[![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:2309.03883 (cs)

[Submitted on 7 Sep 2023 ([v1](https://arxiv.org/abs/2309.03883v1)), last revised 11 Mar 2024 (this version, v2)]

Title:DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models

Authors:[Yung-Sung Chuang](https://arxiv.org/search/cs?searchtype=author&query=Chuang,+Y), [Yujia Xie](https://arxiv.org/search/cs?searchtype=author&query=Xie,+Y), [Hongyin Luo](https://arxiv.org/search/cs?searchtype=author&query=Luo,+H), [Yoon Kim](https://arxiv.org/search/cs?searchtype=author&query=Kim,+Y), [James Glass](https://arxiv.org/search/cs?searchtype=author&query=Glass,+J), [Pengcheng He](https://arxiv.org/search/cs?searchtype=author&query=He,+P)

View a PDF of the paper titled DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models, by Yung-Sung Chuang and 5 other authors

[View PDF](/pdf/2309.03883) [HTML (experimental)](https://arxiv.org/html/2309.03883v2)

> Abstract:Despite their impressive capabilities, large language models (LLMs)
> are prone to hallucinations, i.e., generating content that deviates from
> facts seen during pretraining. We propose a simple decoding strategy for
> reducing hallucinations with pretrained LLMs that does not require
> conditioning on retrieved external knowledge nor additional fine-tuning. Our
> approach obtains the next-token distribution by contrasting the differences
> in logits obtained from projecting the later layers versus earlier layers to
> the vocabulary space, exploiting the fact that factual knowledge in an LLMs
> has generally been shown to be localized to particular transformer layers.
> We find that this Decoding by Contrasting Layers (DoLa) approach is able to
> better surface factual knowledge and reduce the generation of incorrect
> facts. DoLa consistently improves the truthfulness across multiple choices

> tasks and open-ended generation tasks, for example improving the performance

> of LLaMA family models on TruthfulQA by 12-17% absolute points,

> demonstrating its potential in making LLMs reliably generate truthful facts.

Comments: | ICLR 2024 main conference paper. The source code is available at [this https URL](https://github.com/voidism/DoLa)

---|---

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

Cite as: | [arXiv:2309.03883](https://arxiv.org/abs/2309.03883) [cs.CL]

| (or [arXiv:2309.03883v2](https://arxiv.org/abs/2309.03883v2) [cs.CL] for this version)

| <https://doi.org/10.48550/arXiv.2309.03883> Focus to learn more arXiv-issued DOI via DataCite

Submission history

From: Yung-Sung Chuang [[view email](/show-email/23e01694/2309.03883)]

[[v1]](/abs/2309.03883v1) Thu, 7 Sep 2023 17:45:31 UTC (238 KB)

[v2] Mon, 11 Mar 2024 02:01:09 UTC (243 KB)

Full-text links:

Access Paper:

View a PDF of the paper titled DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models, by Yung-Sung Chuang and 5 other authors

* [View PDF](/pdf/2309.03883)
* [HTML (experimental)](https://arxiv.org/html/2309.03883v2)
* [TeX Source](/src/2309.03883)
* [Other Formats](/format/2309.03883)

Current browse context:

cs.CL

[< prev](/prevnext?id=2309.03883&function=prev&context=cs.CL "previous in cs.CL \(accesskey p\)") | [next >](/prevnext?id=2309.03883&function=next&context=cs.CL "next in cs.CL \(accesskey n\)")

[new](/list/cs.CL/new) | [recent](/list/cs.CL/recent) | [2023-09](/list/cs.CL/2023-09)

Change to browse by:

[cs](/abs/2309.03883?context=cs)
[cs.AI](/abs/2309.03883?context=cs.AI)
[cs.LG](/abs/2309.03883?context=cs.LG)

References & Citations

* [NASA ADS](https://ui.adsabs.harvard.edu/abs/arXiv:2309.03883)

* [Google Scholar](https://scholar.google.com/scholar_lookup?arxiv_id=2309.03883)

* [Semantic Scholar](https://api.semanticscholar.org/arXiv:2309.03883)

[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/2309.03883&

description=DoLa:

Decoding by Contrasting Layers Improves Factuality in 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/2309.03883&title=DoLa:

Decoding by Contrasting Layers Improves Factuality in 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/2309.03883) | [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)