Pre^3: Enabling Deterministic Pushdown Automata for Faster Structured LLM Generation

Anonymous ACL submission

Abstract

Extensive LLM applications demand efficient structured generations, particularly for LR(1) grammars, to produce outputs in specified formats (*e.g.*, JSON). Existing methods primarily parse LR(1) grammars into a pushdown automaton (PDA), leading to runtime execution overhead for context-dependent token processing, especially inefficient under large inference batches. We therefore propose Pre^3 that exploits deterministic pushdown automata (DPDA) to optimize the constrained LLM decoding efficiency. First, by **pre**computed **pre**fix-conditioned edges during the **pre**processing, Pre^3 enables additional preprocessing optimizations for edges and supports parallel transition processing. Second, Pre^3 proposes an algorithm to transform LR(1) transition graphs into DPDA, eliminating the need for runtime path exploration, enabling edge transitions with minimal overhead. Pre^3 can be seamlessly integrated into standard LLM inference frameworks, improving time per output token (TPOT) by up to 40% and throughput by up to 36% in our experiments.

1 Introduction

The recent remarkable development of Large Language Models (LLM) has ushered in new opportunities for a wide array of intelligent applications such as automated external tool invocations through function calls (Cai et al., 2023; Li et al., 2024a; Zhuo et al., 2024), chain of thoughts (Wei et al., 2022; Wang et al., 2022; OpenAI, 2024; Guo et al., 2025), embodied AI (Duan et al., 2022; Brohan et al., 2023; Yang et al., 2024b) et al. These applications created substantial demand for LLM systems to perform structured generation and produce outputs adhering to specific formats, such as JSON or other structures. Downstream applications can accordingly utilize these structured outputs to engage in downstream system interactions (Cho et al., 2023).

Constrained decoding (Hu et al., 2019; Scholak et al., 2021) is a widely used method in structured generation tasks (Willard and Louf, 2023b; Dong et al., 2023; Rückstieß et al., 2024) that excludes invalid tokens at each step by applying a *probability mask* to zero out their sample possibility. Flexible mechanisms like LR(1) grammars (Francis, 1961; Knuth, 1965) are often employed to handle diverse and complex structural constraints, as they allow recursive rule definitions that surpass the limitations of regular expressions. However, this flexibility comes at the cost of degraded efficiency: Each decoding step requires parsing the grammar for all candidate tokens in a potentially large vocabulary. Additionally, tokens generated by LLM may consist of multiple characters that span across grammar rule boundaries, further complicating the generation process and demanding dedicated execution stack management. Both of them lead to significant computational overhead. These challenges raise the need to optimize constrained decoding efficiency without affecting LLM generation fidelity, making it more applicable in real-world applications.

Current state-of-the-art (SOTA) methods for constrained decoding acceleration, such as XGrammar (Dong et al., 2024), primarily focus on parsing LR(1) grammars into a pushdown automaton (PDA) (Nederhof and Satta, 1996). A PDA consists of multiple finite state automata (FSA), each representing a grammar rule, with the stack handling recursive rule expansions. These methods achieve substantial speedups by precomputing masks while managing transitions through pushdown automata. However, they overlook the inherent properties of LR(1) grammars, which can be equivalently transformed into a deterministic pushdown automaton (DPDA) (Valiant, 1973, 1975).

The primary issue with traditional PDA-based approaches (Koo et al., 2024; Park et al., 2025a; Dong et al., 2022; Willard and Louf, 2023a; Li et al., 2024b) stems from the non-deterministic nature of the PDA's edges. Although these methods precompute masks based on the PDA structure, this design introduces two critical limitations. First, the non-deterministic edges depend on runtime contextual information to resolve transitions, resulting in incomplete precomputed masks for *context-dependent tokens*. The computation of context-dependent tokens necessitates backtracking, speculative operations, and the maintenance of a *persistent stack* (merges all past stacks into a tree, with each stack as a root-to-node path) during runtime. As batch sizes increase, the overhead from these runtime computations grows significantly, severely degrading decoding efficiency. Second, previous methods cannot effectively optimize non-deterministic transitions during preprocessing for they will dynamically change during runtime. This limitation hinders their ability to fully exploit the potential of the parsing method, leading to suboptimal performance.

To address these challenges, we propose Pre^3, a constrained LLM decoding approach based on a deterministic pushdown automaton (DPDA). Unlike traditional methods , we design an algorithm to directly build a DPDA from the LR(1) grammar. Leveraging the deterministic nature of the DPDA's edges, our approach resolves the aforementioned limitations. First, the determined transitions in the DPDA eliminate the context-dependent tokens, further entirely eliminating the need for backtracking, speculative exploration, and the maintenance of a persistent stack. This fundamentally reduces the runtime computational overhead associated with transitions. Second, since all transition edges in the DPDA are available during preprocessing, we can perform comprehensive optimizations on the automaton in advance. Additionally, for the stack-matched transition mechanism of the DPDA, we design a parallel verification method for transitions, which accelerates inference. Together, these innovations result in a more efficient and scalable constrained decoding framework.

