[![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:2306.00978** (cs)

[Submitted on 1 Jun 2023 ([v1](https://arxiv.org/abs/2306.00978v1)), last revised 18 Jul 2024 (this version, v5)]

# Title:AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration

Authors:[Ji Lin](https://arxiv.org/search/cs?searchtype=author&query=Lin,+J),
[Jiaming Tang](https://arxiv.org/search/cs?searchtype=author&query=Tang,+J),
[Haotian Tang](https://arxiv.org/search/cs?searchtype=author&query=Tang,+H),
[Shang Yang](https://arxiv.org/search/cs?searchtype=author&query=Yang,+S),
[Wei-Ming Chen](https://arxiv.org/search/cs?searchtype=author&query=Chen,+W),
[Wei-Chen Wang](https://arxiv.org/search/cs?searchtype=author&query=Wang,+W),
[Guangxuan Xiao](https://arxiv.org/search/cs?searchtype=author&query=Xiao,+G),
[Xingyu Dang](https://arxiv.org/search/cs?searchtype=author&query=Dang,+X),
[Chuang Gan](https://arxiv.org/search/cs?searchtype=author&query=Gan,+C),
[Song Han](https://arxiv.org/search/cs?searchtype=author&query=Han,+S)

View a PDF of the paper titled AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration, by Ji Lin and 9 other authors

[View PDF](/pdf/2306.00978) [HTML (experimental)](https://arxiv.org/html/2306.00978v5)

> Abstract:Large language models (LLMs) have transformed numerous AI
> applications. On-device LLM is becoming increasingly important: running LLMs
> locally on edge devices can reduce the cloud computing cost and protect
> users' privacy. However, the astronomical model size and the limited
> hardware resource pose significant deployment challenges. We propose
> Activation-aware Weight Quantization (AWQ), a hardware-friendly approach for
> LLM low-bit weight-only quantization. AWQ finds that not all weights in an
> LLM are equally important. Protecting only 1% salient weights can greatly
> reduce quantization error. To identify salient weight channels, we should

> refer to the activation distribution, not weights. To avoid the hardware-
> inefficient mix-precision quantization, we mathematically derive that
> scaling up the salient channels can reduce the quantization error. AWQ
> employs an equivalent transformation to scale the salient weight channels to
> protect them. The scale is determined by collecting the activation
> statistics offline. AWQ does not rely on any backpropagation or
> reconstruction, so it generalizes to different domains and modalities
> without overfitting the calibration set. AWQ outperforms existing work on
> various language modeling and domain-specific benchmarks (coding and math).
> Thanks to better generalization, it achieves excellent quantization
> performance for instruction-tuned LMs and, for the first time, multi-modal
> LMs. Alongside AWQ, we implement TinyChat, an efficient and flexible
> inference framework tailored for 4-bit on-device LLM/VLMs. With kernel
> fusion and platform-aware weight packing, TinyChat offers more than 3x
> speedup over the Huggingface FP16 implementation on both desktop and mobile
> GPUs. It also democratizes the deployment of the 70B Llama-2 model on mobile
> GPUs.

Comments: | MLSys 2024 Best Paper Award. Code available at: [this https URL](https://github.com/mit-han-lab/llm-awq)

---|---

Subjects: | Computation and Language (cs.CL)

Cite as: | [arXiv:2306.00978](https://arxiv.org/abs/2306.00978) [cs.CL]

| (or [arXiv:2306.00978v5](https://arxiv.org/abs/2306.00978v5) [cs.CL] for this version)

| <https://doi.org/10.48550/arXiv.2306.00978> Focus to learn more arXiv-issued DOI via DataCite

## Submission history

From: Haotian Tang [[view email](/show-email/9f84e992/2306.00978)]

**[[v1]](/abs/2306.00978v1)** Thu, 1 Jun 2023 17:59:10 UTC (2,783 KB)

**[[v2]](/abs/2306.00978v2)** Tue, 3 Oct 2023 18:20:01 UTC (4,384 KB)

**[[v3]](/abs/2306.00978v3)** Sun, 21 Apr 2024 03:47:49 UTC (24,553 KB)

**[[v4]](/abs/2306.00978v4)** Tue, 23 Apr 2024 19:51:53 UTC (24,552 KB)

**[v5]** Thu, 18 Jul 2024 17:51:33 UTC (18,170 KB)

Full-text links:

## Access Paper:

View a PDF of the paper titled AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration, by Ji Lin and 9 other authors

* [View PDF](/pdf/2306.00978)
* [HTML (experimental)](https://arxiv.org/html/2306.00978v5)
* [TeX Source](/src/2306.00978)
* [Other Formats](/format/2306.00978)



Current browse context:

cs.CL

[< prev](/prevnext?id=2306.00978&function=prev&context=cs.CL "previous in cs.CL \(accesskey p\)") | [next >](/prevnext?id=2306.00978&function=next&context=cs.CL "next in cs.CL \(accesskey n\)")

[new](/list/cs.CL/new) | [recent](/list/cs.CL/recent) | [2023-06](/list/cs.CL/2023-06)

Change to browse by:

[cs](/abs/2306.00978?context=cs)

### References & Citations

* [NASA ADS](https://ui.adsabs.harvard.edu/abs/arXiv:2306.00978)
* [Google Scholar](https://scholar.google.com/scholar_lookup?arxiv_id=2306.00978)
* [Semantic Scholar](https://api.semanticscholar.org/arXiv:2306.00978)

[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/2306.00978&
description=AWQ:
Activation-aware Weight Quantization for LLM Compression and Acceleration
"Bookmark on BibSonomy") [ ![Reddit
logo](/static/browse/0.3.4/images/icons/social/reddit.png)
](https://reddit.com/submit?url=https://arxiv.org/abs/2306.00978&title=AWQ:
Activation-aware Weight Quantization for LLM Compression and Acceleration
"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/2306.00978) | [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)