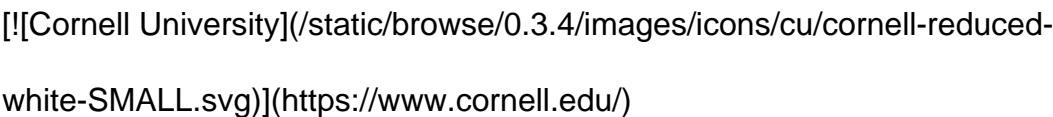


Skip to main content


 (https://www.cornell.edu/)

In just 3 minutes help us improve arXiv:

[Annual Global Survey](https://cornell.ca1.qualtrics.com/jfe/form/SV_6m22mbqW9GQ3pQO) (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) (https://info.arxiv.org/about/ourmembers.html), and all contributors. [Donate](https://info.arxiv.org/about/donate.html) (https://info.arxiv.org/about/donate.html)

[IgnoreMe](#)

 (/static/browse/0.3.4/images/arxiv-logo-one-color-white.svg) (/)
> [\[cs\]](/list/cs/recent) (/list/cs/recent) > arXiv:2401.03462

[\[Help\]](https://info.arxiv.org/help) (https://info.arxiv.org/help) | [\[Advanced Search\]](https://arxiv.org/search/advanced) (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:2401.03462 (cs)

[Submitted on 7 Jan 2024 ([v1](https://arxiv.org/abs/2401.03462v1)), last revised 11 Oct 2024 (this version, v3)]

Title:Long Context Compression with Activation Beacon

Authors:[Peitian

Zhang](<https://arxiv.org/search/cs?searchtype=author&query=Zhang,+P>), [Zheng

Liu](<https://arxiv.org/search/cs?searchtype=author&query=Liu,+Z>), [Shitao

Xiao](<https://arxiv.org/search/cs?searchtype=author&query=Xiao,+S>), [Ninglu

Shao](<https://arxiv.org/search/cs?searchtype=author&query=Shao,+N>), [Qiwei

Ye](<https://arxiv.org/search/cs?searchtype=author&query=Ye,+Q>), [Zhicheng

Dou](<https://arxiv.org/search/cs?searchtype=author&query=Dou,+Z>)

View a PDF of the paper titled Long Context Compression with Activation

Beacon, by Peitian Zhang and 5 other authors

[View PDF](/pdf/2401.03462) [HTML

(experimental)](<https://arxiv.org/html/2401.03462v3>)

> Abstract:Long context compression is a critical research problem due to its
> significance in reducing the high computational and memory costs associated
> with LLMs. In this paper, we propose Activation Beacon, a plug-in module for
> transformer-based LLMs that targets effective, efficient, and flexible
> compression of long contexts. To achieve this, our method introduces the
> following technical designs. 1) We directly compress the activations (i.e.
> keys and values at every layer), rather than leveraging soft prompts to
> relay information (which constitute a major bottleneck to encapsulate the
> complex information within long contexts). 2) We tailor the compression
> workflow, where each fine-grained input unit is progressively compressed,
> enabling high-quality compression and efficient computation during both
> training and inference. 3) We train the model through compression-based
> auto-regression, making full use of plain texts and instructional data to

> optimize the model's compression performance. 4) During training, we

> randomly sample a compression ratio at each step, teaching the model to

> support a wide range of compression configurations. Extensive evaluations

> are conducted on various long-context tasks whose lengths (e.g., 128K) may

> far exceed the maximum training length (20K), such as document

> understanding, few-shot learning, and Needle-in-a-Haystack. Whilst existing

> methods struggle to handle these challenging tasks, Activation Beacon

> maintains a comparable performance to the uncompressed baseline across

> various scenarios, achieving a 2x acceleration in inference time and an 8x

> reduction of memory costs for KV cache. Our data, model, and code have been

> released at [\url{this https](https://github.com/FlagOpen/FlagEmbedding/)

> URL](<https://github.com/FlagOpen/FlagEmbedding/>)}.