In summary, the paper's main contributions are:

- We firstly propose an algorithm to transform LR(1) state transition graphs into DPDA, eliminating runtime exploration and enabling edge transitions with minimal overhead.
- We enables additional optimizations for edges and supports parallel transition processing by pre-computing prefix-conditioned edges.
- We integrate Pre^3 into mainstream LLM inference systems and achieve up to 40% improvement in time per output token (TPOT) and increase throughput by up to 36% with high scalability into large batch sizes.

2 Preliminaries and Background

2.1 LLM Constrained Decoding

Constrained decoding (Hu et al., 2019) enforces strict adherence to predefined structural grammars by aligning the LLM's output with syntactic rules. At each decoding step, it assigns a probability of negative infinity to tokens that violate the grammar, ensuring valid token selection. This guarantees structurally compliant outputs but faces challenges like grammar diversity, large vocabulary sizes, and complex token-to-text mappings, which complicate implementation and increase computational overhead.

Several constrained decoding implementations have been proposed, but most exhibit limitations in large batch-size inference scenarios. For example, frameworks like llama.cpp (Gerganov, 2023) inefficiently verify tokens during runtime, causing computational bottlenecks. Approaches like Outlines (Willard and Louf, 2023b) and SynCode (Ugare et al., 2024) suffer from boundary mismatch issues and suboptimal efficiency. The current SOTA work XGrammer (Dong et al., 2024) excels in correctness and speed for batch size=1, but its overhead increases with larger batch sizes. GreatGramma (Park et al., 2025b) efficiently supports complex grammars but only discusses scenario where batch size equals 1.

2.2 LR(1) Grammar and State Transition Graphs

In constrained decoding scenarios, most grammars can be classified as LR(1) grammars, which are fundamental to bottom-up parsing and align naturally with the token-by-token generation process of large language models (LLMs). LR(1) grammars are a powerful subset of context-free grammars capable of describing the syntax of most programming languages. They are characterized by their ability to handle deterministic parsing with a single lookahead symbol, making them highly expressive and widely applicable. Nearly all context-free grammars can be converted into LR(1) form, which ensures their versatility in modeling structured languages. This property, combined with their alignment with bottom-up parsing methods, makes

LR(1) grammars a cornerstone in constrained decoding and syntactic analysis tasks.

LR(1) items are tuples of the form $[A \to \alpha \cdot B\beta, a]$, where $A \to \alpha \cdot B\beta$ represents the parsing progress of a production rule, and a is a lookahead symbol used to determine when a reduction should occur. The **closure** operation constructs LR(1) item sets by adding items for non-terminals and their productions, ensuring all possible derivations are considered. The **Goto function** generates the LR(1) state transition graph by moving the dot in items past a grammar symbol X and computing the closure of the resulting items, thereby connecting states to form the LR(1) automata. This process continues until no new states are generated, creating a complete parsing structure for the grammar.

2.3 Pushdown Automata and Deterministic Pushdown Automata

Pushdown automata (PDA) are a class of abstract machines that extend finite automata with an unbounded stack memory, enabling them to recognize context-free languages (CFLs) (Hopcroft et al., 2001). Formally, a PDA is defined as a 7-tuple $(Q, \Sigma, \Gamma, \delta, q_0, Z_0, F)$, where Q is a finite set of states, Σ is the input alphabet, Γ is the stack alphabet, $\delta : Q \times (\Sigma \cup \{\epsilon\}) \times \Gamma \to \mathcal{P}(Q \times \Gamma^*)$ is the transition function, q_0 is the initial state, Z_0 is the initial stack symbol, and $F \subseteq Q$ is the set of accepting states. The non-deterministic transition function δ allows PDAs to handle ambiguous structures inherent to context-free grammars (CFGs), such as nested parentheses or recursive syntactic patterns.

A deterministic pushdown automaton (DPDA) is a restricted variant where, for every state $q \in Q$, input symbol $a \in \Sigma$, and stack symbol $Z \in \Gamma$, the transition function $\delta(q, a, Z)$ yields at most one possible move, and ϵ-transitions (stack operations without consuming input) are permitted only if no input-consuming transition is available (Sipser, 1996). This determinism ensures unique computation paths, making DPDAs equivalent to the class of deterministic context-free languages (DCFLs), which are unambiguous and efficiently parsable. As mentioned earlier, the vast majority of grammars in the constrained decoding scenario can be represented by LR(1), which is a true subset of DCFL and can be recognized by DPDA (ASU86 et al., 1986; Sipser, 1996). Compared to PDA, DPDA avoided backtracking and non-deterministic search overhead, which can significantly improve the efficiency of constrained decoding.

3 Pre^3 Design

Our proposed method, Pre^3, is a DPDA-based constrained decoding solution that leverages a novel approach for constructing a DPDA from a given LR(1) grammar. The method operates by first transforming the LR(1) grammar into an LR(1) state transition graph, which is then converted into a DPDA using the techniques introduced in this section. This DPDA can be directly utilized for constrained decoding, enabling efficient and effective decoding. The complete workflow of our method is illustrated in Figure 1.

