Skip to main content

[![Cornell University](/static/browse/0.3.4/images/icons/cu/cornell-reduced-white-SMALL.svg)](https://www.cornell.edu/)

We gratefully acknowledge support from the Simons Foundation, [member institutions](https://info.arxiv.org/about/ourmembers.html), and all contributors. [Donate](https://info.arxiv.org/about/donate.html)

[](/IgnoreMe)

[![arxiv logo](/static/browse/0.3.4/images/arxiv-logo-one-color-white.svg)](/) > [cs](/list/cs/recent) > arXiv:2411.15100

[Help](https://info.arxiv.org/help) | [Advanced Search](https://arxiv.org/search/advanced)

All fields Title Author Abstract Comments Journal reference ACM classification MSC classification Report number arXiv identifier DOI ORCID arXiv author ID Help pages Full text

Search

[![arXiv logo](/static/browse/0.3.4/images/arxiv-logomark-small-white.svg)](https://arxiv.org/)

[![Cornell University Logo](/static/browse/0.3.4/images/icons/cu/cornell-reduced-white-SMALL.svg)](https://www.cornell.edu/)

open search

GO

open navigation menu

## quick links

  * [Login](https://arxiv.org/login)

  * [Help Pages](https://info.arxiv.org/help)

  * [About](https://info.arxiv.org/about)

# Computer Science > Computation and Language

arXiv:2411.15100 (cs)

[Submitted on 22 Nov 2024 ([v1](https://arxiv.org/abs/2411.15100v1)), last revised 27 Nov 2024 (this version, v2)]

# Title:XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models

Authors:[Yixin Dong](https://arxiv.org/search/cs?searchtype=author&query=Dong,+Y), [Charlie F. Ruan](https://arxiv.org/search/cs?searchtype=author&query=Ruan,+C+F), [Yaxing Cai](https://arxiv.org/search/cs?searchtype=author&query=Cai,+Y),

[Ruihang Lai](https://arxiv.org/search/cs?searchtype=author&query=Lai,+R), [Ziyi Xu](https://arxiv.org/search/cs?searchtype=author&query=Xu,+Z), [Yilong Zhao](https://arxiv.org/search/cs?searchtype=author&query=Zhao,+Y), [Tianqi Chen](https://arxiv.org/search/cs?searchtype=author&query=Chen,+T)

View a PDF of the paper titled XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models, by Yixin Dong and 6 other authors

[View PDF](/pdf/2411.15100) [HTML (experimental)](https://arxiv.org/html/2411.15100v2)

> Abstract:The applications of LLM Agents are becoming increasingly complex
> and diverse, leading to a high demand for structured outputs that can be
> parsed into code, structured function calls, and embodied agent commands.
> These developments bring significant demands for structured generation in
> LLM inference. Context-free grammar is a flexible approach to enable
> structured generation via constrained decoding. However, executing context-
> free grammar requires going through several stack states over all tokens in
> vocabulary during runtime, bringing non-negligible overhead for structured
> generation. In this paper, we propose XGrammar, a flexible and efficient
> structure generation engine for large language models. XGrammar accelerates
> context-free grammar execution by dividing the vocabulary into context-
> independent tokens that can be prechecked and context-dependent tokens that
> need to be interpreted during runtime. We further build transformations to
> expand the grammar context and reduce the number of context-independent
> tokens. Additionally, we build an efficient persistent stack to accelerate
> the context-dependent token checks. Finally, we co-design the grammar engine

> with LLM inference engine to overlap grammar computation with GPU

> executions. Evaluation results show that XGrammar can achieve up to 100x

> speedup over existing solutions. Combined with an LLM inference engine, it

> can generate near-zero overhead structure generation in end-to-end low-LLM

> serving.

Subjects: | Computation and Language (cs.CL); Artificial Intelligence (cs.AI); Programming Languages (cs.PL)

---|---

Cite as: | [arXiv:2411.15100](https://arxiv.org/abs/2411.15100) [cs.CL]

| (or [arXiv:2411.15100v2](https://arxiv.org/abs/2411.15100v2) [cs.CL] for this version)

| <https://doi.org/10.48550/arXiv.2411.15100> Focus to learn more arXiv-issued DOI via DataCite

## Submission history

From: Ruihang Lai [[view email](/show-email/a9d70a6f/2411.15100)]

[[v1]](/abs/2411.15100v1) Fri, 22 Nov 2024 18:01:37 UTC (527 KB)

[v2] Wed, 27 Nov 2024 18:59:28 UTC (531 KB)

Full-text links:

## Access Paper:

View a PDF of the paper titled XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models, by Yixin Dong and 6 other authors

* [View PDF](/pdf/2411.15100)

* [HTML (experimental)](https://arxiv.org/html/2411.15100v2)

* [TeX Source](/src/2411.15100)

* [Other Formats](/format/2411.15100)

