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:2312.15503

[![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:2312.15503 (cs)

[Submitted on 24 Dec 2023]

Title:Making Large Language Models A Better Foundation For Dense Retrieval

Authors:[Chaofan

Li](https://arxiv.org/search/cs?searchtype=author&query=Li,+C), [Zheng Liu](https://arxiv.org/search/cs?searchtype=author&query=Liu,+Z), [Shitao Xiao](https://arxiv.org/search/cs?searchtype=author&query=Xiao,+S), [Yingxia Shao](https://arxiv.org/search/cs?searchtype=author&query=Shao,+Y)

View a PDF of the paper titled Making Large Language Models A Better Foundation For Dense Retrieval, by Chaofan Li and 3 other authors

[View PDF](/pdf/2312.15503) [HTML (experimental)](https://arxiv.org/html/2312.15503v1)

> Abstract:Dense retrieval needs to learn discriminative text embeddings to
> represent the semantic relationship between query and document. It may
> benefit from the using of large language models (LLMs), given LLMs' strong
> capability on semantic understanding. However, the LLMs are pre-trained by
> text generation tasks, whose working pattern is completely different from
> representing texts as embeddings. As a result, it is imperative to study how
> to adapt LLMs properly so that they can be effectively initialized as the
> backbone encoder for dense retrieval.
> In this paper, we propose a novel approach, called LLaRA (LLM adapted for
> dense RetrievAl), which works as a post-hoc adaptation of LLM for the dense
> retrieval application. LLaRA consists of two pretext tasks: EBAE (Embedding-
> Based Auto-Encoding) and EBAR (Embedding-Based Auto-Regression), where the
> text embeddings from LLM are used to reconstruct the tokens for the input
> sentence and predict the tokens for the next sentence, respectively. LLaRA
> turns out to be simple, lightweight, and highly effective. It is applied to
> adapt LLaMA-2-7B (base) on the Wikipedia corpus, where it substantially

> improves the model's fine-tuned performances on a variety of dense retrieval

> benchmarks, like MSMARCO and BEIR. Our model and code will be made publicly

> available at BGE repository.

Subjects: | Computation and Language (cs.CL)

---|---

Cite as: | [arXiv:2312.15503](https://arxiv.org/abs/2312.15503) [cs.CL]

| (or [arXiv:2312.15503v1](https://arxiv.org/abs/2312.15503v1) [cs.CL] for this version)

| <https://doi.org/10.48550/arXiv.2312.15503> Focus to learn more arXiv-issued DOI via DataCite

Submission history

From: Zheng Liu [[view email](/show-email/30c7c6f4/2312.15503)]

[v1] Sun, 24 Dec 2023 15:10:35 UTC (7,081 KB)

Full-text links:

Access Paper:

View a PDF of the paper titled Making Large Language Models A Better

Foundation For Dense Retrieval, by Chaofan Li and 3 other authors

* [View PDF](/pdf/2312.15503)
* [HTML (experimental)](https://arxiv.org/html/2312.15503v1)
* [TeX Source](/src/2312.15503)
* [Other Formats](/format/2312.15503)

[![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=2312.15503&function=prev&context=cs.CL "previous in cs.CL \(accesskey p\)") | [next >](/prevnext?id=2312.15503&function=next&context=cs.CL "next in cs.CL \(accesskey n\)")

[new](/list/cs.CL/new) | [recent](/list/cs.CL/recent) | [2023-12](/list/cs.CL/2023-12)

Change to browse by:

[cs](/abs/2312.15503?context=cs)

References & Citations

* [NASA ADS](https://ui.adsabs.harvard.edu/abs/arXiv:2312.15503)
* [Google Scholar](https://scholar.google.com/scholar_lookup?arxiv_id=2312.15503)
* [Semantic Scholar](https://api.semanticscholar.org/arXiv:2312.15503)

[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/2312.15503&
description=Making
Large Language Models A Better Foundation For Dense Retrieval "Bookmark on
BibSonomy") [![Reddit
logo](/static/browse/0.3.4/images/icons/social/reddit.png)
](https://reddit.com/submit?url=https://arxiv.org/abs/2312.15503&title=Making
Large Language Models A Better Foundation For Dense Retrieval "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/2312.15503) | [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)