In Section 3.1, we introduce the Prefix-conditioned Edge, a novel mechanism ensuring uniqueness by matching both prefix information and input symbols, unlike traditional PDA transitions. In Section 3.2, we design an algorithm to compute all LR(1) state transitions, incorporating Prefix-conditioned Edge and addressing cyclic structures, successfully constructing a DPDA. In Section 3.3, we optimize the DPDA's structure and performance through preprocessing, leveraging its pre-determined edges.

3.1 Prefix-conditioned Edges

Constrained decoding with LLMs faces challenges due to non-deterministic transitions in PDA, where the same input symbol can trigger multiple transitions based on prior symbol sequences. This non-determinism complicates computation by requiring speculative exploration, backtracking, and a persistent stack to store historical context, increasing overhead. To resolve these issues, eliminating non-determinism in transitions is crucial for enabling preprocessing optimizations and efficient runtime computation.

A fundamental property of LR(1) grammars is that **the current stack configuration and a single lookahead symbol are sufficient to uniquely determine the next action**. This property provides a theoretical foundation for introducing determinism into the automaton's transition edges. Building on this insight, we propose the Prefix-conditioned Edge, as illustrated in Figure 2.

By simultaneously considering the input symbol and the prefix of accepted symbols (represented by the stack's state), we uniquely determine the target state for each transition. To achieve this, our method enhances each edge with three key

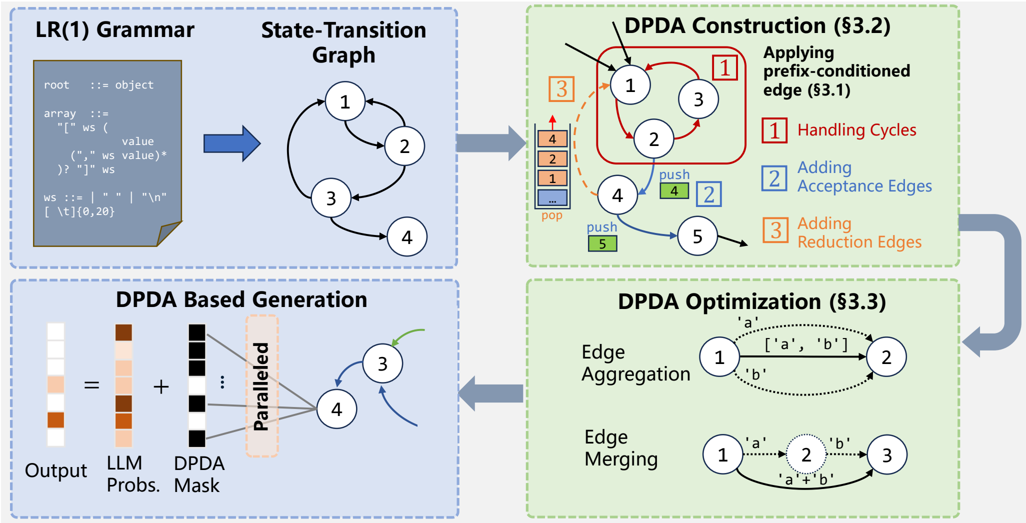

Figure 1: Overview of Pre^3: The figure depicts the workflow from LR(1) grammar to DPDA-based generation, encompassing DPDA construction and optimization steps.

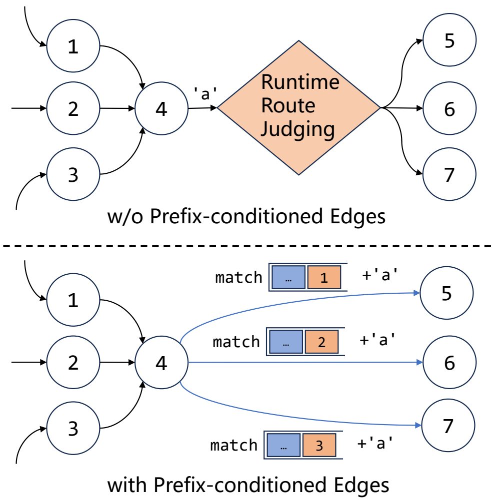

Figure 2: This diagram illustrates prefix-conditioned edges: above shows the case before calculation, where ‘a’ is a context-dependent token requiring runtime context for transition; below shows the precomputed case, where each edge includes a stack-matching condition, uniquely determining the transition path via the condition and transition symbol.

components:

- **Accepted Symbol**: The input symbol that triggers the transition.
- **Stack Matching Condition**: The specific prefix of the stack required for the transition to be valid.
- **Stack Operations**: Actions such as push to update the stack during the transition, which is both required by PDAs and DPDAs.

Notably, although the additional stack-matching conditions introduced to the edges increase complexity, we address this challenge by implementing a parallel algorithm capable of simultaneously verifying multiple stack-matching conditions, effectively resolving this issue.

3.2 Cycle-aware Deterministic Pushdown Automata Construction

To avoid the additional exploration overhead at runtime, we aim to construct a DPDA based on LR(1) grammars. However, building a DPDA is non-trivial and requires a systematic approach. In this section, we introduce our step-by-step algorithm for constructing a DPDA from an LR(1) state transition graph, leveraging the prefix-conditioned edge to ensure determinism.

