache miss, span queries are much faster.

Finally, CIDRA optimizes tensor operations by concatenating layers for small batches. Figure 13a plots the repositioning throughput of CIDRA, showing a maximum throughput of 500 tokens per millisecond.

5.6 Accuracy on RAG

Having generalized prior work and implemented it in a production inference server, we now report on experiments that evaluate whether we have harmed accuracy. We do so by comparing to prior work on Block Attention (Ma et al., 2025) using the model from that prior work's experiments: `ldsjmdy/Tulu3-Block-FT`.

Table 4 summarizes the results on three RAG datasets. First is 2Wiki (Ho et al., 2020), which consists of 3000 documents. Second is NQ (12,000 documents). Both are prepared using the same methodology as in (Ma et al., 2025). Third is LongBenchV2 (Bai et al., 2025) (86 documents). For LongBenchV2, we had to filter the dataset to documents that fit in the 128,000-token context limit of the tested model. The first two use rather short documents (in the hundreds of tokens), while the last one emphasizes documents in the tens to hundreds of thousands of tokens. The experiments were run on an NVIDIA H100. All told, accuracy with span queries is on par with prior work. We also experimented with larger batch sizes, to probe effects along that dimension. As the table shows, no such degradations are apparent. Note that on LongBenchV2, we were limited to a batch size of 2 due to GPU memory limitations.

5.7 TTFT Speedup on RAG

Next, we quantify the benefits to KV cache locality on RAG. Our first experiment uses a microbenchmark to probe the limits of what is possible. In the microbenchmark, doc-

Implementation	Batch Size	Average Accuracy on Datasets 2Wiki (4 samples)	NQ (10 samples)	LongBenchV2 (10 samples)
BlockAttention	1	$68.36 \pm 0.09\%$	$62.36 \pm 0.24\%$	$20.69 \pm 2.37\%$
Span Queries	1	$69.39 \pm 0.07\%$	$61.58 \pm 0.28\%$	$20.57 \pm 2.95\%$
Span Queries	2	-	-	$21.16 \pm 1.80\%$
Span Queries	16	$68.14 \pm 0.80\%$	$61.63 \pm 0.45\%$	-

Table 4. RAG accuracy on three benchmarks (2Wiki, NQ, and LongBenchV2), comparing prior work on Block Attention (Ma et al., 2025) which used a custom inference library for their experiments, to Span Queries run on vLLM using an NVIDIA H100. The numbers denote average accuracy over the indicated number of samples. Each number includes the standard deviation across those samples. The Span Query experiments were run with Batch Sizes 1, 2, and 16. Due to VRAM limitations, we were limited to batch size 2 for LongBenchV2. The Block Attention implementation does not support batching.

uments have 2857 tokens (the mean length of a Python source file in vLLM) of randomly generated content. The microbenchmark runs three variants across a range of 1–32 documents and measures TTFT for each experiment. Each tuple in this explored space is measured 10 times. The first (baseline) variant runs vLLM in its stock configuration. The documents will never be cache local in this configuration. The second variant measures the TTFT of our stack, but when the documents are not cache local. The third variant is the same, but in the case where the documents are served by cache. This variant will help us to understand our overheads.

We expect the TTFT of the first two variants to grow at worst quadratically as the number of documents increases. In the baseline (stock vLLM), every token in every document needs to attend to every prior token. For the second variant (span queries, cache miss), the presence of spans means that attention is sparse. The tokens within a document only attend to prior tokens in that same document. Both are quadratic at worst, but the second variant should have a smaller constant factor. Therefore, we expect a large performance benefit, even if we miss cache. The third variant (span queries, cache hit) should grow at worst linearly, as we can skip prefill but pay for repositioning. The experimental results, shown in Figure 13b, bear this out.

5.8 Accuracy on Nested Generation

Next, we experiment with the accuracy of span query execution on the nested generation use case. The main threat here is with respect to the cropping done in subsubsection 5.3.2. Since vLLM does not cache trailing partial blocks, the logic to find the longest cached prefix match will stop at that first cropping point. This should not harm us if the important information comes before the cropping point. But will certainly be harmful to accuracy if the important information lies at the very end of the inner generated sequences. How often is this the case in practice?