Comments: | Newer version of Activation Beacon

---|---

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

Cite as: | [arXiv:2401.03462](<https://arxiv.org/abs/2401.03462>) [cs.CL]

| (or [arXiv:2401.03462v3](<https://arxiv.org/abs/2401.03462v3>) [cs.CL] for this version)

| <<https://doi.org/10.48550/arXiv.2401.03462>> Focus to learn more arXiv-issued DOI via DataCite

Submission history

From: Peitian Zhang [[view email]](/show-email/b64e71d2/2401.03462)]

[[v1]](/abs/2401.03462v1) Sun, 7 Jan 2024 11:57:40 UTC (492 KB)

[[v2]](/abs/2401.03462v2) Fri, 2 Feb 2024 12:34:25 UTC (162 KB)


[v3] Fri, 11 Oct 2024 02:18:24 UTC (202 KB)

Full-text links:

Access Paper:

View a PDF of the paper titled Long Context Compression with Activation Beacon, by Peitian Zhang and 5 other authors

- * [View PDF](/pdf/2401.03462)
- * [HTML (experimental)](https://arxiv.org/html/2401.03462v3)
- * [TeX Source](/src/2401.03462)
- * [Other Formats](/format/2401.03462)

[ (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=2401.03462&function=prev&context=cs.CL "previous in cs.CL \(\accesskey p\)") | [next >](/prevnext?id=2401.03462&function=next&context=cs.CL "next in cs.CL \(\accesskey n\)")

[new](/list/cs.CL/new) | [recent](/list/cs.CL/recent) | [2024-01](/list/cs.CL/2024-01)

Change to browse by:

[cs](/abs/2401.03462?context=cs)

[cs.AI](/abs/2401.03462?context=cs.AI)

References & Citations

- * [NASA ADS](https://ui.adsabs.harvard.edu/abs/arXiv:2401.03462)
- * [Google Scholar](https://scholar.google.com/scholar_lookup?arxiv_id=2401.03462)
- * [Semantic Scholar](https://api.semanticscholar.org/arXiv:2401.03462)

[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/2401.03462&description=Long

Context Compression with Activation Beacon "Bookmark on BibSonomy") [![Reddit logo](/static/browse/0.3.4/images/icons/social/reddit.png)

](https://reddit.com/submit?url=https://arxiv.org/abs/2401.03462&title=Long
Context Compression with Activation Beacon "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 [_\(\[What is alphaXiv?\]\(https://alphaxiv.org/\)\)_](https://alphaxiv.org/)

Links to Code Toggle

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

DagsHub Toggle

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

GotitPub Toggle

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

Huggingface Toggle

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

Links to Code Toggle

Papers with Code [_\(\[What is Papers with Code?\]\(https://paperswithcode.com/\)\)_](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 [_\(\[What are Influence Flowers?\]\)\(https://influencemap.cmlab.dev/\)\)_](https://influencemap.cmlab.dev/)

Core recommender toggle

CORE Recommender [_\(\[What is CORE?\]\)\(https://core.ac.uk/services/recommender\)\)_](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? [\[**Learn more about arXivLabs**\]\(https://info.arxiv.org/labs/index.html\)](https://info.arxiv.org/labs/index.html).

[\[Which authors of this paper are endorsers?\]](/auth/show-endorsers/2401.03462) | [\[Disable MathJax\]](#)([javascript:setMathjaxCookie\\(\)](#)) [\(\[What is MathJax?\]](#)(<https://info.arxiv.org/help/mathjax.html>))

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

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

* [contact arXiv](#)[Click here to contact arXiv](#) [\[Contact\]](https://info.arxiv.org/help/contact.html)(<https://info.arxiv.org/help/contact.html>)

* [subscribe to arXiv mailings](#)[Click here to subscribe](#) [\[](#)
[Subscribe\]](https://info.arxiv.org/help/subscribe)(<https://info.arxiv.org/help/subscribe>)

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

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

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

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

Get status notifications via

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

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