3.2.1 DPDA Structure

We begin our algorithm with the state transition graph generated from the LR(1) grammar, where the nodes represent the LR(1) item set family and the edges indicate the acceptance of a symbol when traversing from one node to another. Building on this foundation, we construct the DPDA by retaining the node definitions from the LR(1) transition graph but redefining the edges into two distinct types: *acceptance edges* and *reduction edges*, as shown in Figure 3.

- **Acceptance Edges** are the simplest type of transition in our DPDA. These edges are directly derived from the original state transition graph of the LR(1) grammar. In the context of LR(1) parsing, an acceptance edge corresponds to a shift operation, where the automaton consumes an input symbol from the input stream and pushes it onto the stack while transitioning to a new state. This operation reflects the fundamental step of recognizing and accepting a terminal symbol in the input, advancing the parsing process.

- **Reduction Edges** model reduction operations in

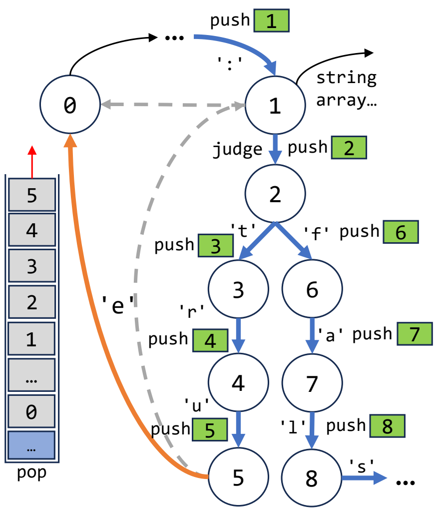

Figure 3: This diagram shows the two edge types for DPDA computation: blue edges are acceptance edges (existing in the original LR(1) graph, handling stack operations for acceptance); orange edges are reduction edges (added to the DPDA, matching and popping stack operations for reductions); gray edges depict LR(1) reduction paths, demonstrating fewer nodes needed for reduction after state machine construction.

LR(1) parsing. In traditional LR(1) parsing, reductions involve replacing a sequence of terminal symbols with a non-terminal symbol according to the grammar rules. However, nested grammar rules often require multiple reduction steps, leading to inefficiencies. Reduction edges address this by directly encoding reduction operations as single-step transitions during the pre-processing phase. These edges connect reduction targets, enabling the automaton to handle nested reductions efficiently.

3.2.2 Acceptance Edges and Reduction Edges Integration

The state transition graph alone cannot function as a DPDA because it only supports shift operations (*i.e.*, symbol acceptance) and lacks reduction operations, while some edges also suffer from nondeterminism. To address these issues, we not only compute all possible transition edges, including both shift and reduction edges, to complete the missing reduction paths, but also leverage prefix-conditioned edges to incorporate stack conditions into each transition, resolving nondeterminism and enabling the transformation of the nondeterministic state transition graph into a DPDA.

Adding Acceptance Edges: Acceptance edges do not need to consider determinism because the construction of the LR(1) state transition graph ensures that no node will have two identical transitions. As a result, when an acceptance edge is encountered, the target node's state information is simply pushed onto the runtime stack. The algorithmic flow of this operation is described in Algorithm 1, lines 6–8.

Adding Reduction Edges: Based on the definition of reduction edges, we can employ a two-step method to add all necessary reduction edges to the automaton, which is described in Algorithm 1, lines 9–18.

First, we identify ϵ-reduction transitions, representing unconditional reductions, and add them to the automaton to handle mandatory reductions. These transitions backtrack along their path, popping states until reaching the reduction endpoint. However, their lack of accept symbols introduces ambiguity, violating the DPDA's determinism. To ensure completeness, this process is applied recursively, generating all necessary reduction edges by traversing the state transition graph.

Second, we resolve indeterminism by merging ϵ-reduction edges with compatible acceptance edges, ensuring aligned stack operations and reduction targets, and assigning appropriate accept tokens to satisfy the Prefix-condition.

3.2.3 Solving Issues with Automaton Cycles

LR(1) grammars are highly expressive and can handle complex language constructs, including the acceptance of cyclic symbol sequences. However, cycles introduce significant challenges when constructing a DPDA.

During the precomputation of reduction edges, cycles create a critical issue: repeatedly traversing a cycle generates an infinite number of potential reduction paths. This makes it computationally infeasible to add all necessary reduction edges. Figure 4 visually illustrates how cycles in the automaton can lead to infinite reduction paths.

Through further observation, we note that during the reduction process, specifying an entry node and an exit node uniquely determines the path along which the reduction occurs. This property allows us to disregard the number of cycle traversals, as even a single traversal of the cycle does not need to be explicitly recorded.