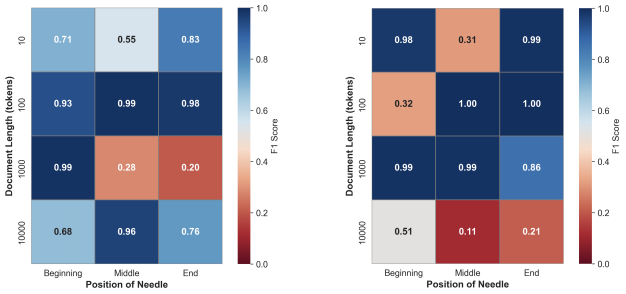
Unfortunately, existing benchmarks do not isolate this scenario. Therefore, we constructed a variant of the needle-in-haystack style of benchmark that allows us to place the needle either at the front, in the middle, or at the very end of tokens generated by the inner generates.

The without-cropping experiments use the normal OpenAI chat completion endpoint to process each of the inner and outer generate requests. The with-cropping experiments use our span query execution endpoint. Both use the model fine-tuned by prior work (Ma et al., 2025), ldsjmdy/Tulu3-Block-FT. That prior work proves that their fine tuning benefits accuracy, as it is trained to respond well to the "plus" tokens inserted by the low-level optimizer. It seems to embody the attention sparsity implied by a plus token.

Figure 14 shows the results of this nested generation accuracy experiment. Each cell of the heat maps represents the mean, over 100 experiments, of the F1 score of actual versus expected output. The expected output is the needle and the actual output is what is emitted by the outer generate. The grid of the heat map represents where the needle is placed (beginning, middle, or end) and the length of the documents in which the needle was inserted (10, 100, 1000, 10_000 characters). The heat maps are colored by the

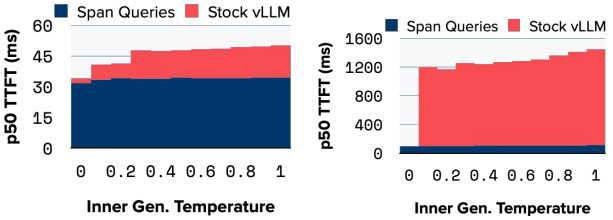
Due to the model fine tuning, span queries repeatably excel for shorter documents. In short documents, the beginning/middle/end distinction matters less, and the fine-tuned model responds well. For longer documents, especially when the needle occurs towards the end of the documents, accuracy suffers. This is unsurprising, as we may unceremoniously crop the needle.

In future work, we will explore an improved approach. Rather than cropping, we can use a custom logit processor to bias vLLM towards padding the output of inner generates so that they are block-aligned. This is in the same spirit as vLLM's `min_tokens` logit processor, which biases the sampler to continue until a certain number of tokens have been generated.



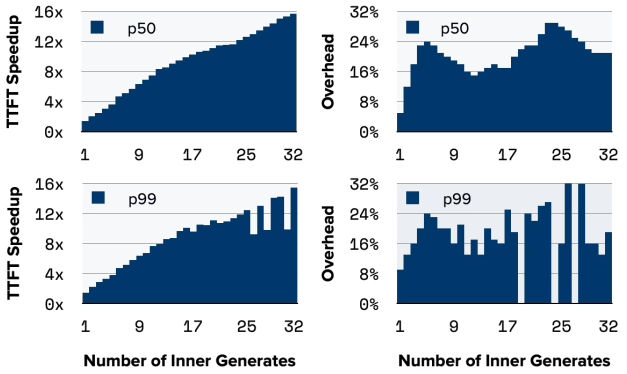
(a) Without cropping (b) With cropping

Figure 14. Accuracy on nested generation, without and with cropping When the important information is towards the end of longer documents, cropping hurts accuracy.



(a) 1 inner generate (b) 24 inner generates

Figure 15. Nested generation microbenchmark, showing median TTFT and varying inner generate temperature.



(a) TTFT speedup (b) Repositioning overhead

Figure 16. Nested generation microbenchmark, showing the distribution of TTFT speedup from span queries and CIDRA overhead.

Table 5. Bulk span query execution allows for scheduling the queries such that reused fragments are clustered temporally. Showing average speedup of TTFT by using span queries, for various bulk request sizes on two benchmarks.

