Skip to main content

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

In just 3 minutes help us improve arXiv:

[Annual Global Survey](https://cornell.ca1.qualtrics.com/jfe/form/SV_6m22mbqW9GQ3pQO)

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:2405.04304

[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

# Computer Science > Computation and Language

arXiv:2405.04304 (cs)

[Submitted on 7 May 2024 ([v1](https://arxiv.org/abs/2405.04304v1)), last revised 7 Nov 2024 (this version, v5)]

# Title:Dynamic Speculation Lookahead Accelerates Speculative Decoding of Large Language Models

Authors:[Jonathan Mamou](https://arxiv.org/search/cs?searchtype=author&query=Mamou,+J), [Oren Pereg](https://arxiv.org/search/cs?searchtype=author&query=Pereg,+O), [Daniel Korat](https://arxiv.org/search/cs?searchtype=author&query=Korat,+D), [Moshe Berchansky](https://arxiv.org/search/cs?searchtype=author&query=Berchansky,+M), [Nadav Timor](https://arxiv.org/search/cs?searchtype=author&query=Timor,+N), [Moshe Wasserblat](https://arxiv.org/search/cs?searchtype=author&query=Wasserblat,+M), [Roy Schwartz](https://arxiv.org/search/cs?searchtype=author&query=Schwartz,+R)

View a PDF of the paper titled Dynamic Speculation Lookahead Accelerates Speculative Decoding of Large Language Models, by Jonathan Mamou and Oren Pereg and Daniel Korat and Moshe Berchansky and Nadav Timor and Moshe Wasserblat and Roy Schwartz

[View PDF](/pdf/2405.04304) [HTML (experimental)](https://arxiv.org/html/2405.04304v5)

> Abstract:Speculative decoding is commonly used for reducing the inference
> latency of large language models. Its effectiveness depends highly on the
> speculation lookahead (SL)-the number of tokens generated by the draft model
> at each iteration. In this work we show that the common practice of using
> the same SL for all iterations (static SL) is suboptimal. We introduce DISCO
> (DynamIc SpeCulation lookahead Optimization), a novel method for dynamically
> selecting the SL. Our experiments with four datasets show that DISCO reaches

> an average speedup of 10% compared to the best static SL baseline, while

> generating the exact same text.

Subjects: | Computation and Language (cs.CL)

---|---

Cite as: | [arXiv:2405.04304](https://arxiv.org/abs/2405.04304) [cs.CL]

| (or [arXiv:2405.04304v5](https://arxiv.org/abs/2405.04304v5) [cs.CL] for this version)

| <https://doi.org/10.48550/arXiv.2405.04304> Focus to learn more arXiv-issued DOI via DataCite

## Submission history

From: Jonathan Mamou [[view email](/show-email/65b647c9/2405.04304)]

[[v1]](/abs/2405.04304v1) Tue, 7 May 2024 13:27:52 UTC (674 KB)

[[v2]](/abs/2405.04304v2) Tue, 18 Jun 2024 12:34:35 UTC (677 KB)

[[v3]](/abs/2405.04304v3) Wed, 19 Jun 2024 08:54:51 UTC (677 KB)

[[v4]](/abs/2405.04304v4) Sun, 23 Jun 2024 13:46:23 UTC (677 KB)

[v5] Thu, 7 Nov 2024 12:59:39 UTC (676 KB)

Full-text links:

## Access Paper:

View a PDF of the paper titled Dynamic Speculation Lookahead Accelerates Speculative Decoding of Large Language Models, by Jonathan Mamou and Oren Pereg and Daniel Korat and Moshe Berchansky and Nadav Timor and Moshe Wasserblat and Roy Schwartz

* [View PDF](/pdf/2405.04304)
* [HTML (experimental)](https://arxiv.org/html/2405.04304v5)
* [TeX Source](/src/2405.04304)
* [Other Formats](/format/2405.04304)

[![license icon](https://arxiv.org/icons/licenses/by-4.0.png) view license](http://creativecommons.org/licenses/by/4.0/ "Rights to this article")

Current browse context:

cs.CL

[< prev](/prevnext?id=2405.04304&function=prev&context=cs.CL "previous in cs.CL \(accesskey p\)") | [next >](/prevnext?id=2405.04304&function=next&context=cs.CL "next in cs.CL \(accesskey n\)")

[new](/list/cs.CL/new) | [recent](/list/cs.CL/recent) | [2024-05](/list/cs.CL/2024-05)

Change to browse by:

[cs](/abs/2405.04304?context=cs)

### References & Citations

* [NASA ADS](https://ui.adsabs.harvard.edu/abs/arXiv:2405.04304)
* [Google Scholar](https://scholar.google.com/scholar_lookup?arxiv_id=2405.04304)

  * [Semantic Scholar](https://api.semanticscholar.org/arXiv:2405.04304)

[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/2405.04304&
description=Dynamic
Speculation Lookahead Accelerates Speculative Decoding of 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/2405.04304&title=Dynamic
Speculation Lookahead Accelerates Speculative Decoding of 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/2405.04304) | [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)