We propose a solution that simplifies the reduction process as follows: Suppose we have a detected cycle with the reduction problem $C = (s_1, s_2, s_3, \ldots, s_n, s_1)$. We define the *back-edge* as $s_n \rightarrow s_1$. While handling the cycle, we modify this back-edge by introducing an additional

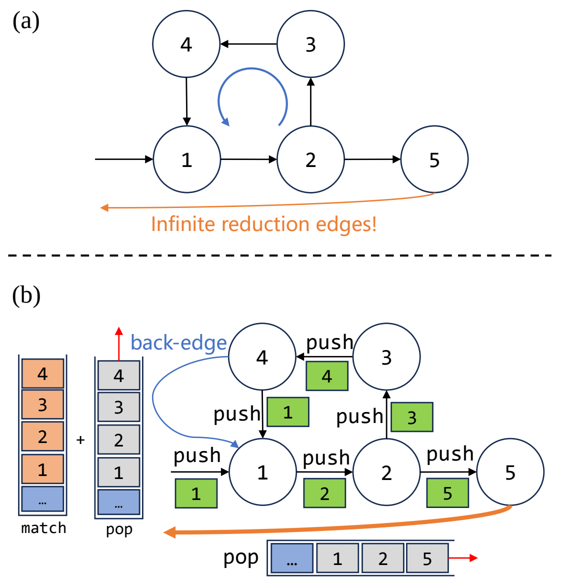

Figure 4: (a) illustrates pushdown automaton with an infinite cycle between State 1, 2, 3, 4, leading to an infinite number of possible paths and indeterminable transition paths when adding reduction edges at State 5. (b) shows how our method handles the cycle issue: The back-edge from State 4 to State 1 is modified to check for complete cycle traversal information (*e.g.*, [1, 2, 3, 4]) in the stack. If detected, it pops the redundant state (*e.g.*, [1, 2, 3, 4]), ensuring reduction edges at State 5 only need to account for traversals without cycles.

stack operation: a pop operation for the sequence $(s_1, s_2, \ldots, s_n)$. This modification enables efficient handling of cyclic traversals.

Furthermore, by checking whether all vertices traversed in a single cycle are fully present in the execution stack, we ensure that the stack retains only the necessary information from outside the cycle traversal. Specifically, if a complete traversal of the cycle is detected, the stack information corresponding to the current traversal is popped immediately. This guarantees that the stack never accumulates redundant context from repeated cycle traversals.

This approach, described in Algorithm 1, lines 1–5, guarantees that the system reverts to an equivalent state after each complete traversal, avoiding infinite reduction edges. As a result, the automaton can handle cycles efficiently without compromising determinism or computational feasibility.

3.3 Edge Optimization with Prefix-condition

Building on the DPDA constructed in Section 3.2, we can further perform various optimizations. Since all transition edges in the DPDA are deterministic and can be uniquely resolved by matching both the stack state and input symbols, we are able to analyze and optimize the automaton's structure

---

Algorithm 1: Construct DPDA from LR(1) Transition Graph

Input: LR(1) State Transition Graph $G = (S, E)$
Output: Deterministic Pushdown Automata (DPDA)

`/* Step 1: Cycle Handling */`
1 $C \leftarrow$Detect cycles with reduction problem in G
2 **foreach** *detected cycle* $C = (s_1, s_2, ..., s_n, s_1)$ **do**
3   **if** *C corresponds to recursive reduction of non-terminal A* **then**
4   Define the back-edge: $s_n \xrightarrow{\text{back}} s_1$
5   Modify the back-edge to check for complete cycle traversal in the stack: match and pop $(s_1, s_2, ..., s_n)$, push(s_1)

`/* Step 2: Acceptance Edge Generation */`
6 **foreach** *state* $s_i \in S$ **do**
7   **foreach** *valid transition* $s_i \xrightarrow{X} s_j$ *in* E **do**
8   Add stack operation: push(s_j)

`/* Step 3: Reduction Edge Generation */`
9 **Function** `GenerateReductionEdges(`*state* s_i`):`
10   **foreach** *reduction sequence* $s_i \xrightarrow{\textit{reduce } A} s_j \xrightarrow{\textit{reduce } B} s_k$ **do**
11   Merge into a direct transition: $s_i \xrightarrow{\text{reduce } A \to B} s_k$
12   Validate stack compatibility
13   `GenerateReductionEdges(`s_k`)`
14   **foreach** *ϵ-reduction edge from s_i* **do**
15   Merge the ϵ-reduction edge with appropriate acceptance edges that share the same stack operations
16   Assign suitable accept tokens to ensure the Prefix-condition is matched
17   `GenerateReductionEdges(`*target state of the merged edge*`)`
18 `GenerateReductionEdges(`*initial state* s_0`)`

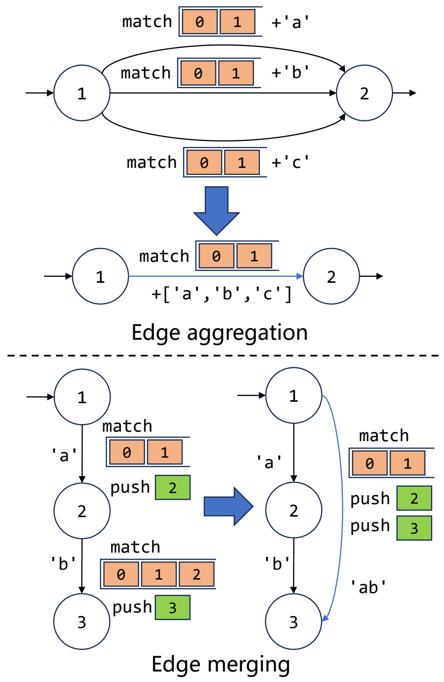

Figure 5: Two different types of edge optimization.

during the preprocessing phase. In contrast, traditional methods based on non-deterministic push-down automata (PDA) cannot achieve such optimizations during preprocessing due to the ambiguity of transition edges—where the same input symbol may lead to multiple possible transition targets. As a result, we can aggregate and merge transition edges as shown in Figure 5.