BULK SIZE	2WIKI	NQ
1	1.21X	1.03X
1024	1.31X	1.05X
WHOLE CORPUS	1.59X	1.13X

5.9 TTFT Speedup on Nested Generation

Next, we expand the nested generation experiments to quantify the effects on prefill speed. The goal is to compare TTFT for stock vLLM with the span query stack, for varying number of inner generates (this corresponds to the fan-out of the + node) and for varying temperature of the inner generate. We should expect stock VLLM to perform well if the inner temperature is 0. In this case the output of the inner generates is the constant, and hence normal prefix caching works well. We should also expect stock vLLM to perform well if the fan-out is very small, because the cost of recomputation at some point becomes comparable to the cost of repositioning.

Figure 15 corroborates this hypothesis. In this experiment, we executed a judge/generator span query 100 times with varying inner generate temperature. As shown, when inner temperature is 0, span query benefit to TTFT is near 0. For any non-zero inner temperature, the TTFT benefit depends on the inner fan-out. For an inner fan-out of 1 (Figure 15a, the span query stack lowers TTFT by 7–32% depending on inner temperature. For an inner fan-out of 24 (Figure 15b), the span query stack has 12-13x faster TTFT.

Figure 16a digs into this data to verify that the span query stack has stable performance. In this experiment, we used an inner temperature of 0.5, varied the inner fan-out, and ran 500 query executions for each valuoe of fan-out. We again compared stock vLLM to our stack. As shown, the TTFT speedup is stable up to the 99th percentile (p99), at which point stability begins to suffer. Further work is therefore necessary to ensure stability beyond two 9's.

Figure 16b shows the overhead of CIDRA repositioning. Observe the non-monotonic shape to the curves, and how these nicely trace the non-monotonic shape to the CIDRA repositioning throughput shown in Figure 13a.

5.10 TTFT Speedups from Bulk Span Query Execution

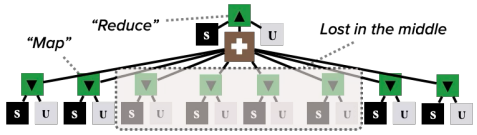
We have explored the TTFT benefits of executing one span query at a time. We now consider an additional layer of potential offered by a bulk query execution API. If the queries in a bulk request are executed in the given order, KV cache locality may suffer if the working set of the bulk is larger than KV cache capacity. To experiment with this possibility, we implemented a greedy heuristic that clusters the requests in a given bulk to increase temporal locality. Table 5 shows the average TTFT speedup (versus a stock vLLM baseline) of executing span queries in bulk. We explored two datasets: 2Wiki (Ho et al., 2020) and NaturalQuestions (Kwiatkowski et al., 2019) Both datasets are very small (each has around 100 tokens per fragment, placing them at the extreme left of

Figure 13b), hence as expected the speedup numbers align well with those presented in that figure.

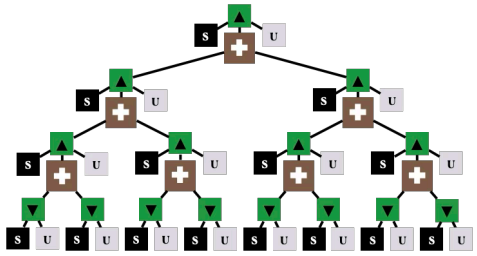
6 Improving Attention Locality

Finally, we show that the high-level optimizer from subsection 5.1 can be extended to target the "lost in the middle" problem (Liu et al., 2024). In this well-studied phenomenon, model accuracy is U-shaped as a function of the position of relevant information within a given input. Models perform well if the "needle" is located near the beginning or end of the sequence, but poorly if it is buried in the middle of the "hay". We consider *attention locality* as the success in finding the needle in the haystack.

Prior work observes that one can increase attention locality by increasing the sparsity of the attention matrix (Sun et al., 2025). We observe that a span query can be optimized to achieve the same effect. Unlike KV cache locality, where cache-optimized span queries rely on vLLM support, we now show how an attention-optimized span query can work without requiring any changes to the model server.



(a) Unoptimized query, but with commutativity hint

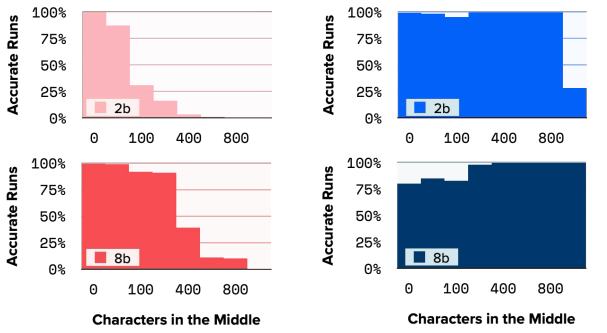


(b) Optimized for attention locality

Figure 17. High-level optimization for attention locality can transform a single 8-way judge/generator into 3 2-way judge steps.