Current browse context:

cs.CL

[< prev](/prevnext?id=2411.15100&function=prev&context=cs.CL "previous in cs.CL \(accesskey p\)") | [next >](/prevnext?id=2411.15100&function=next&context=cs.CL "next in cs.CL \(accesskey n\)")

[new](/list/cs.CL/new) | [recent](/list/cs.CL/recent) | [2024-11](/list/cs.CL/2024-11)

Change to browse by:

[cs](/abs/2411.15100?context=cs)

[cs.AI](/abs/2411.15100?context=cs.AI)

[cs.PL](/abs/2411.15100?context=cs.PL)

### References & Citations

* [NASA ADS](https://ui.adsabs.harvard.edu/abs/arXiv:2411.15100)

  * [Google Scholar](https://scholar.google.com/scholar_lookup?arxiv_id=2411.15100)

  * [Semantic Scholar](https://api.semanticscholar.org/arXiv:2411.15100)

[a](/static/browse/0.3.4/css/cite.css) export BibTeX citation Loading...

## BibTeX formatted citation

×

loading...

Data provided by:

### Bookmark

[![BibSonomy logo](/static/browse/0.3.4/images/icons/social/bibsonomy.png)
](http://www.bibsonomy.org/BibtexHandler?requTask=upload&url=https://arxiv.org/abs/2411.15100&
description=XGrammar:
Flexible and Efficient Structured Generation Engine for Large Language Models
"Bookmark on BibSonomy") [![Reddit
logo](/static/browse/0.3.4/images/icons/social/reddit.png)
](https://reddit.com/submit?url=https://arxiv.org/abs/2411.15100&title=XGrammar:
Flexible and Efficient Structured Generation Engine for Large Language Models
"Bookmark on Reddit")

Bibliographic Tools

# Bibliographic and Citation Tools

Bibliographic Explorer Toggle

Bibliographic Explorer _([What is the Explorer?](https://info.arxiv.org/labs/showcase.html#arxiv-bibliographic-explorer))_

Connected Papers Toggle

Connected Papers _([What is Connected Papers?](https://www.connectedpapers.com/about))_

Litmaps Toggle

Litmaps _([What is Litmaps?](https://www.litmaps.co/))_

scite.ai Toggle

scite Smart Citations _([What are Smart Citations?](https://www.scite.ai/))_

Code, Data, Media

# Code, Data and Media Associated with this Article

alphaXiv Toggle

alphaXiv _([What is alphaXiv?](https://alphaxiv.org/))_

Links to Code Toggle

CatalyzeX Code Finder for Papers _([What is CatalyzeX?](https://www.catalyzex.com))_

DagsHub Toggle

DagsHub _([What is DagsHub?](https://dagshub.com/))_

GotitPub Toggle

Gotit.pub _([What is GotitPub?](http://gotit.pub/faq))_

Huggingface Toggle

Hugging Face _([What is Huggingface?](https://huggingface.co/huggingface))_

Links to Code Toggle

Papers with Code _([What is Papers with Code?](https://paperswithcode.com/))_

ScienceCast Toggle

ScienceCast _([What is ScienceCast?](https://sciencecast.org/welcome))_

Demos

# Demos

Replicate Toggle

Replicate _([What is Replicate?](https://replicate.com/docs/arxiv/about))_

Spaces Toggle

Hugging Face Spaces _([What is Spaces?](https://huggingface.co/docs/hub/spaces))_

Spaces Toggle

TXYZ.AI _([What is TXYZ.AI?](https://txyz.ai))_

Related Papers

# Recommenders and Search Tools

Link to Influence Flower

Influence Flower _([What are Influence Flowers?](https://influencemap.cmlab.dev/))_

Core recommender toggle

CORE Recommender _([What is CORE?](https://core.ac.uk/services/recommender))_

* Author
* Venue
* Institution
* Topic

About arXivLabs

# arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? [**Learn more about arXivLabs**](https://info.arxiv.org/labs/index.html).

[Which authors of this paper are endorsers?](/auth/show-endorsers/2411.15100) | [Disable MathJax](javascript:setMathjaxCookie\(\)) ([What is MathJax?](https://info.arxiv.org/help/mathjax.html))

* [About](https://info.arxiv.org/about)

* [Help](https://info.arxiv.org/help)

* contact arXivClick here to contact arXiv [Contact](https://info.arxiv.org/help/contact.html)

* subscribe to arXiv mailingsClick here to subscribe [Subscribe](https://info.arxiv.org/help/subscribe)

* [Copyright](https://info.arxiv.org/help/license/index.html)

* [Privacy Policy](https://info.arxiv.org/help/policies/privacy_policy.html)

* [Web Accessibility Assistance](https://info.arxiv.org/help/web_accessibility.html)

* [arXiv Operational Status](https://status.arxiv.org)

Get status notifications via

[email](https://subscribe.sorryapp.com/24846f03/email/new) or

[slack](https://subscribe.sorryapp.com/24846f03/slack/new)