- **Edge Aggregation:** Edges with the same stack prefix condition and stack operations but different accepted symbols can be combined. For example, in grammars describing numbers, edges for digits 0-9 can be merged into a single edge accepting all digits to simplify the automaton.

- **Edge Merging:** If two edges share the matched stack prefix condition and operations, we connect them directly, skip intermediate states, and reduce transitions. This is important for LLM with large vocabularies, as it allows "jumping" to the desired state in fewer steps, leveraging the LLM's vocabulary for efficient parallel validation of transition conditions.

These optimizations are enabled by precomputed prefix-conditioned edges for all stack conditions, so eliminate runtime decisions. By combining these techniques, we further optimize the DPDA, achieving deterministic and efficient grammar parsing.

4 Evaluation

4.1 Experiment Setup

Implementation: We implemented our approach in 2,000 lines of Python code and about 1,000 lines of C++ code, and we seamlessly integrated with LightLLM (ModelTC, 2023), a popular LLM inference framework.

Hardware Setup: All the experiments are tested on a server with Intel(R) Xeon(R) Gold 6448Y CPU and 8 NVIDIA H800 GPUs. Depending on the scale of the experiment, we use different numbers of GPUs.

Baselines: We choose the following representative works on LR(1) grammar constraint decoding.

- **XGrammar:** An open-source library for structured generation in large-language models. It significantly enhances performance in tasks like JSON grammar generation with reduced latency and storage.
- **Outlines:** A text generation library, it offers a Python tool for grammar-guided generation, offering a fast generation method. We use vLLM integrated with Outlines for evaluation.
- **Llama.cpp**: A C/C++-based LLM inference tool, and also includes support for LR(1) grammar constraint decoding.

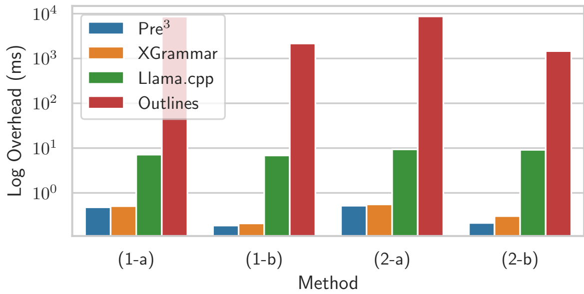

Figure 6: Per-step decoding overhead cross different grammar and models. Outlines incurs an overhead of up to several seconds per step. Experiments contain (1) Chain-of-Thought grammar (2) JSON grammar. Models contain (a) Meta-Llama-3-8B on 1×H800 (b) Meta-Llama-2-70B on 4×H800.

Datasets: In our experiments, we utilized the JSON-mode-eval (NousResearch, 2024) dataset from NousResearch and jsonschemabench (Geng et al., 2025) from guided-ai as prompts. As there is a scarcity of datasets for structured output, we collected some private data additionally and incorporated it into the test dataset.

4.2 Per-step Decoding Efficiency

To evaluate the improvement of our system, we first examine the per-step decoding overhead, which is defined by subtracting the original decoding time from the grammar-based decoding time. We design four experiment setups including two models, Meta-Llama-3-8B and Meta-Llama-2-70B, and two grammars, JSON and chain-of-thought. For comparison, we benchmark our method against several state-of-the-art and popular structure generation engines, including XGrammar, Outlines, and llama.cpp-Grammar, to demonstrate the efficiency of our system at a per-step scale.

The results are shown in Figure 6 and Table 2. Pre^3 demonstrates a superiority over Outlines and llama.cpp with approximately 1000× reduction, and Pre^3 remains a consistent advantage over XGrammar. The results indicate that Pre^3 introduces less overhead than previous SOTA systems.

4.3 Large-batch Inference Efficiency

In real-world serving scenarios, inference often handles large batches of requests simultaneously, making large-batch efficiency crucial for deploying language models at scale. We evaluate performance in such settings, where efficiency gains

Table 1: Decode batch inference time comparison between our method and XGrammar. The "-" marker stands for the batch size cannot be executed on the given hardware setup.