Figure 17a shows a judge/generator span query that generates 8 candidates and then judges them with an outer generate (c.f. Figure 6d). The inner generates commute, as specified via +. Alternatively, we could perform a tree reduction: e.g. judge 2 at a time, then judge the output of each pair of 2-way judgments, and so on through 3 plies. This query is visualized in Figure 17b. We have automated this by adding a query rewrite rule to the high-level optimizer.

We implemented a simple needle-in-the-haystack microbenchmark. The benchmark executes the two queries shown above, where the inner generates produce random names (the needles) interspersed with randomly generated content (the hay). The judge is tasked to extract the names from the hay. We run each query 1000 times, varying the amount of hay between the needles and using two variants of the granite3.3 model: 2b and 8b. Our hypothesis is that the unoptimized queries should perform well for smaller values of hay, and then drop off a cliff — i.e. once the attention mechanism gets lost in the middle. Furthermore, the larger 8b model should tolerate more hay, as it is trained with a longer context length. And lastly, we hypothesize that the optimized queries should tolerate significantly higher amounts of hay. Figure 18 shows the results, which bear this out. Note how the 2b model, using the optimized query, fares better than the 8b model using the unoptimized query.



(a) Unoptimized query (b) Attention-optimized $k = 2$

Figure 18. Fraction of runs with good attention locality, as the length of the "lost in the middle" data increases.

7 Related Work

Researchers are actively looking at techniques to improve cache performance, which is especially critical as cache size grows beyond GPU memory limits. Block Attention (Ma et al., 2025) decomposes operations into discrete computational blocks, enhancing locality and enabling modular cache segmentation. CacheBlend (Yao et al., 2025) allows repositioning of fixed size cache blocks with greatly reduced re-calculation of attention between blocks while maintaining high accuracy. Superposition prompting improves RAG cache performance by breaking a prompt into parallel paths which allows for more selective attention. SGLang (Zheng et al., 2024) introduces a programming library to simplify LLM calling and allow more cache-aware optimizations. Research in Graphs-of-Thought (Besta et al., 2024) structures LLM processing into a graph structure that is inherently multi-generate and potentially allows for cache-aware placement of computation. Span Queries captures and extends these ideas into a declarative framework to co-optimize selective application of attention and cache locality.

8 Future Work

We believe that nested model calls are an increasingly relevant pattern. For example, nested tool calls (Basu et al., 2025), calls to sub-agents or other compound-AI decompositions (Davis et al., 2024), and techniques that combine multiple sampled or parallel model outputs to improve accuracy and reliability (Snell et al., 2025; Hu et al., 2026) are all naturally expressed as span queries. This suggests that span queries could help avoid large amounts of repeated prefill work in these settings.

Another natural fit for span queries would be integration with datacenter-scale LLM inference engines such as DeepServe (Hu et al., 2025), Mooncake (Qin et al., 2024), and llm-d (Ayoub et al., 2025), where entire corpora of documents or chunks could be uploaded, prefilled, and stored in anticipation of retrieval requests through RAG or agentic systems.

9 Conclusion

We introduced the span query as an intermediate representation for analysis and optimization. Our focus to date has been on locality, for KV cache and for attention. There are many prospects for other optimization avenues. For example, is there any reason a nested generation structure should flow through KV cache at all? One could imagine analyzing a span query with nested generation to produce an optimized scheduler, where the inner and outer generates are arranged in a gather/scatter structure (c.f. SQL). This kind of optimization would benefit from span queries, because the query provides the full scope of the connectivity and ordering constraints — the kind of information needed to make gather/scatter work well. This is one of many exciting avenues of exploration enabled by span queries.

References

Ayoub, M., Harnik, D., Smith, T., Swain, K., Wang, X., Yin, H., and Yan, K. From prefix caching in vllm to distributed scheduling with llm-d. https://llm-d.ai/blog/kvcache-wins-you-can-see, September 2025. URL https://llm-d.ai/blog/kvcache-wins-you-can-see. Accessed: 2025-10-25.

