Neuralese

Contents hide

(Top)
See also
References

From Wikipedia, the free encyclopedia

This article **is an orphan, as no other articles link to it**. Please introduce links to this page from related articles. *(November 2026)*

In artificial intelligence (AI) research, **neuralese** is a method where large language models (LLMs) perform intermediate reasoning steps in their high-dimensional latent space (vector embeddings) rather than outputting human-readable text tokens. While standard chain-of-thought reasoning forces a model to generate a sequence of words, neuralese allows the model to pass raw, continuous vectors between computational layers, creating a high-bandwidth, non-linguistic reasoning channel.[1][2]

See also [edit]

- Artificial intelligence safety
- Interpretability (machine learning)

Look up ***neuralese*** in Wiktionary, the free dictionary.

References [edit]

1. ^ Hao, Shibo, et al. "Training Large Language Models to Reason in a Continuous Latent Space." *arXiv preprint arXiv:2412.06769*, 9 December 2024.
2. ^ "Translating Neuralese". *ar5iv*. Retrieved 2026-08-24.

This artificial intelligence–related article is a stub. You can help Wikipedia by adding missing information.

Categories: Artificial intelligence stubs | Artificial intelligence | Machine learning | Artificial intelligence engineering | Large language models

This page was last edited on 18 November 2026, at 13:18 (UTC). Page was rendered with Parsoid.

Text is available under the Creative Commons Attribution-ShareAlike 4.0 License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

Privacy policy About Wikipedia Disclaimers Contact Wikipedia Legal & safety contacts Code of Conduct Developers Statistics Cookie statement Mobile view

a WIKIMEDIA project Powered by MediaWiki