	Batch Size	16	32	64	128	256	512	1024
Llama-3-8B (Dubey et al., 2024)	XGrammar (ms)	15.19	43.69	52.07	65.21	90.98	147.64	272.77
	Pre^3 (ms)	11.77	31.12	35.88	45.32	64.42	104.46	201.16
2×H800	Reduction	**↓22.49%**	**↓28.78%**	**↓30.09%**	**↓30.50%**	**↓29.20%**	**↓29.24%**	**↓26.25%**
DeepSeek-V2-Lite-Chat (Liu et al., 2024)	XGrammar (ms)	51.76	59.45	77.74	104.06	121.46	-	-
	Pre^3 (ms)	49.91	53.71	54.41	61.63	75.47	-	-
15.7B 2×H800	Reduction	**↓3.57%**	**↓9.65%**	**↓30.01%**	**↓40.78%**	**↓37.86%**	-	-
Qwen2-14B (Yang et al., 2024a)	XGrammar (ms)	16.77	47.94	57.05	74.54	98.64	162.47	285.42
	Pre^3 (ms)	16.52	47.94	47.89	65.50	90.20	143.83	232.18
INT8 2×H800	Reduction	**↓1.52%**	**↓0.12%**	**↓2.37%**	**↓12.14%**	**↓8.55%**	**↓11.47%**	**↓18.65%**
Llama-2-70B (Touvron et al., 2023)	XGrammar (ms)	28.75	55.12	56.94	68.79	85.92	-	-
	Pre^3 (ms)	27.20	54.24	54.18	62.27	75.72	-	-
4×H800	Reduction	**↓5.39%**	**↓1.60%**	**↓4.85%**	**↓9.48%**	**↓11.87%**	-	-

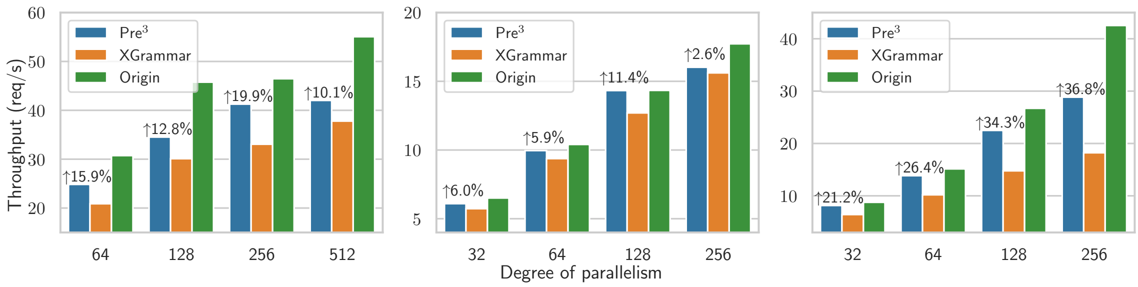

Figure 7: System throughput based on different models and concurrency levels. Left: Llama3-8B, Middle: Llama2-70B, Right: DeepSeek-V2-Lite-Chat.

Table 2: Per-step decode time comparison between our method and XGrammar.

	Llama-3-8B		Llama-2-70B	
Batchsize	Pre^3	XGrammar	Pre^3	XGrammar
1	**0.5172**	0.5531	**0.2163**	0.3030
4	**0.6537**	0.9327	**0.2407**	0.3310

significantly impact system performance.

We benchmark Pre^3 against the state-of-the-art XGrammar, using the JSON grammar for its complexity and challenging recursive structures (*e.g.*, lists and dictionaries). This tests the robustness and scalability of our method under demanding conditions.

Our experiments are conducted on multiple models of varying sizes and architectures. Specifically, we conducted experiments on Llama3-8B and Deepseek-V2 (15.7B) on a 2×H800 setup, and Llama2-70B on a 4×H800 setup. The maximum batch size goes to 1024, large enough to test the scalability of our method. In this experiment, we also measured the average time taken for each step, but the requests are batched in number to test the system's ability to process large batches.

The result is shown in Table 1. The results show that Pre^3 consistently outperforms XGrammar in all scenarios with latency reduction by up to 30%. The advantage is more significant at larger batch sizes, demonstrating the scalability of Pre^3.

4.4 Realworld Deployment

To evaluate the throughput in real-world service environments, we compare the performance of XGrammar and our method, Pre^3 under varying system concurrency levels. We conducted experiments on Meta-Llama-3-8B (2×H800) and Meta-Llama-2-70B (4×H800), measuring the throughput in terms of requests per second across different levels of concurrency.

The results are shown in Figure 7. Both Pre^3 and XGrammar have lower throughput than the Original system due to the added overhead introduced by constraint decoding, while Pre^3 demonstrated a significant improvement over XGrammar, achieving up to 20% higher throughput at higher concurrency levels, showing that Pre^3 provides higher throughput in end-to-end deployment.

5 Conclusion

In this work, we address the limitations of existing structured generation approaches by proposing a DPDA-based methodology (Pre^3), which integrates Cycle-aware Deterministic Pushdown Automata Construction and Prefix-conditioned Edge Optimization, Pre^3 significantly outperforms existing SOTA baselines by up to 40% in throughput and demonstrates greater advantages with large batch sizes.

Limitation

While our work demonstrates significant improvements in constrained LLM decoding efficiency, several limitations and potential areas for improvement remain.