Bai, Y., Tu, S., Zhang, J., Peng, H., Wang, X., Lv, X., Cao, S., Xu, J., Hou, L., Dong, Y., et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Basu, K., Abdelaziz, I., Lastras, L., and Kapanipathi, P. A benchmark for evaluating llms on nested sequences of api calls. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.1702/.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P., and Hoefler, T. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16): 17682–17690, Mar 2024. doi: 10.1609/aaai.v38i16.29720. URL https://ojs.aaai.org/index.php/AAAI/article/view/29720.

Davis, J. Q., Hanin, B., Chen, L., Bailis, P., Stoica, I., and Zaharia, M. Networks of networks: Complexity class principles applied to compound ai systems design, 2024. URL https://arxiv.org/abs/2407.16831.

Dean, J. and Ghemawat, S. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

Ho, X., Duong Nguyen, A.-K., Sugawara, S., and Aizawa, A. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Scott, D., Bel, N., and Zong, C. (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Hu, J., Xu, J., Liu, Z., He, Y., Chen, Y., Xu, H., Liu, J., Meng, J., Zhang, B., Wan, S., Dan, G., Dong, Z., Ren, Z., Liu, C., et al. Deepserve: Serverless large language model serving at scale, 2025. URL https://arxiv.org/abs/2501.14417.

Hu, J., Zhang, Y., Shang, S., Yang, X., et al. Pacore: Learning to scale test-time compute with parallel coordinated reasoning, 2026. URL https://arxiv.org/abs/2601.05593.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Li, H., Li, Y., Tian, A., Tang, T., Xu, Z., Chen, X., Hu, N., Dong, W., Li, Q., and Chen, L. A survey on large language model acceleration based on kv cache management, 2025. URL https://arxiv.org/abs/2412.19442.

Liang, L., Sun, M., Gui, Z., Zhu, Z., Jiang, Z., Zhong, L., Qu, Y., Zhao, P., Bo, Z., Yang, J., Xiong, H., Yuan, L., Xu, J., Wang, Z., Zhang, Z., Zhang, W., Chen, H., Chen, W., and Zhou, J. Kag: Boosting llms in professional domains via knowledge augmented generation, 2024. URL https://arxiv.org/abs/2409.13731.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Ma, D., Wang, Y., and Lan, T. Block-attention for efficient prefilling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7zNYY1E2fq.

Merth, T., Fu, Q., Rastegari, M., and Najibi, M. Superposition prompting: Improving and accelerating retrieval-augmented generation, 2024. URL https://arxiv.org/abs/2404.06910.

PwC. Pwc's ai agent survey. Tech Effect: AI & Analytics, PwC US, May 2025. URL https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-agent-survey.html. Survey of 308 US executives conducted Apr 22–28, 2025.

Qin, R., Li, Z., He, W., Zhang, M., Wu, Y., Zheng, W., and Xu, X. Mooncake: A kvcache-centric disaggregated architecture for llm serving, 2024. URL https://arxiv.org/abs/2407.00079.

Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., and Manning, C. D. Raptor: Recursive abstractive processing for tree-organized retrieval, 2024. URL https://arxiv.org/abs/2401.18059.

Snell, C. V., Lee, J., Xu, K., and Kumar, A. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Sun, Y., Li, Z., Zhang, Y., Pan, T., Dong, B., Guo, Y., and Wang, J. Efficient attention mechanisms for large language models: A survey, 2025. URL https://arxiv.org/abs/2507.19595.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yao, J., Li, H., Liu, Y., Ray, S., Cheng, Y., Zhang, Q., Du, K., Lu, S., and Jiang, J. Cacheblend: Fast large language model serving for rag with cached knowledge fusion, 2025. URL https://arxiv.org/abs/2405.16444.

Zheng, L., Yin, L., Xie, Z., Sun, C., Huang, J., Yu, C. H., Cao, S., Kozyrakis, C., Stoica, I., Gonzalez, J. E., Barrett, C., and Sheng, Y. Sglang: Efficient execution of structured language model programs, 2024. URL https://arxiv.org/abs/2312.07104.

Zhou, Y., Xu, A., Wang, P., Xiong, C., and Joty, S. Evaluating judges as evaluators: The jetts benchmark of llm-as-judges as test-time scaling evaluators, 2025. URL https://arxiv.org/abs/2504.15253.