Firstly, our method is designed and optimized for LR(1) grammars, which are sufficient for many structured generation tasks. However, it may face challenges when scaling to more complex or ambiguous grammars, such as those requiring LR(k) parsing (where $k > 1$). These grammars involve more intricate state transitions and lookahead mechanisms, which could increase the complexity of constructing and processing deterministic pushdown automata (DPDA). Extending the approach to handle such grammars while maintaining efficiency remains an open challenge. Future work could explore hybrid parsing strategies or adaptive mechanisms to dynamically adjust grammar complexity based on the input.

Secondly, our current implementation is primarily a research prototype and has not yet been fully engineered for production-level performance. The method is implemented in Python, which, while suitable for rapid development and experimentation, does not leverage the full potential of low-level optimizations or hardware acceleration. For instance, critical components such as transition processing and stack operations could benefit from parallelization on GPUs or specialized hardware. Additionally, the lack of fine-tuned memory management and efficient data structures limits the method's ability to scale to larger workloads. By reimplementing the approach in a systems-level language like C++ or Rust and incorporating hardware-aware optimizations, we could achieve even greater acceleration and performance gains.

Addressing these limitations could unlock additional performance improvements and broaden the applicability of our approach.

References

AV ASU86, R Sethi Aho, and Ullman JD. 1986. Compilers: Principles, techniques, and tools.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2023. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*.

Sukmin Cho, Soyeong Jeong, Jeong yeon Seo, and Jong Park. 2023. Discrete prompt optimization via constrained generation for zero-shot re-ranker. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 960–971, Toronto, Canada. Association for Computational Linguistics.

Yihong Dong, Xue Jiang, Yuchen Liu, Ge Li, and Zhi Jin. 2022. Codepad: Sequence-based code generation with pushdown automaton. *arXiv preprint arXiv:2211.00818*.

Yihong Dong, Ge Li, and Zhi Jin. 2023. Codep: grammatical seq2seq model for general-purpose code generation. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 188–198.

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. *Preprint*, arXiv:2411.15100.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John GF Francis. 1961. The qr transformation a unitary analogue to the lr transformation—part 1. *The Computer Journal*, 4(3):265–271.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. Generating structured outputs from language models: Benchmark and studies. *arXiv preprint arXiv:2501.10868*.

Georgi Gerganov. 2023. llama.cpp. LLM inference in C/C++.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

John E Hopcroft, Rajeev Motwani, and Jeffrey D Ullman. 2001. Introduction to automata theory, languages, and computation. *Acm Sigact News*, 32(1):60–65.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Donald E. Knuth. 1965. On the translation of languages from left to right. *Information and Control*, 8(6):607–639.

Terry Koo, Frederick Liu, and Luheng He. 2024. Automata-based constraints for language model decoding. *arXiv preprint arXiv:2407.08103*.

Zekun Li, Zhiyu Zoey Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Xin Luna Dong, Adithya Sagar, Xifeng Yan, and Paul A Crook. 2024a. Large language models as zero-shot dialogue state tracker through function calling. *arXiv preprint arXiv:2402.10466*.

Zelong Li, Wenyue Hua, Hao Wang, He Zhu, and Yongfeng Zhang. 2024b. Formal-llm: Integrating formal language and natural language for controllable llm-based agents. *arXiv preprint arXiv:2402.00798*.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

ModelTC. 2023. Lightllm. `https://github.com/ModelTC/lightllm`.

Mark-Jan Nederhof and Giorgio Satta. 1996. Efficient tabular lr parsing. *arXiv preprint cmp-lg/9605018*.

NousResearch. 2024. Nousresearch/json-mode-eval.

OpenAI. 2024. Learning to reason with llms.

Kanghee Park, Timothy Zhou, and Loris D'Antoni. 2025a. Flexible and efficient grammar-constrained decoding. *arXiv preprint arXiv:2502.05111*.

Kanghee Park, Timothy Zhou, and Loris D'Antoni. 2025b. Flexible and efficient grammar-constrained decoding. *Preprint*, arXiv:2502.05111.

Thomas Rückstieß, Alana Huang, and Robin Vujanic. 2024. Origami: A generative transformer architecture for predictions from semi-structured data. *arXiv preprint arXiv:2412.17348*.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. Picard: Parsing incrementally for constrained auto-regressive decoding from language models. *arXiv preprint arXiv:2109.05093*.

Michael Sipser. 1996. Introduction to the theory of computation. *ACM Sigact News*, 27(1):27–29.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shubham Ugare, Tarun Suresh, Hangoo Kang, Sasa Misailovic, and Gagandeep Singh. 2024. Syncode: Llm generation with grammar augmentation. *Preprint*, arXiv:2403.01632.

Leslie Valiant. 1973. *Decision procedures for families of deterministic pushdown automata*. Ph.D. thesis, University of Warwick.

Leslie G Valiant. 1975. Regularity and related problems for deterministic pushdown automata. *Journal of the ACM (JACM)*, 22(1):1–10.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Brandon T Willard and Rémi Louf. 2023a. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Brandon T. Willard and Rémi Louf. 2023b. Efficient guided generation for large language models. *Preprint*, arXiv:2307.09702.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. 2024b. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26275–26285.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. 2024. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877*.